Silhouette-based Object Phenotype Recognition using 3D Shape Priors

Yu Chen[1] Tae-Kyun Kim[2] Roberto Cipolla[1]
Department of Engineering, University of Cambridge, Cambridge, UK[1]
Department of Electrical Engineering, Imperial College, London, UK[2]
yc301@cam.ac.uk tk.kim@imperial.ac.uk rc10001@cam.ac.uk

Abstract

This paper tackles the novel challenging problem of 3D object phenotype recognition from a single 2D silhouette. To bridge the large pose (articulation or deformation) and camera viewpoint changes between the gallery images and query image, we propose a novel probabilistic inference algorithm based on 3D shape priors. Our approach combines both generative and discriminative learning. We use latent probabilistic generative models to capture 3D shape and pose variations from a set of 3D mesh models. Based on these 3D shape priors, we generate a large number of projections for different phenotype classes, poses, and camera viewpoints, and implement Random Forests to efficiently solve the shape and pose inference problems. By model selection in terms of the silhouette coherency between the query and the projections of 3D shapes synthesized using the galleries, we achieve the phenotype recognition result as well as a fast approximate 3D reconstruction of the query. To verify the efficacy of the proposed approach, we present new datasets which contain over 500 images of various human and shark phenotypes and motions. The experimental results clearly show the benefits of using the 3D priors in the proposed method over previous 2D-based methods.

1. Introduction

Recognizing 3D objects from one or more 2D views is a fundamental problem in computer vision. There have been increasing attempts to solve this problem, which embraces a number of research issues such as view-invariant object instance/category recognition [11, 12, 15, 24, 32, 34], object pose recognition [13, 17, 22, 25, 31, 29, 33], object viewpoint classification [9], gait recognition [18], face recognition across pose and expression [20, 36], etc. However, to our best knowledge, the problem of classifying generic object phenotypes (shapes), under 3D object pose and camera view-point changes, has not been tackled. The successful solutions would be widely useful for potential applications such as automatic human body shape monitoring, in relation with recent food recognition studies in computer vision, and wild animal (such as horse and fish) tracking, etc.

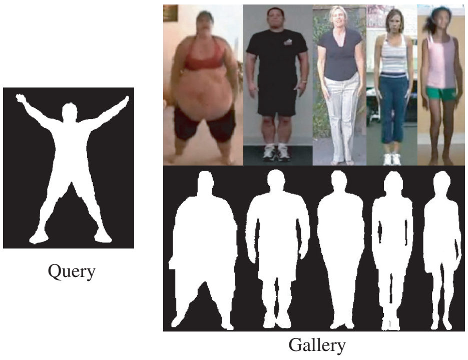


Figure 1. **Phenotype recognition problem.** Given a silhouette gallery of different body shapes, the goal is to classify the body shape of a query silhouette in the presence of pose and/or camera viewpoint changes.

In this work, we address a novel challenging task of shape recognition, i.e. classifying phenotypes of the 3D object from a *single* 2D silhouette input (see Fig. 1 for an example of human body shapes). Here, phenotypes are referred to the intrinsic shape differences across given human instances, e.g., fat vs thin, tall vs short, muscular vs unmuscular. The major difficulty of this problem is that the query silhouette can undergo large pose and camera view-point changes. Traditional 2D-based approaches fail to capture the intrinsic shape (dis-)similarity between the query and gallery silhouettes. In view of this problem, we propose a novel generative+discriminative solution by using 3D shape priors, i.e. the knowledge learnt from previously-seen 3D shapes. Our approach is motivated by the observation that humans can perceive the 3D shape of an object from a single image, provided that they have seen similar 3D shapes. Once 3D shapes are estimated from single images (single view reconstruction), camera view-point/pose invariant object recognition is achievable.

The problem we tackle, therefore, conjoins single view reconstruction with 3D object recognition. The novelties and main contributions lie in:

- Going beyond pose recognition: object pose recog-

978-1-4577-1102-2/11/$26.00 ©2011 IEEE

nition and tracking by 3D template models has been widely studied [13, 22, 25, 27, 29, 35]. This work attempts to capture more subtle 3D shape variations on the top of the estimated pose and camera view-points.

- Recognising generic deformable objects: our framework does not require strong class-specific knowledge such as human body skeleton consisting of a number of joints in [6, 28] or face shape models defined by manual control points [36], and is thus applicable to different object categories. Previous studies [12, 15, 23, 34] are limited to rigid object classes.
- Exploiting shape cues (vs textures): whereas a majority of existing 3D object recognition work relies on image appearance or textures (e.g., affine invariant patches [24, 32]), we exploit shape cues, silhouettes, which are useful when there is no overlap in views between a model and a query, or no consistent textures e.g. changing clothes etc.
- Transferring 3D models to images: we learn from 3D models and perform recognition of images, which contrasts previous work matching only among 3D models [4] or 2D images.

1.1. Related Work

There has been a growing interest for view-invariant object instance or category recognition [24, 32]. Their building blocks are often the image patches that are invariant up to affine transformations, and the structural relations among the patches are then captured. Texture-based object recognition is useful manywhere though, it becomes inherently ambiguous when there are no consistent textures between a model and a query: no overlapping views, changing clothes, or textureless objects.

Shape (silhouette or edge map) is another useful cue which has been long explored for object recognition, however most relevant studies have been done in 2D [5, 30]. They do not explicitly capture 3D shapes, poses, camera view-points of objects, relying on a large number of model images. It basically fails when query images exhibit considerably different poses or view-points from those of models. On the other hand, studies on 2D shape representation [8, 16] have tackled the problem of recognizing articulated objects, but they model the articulation on a 2D basis and have difficulties dealing with self-occlusions and large 3D camera pose changes.

3D templates and shape models have been widely incorporated into object pose recognition problems for hands [31, 25] or human bodies [17, 22, 29, 33, 35], but their models are designed for pose, often without consideration of shape variations. Whereas they do not explicitly handle the classification problem of phenotypes or 3D shapes, we capture and discriminate 3D shape variations in an invariant manner to object poses and camera view-points.

Single view reconstruction is an active research field. Just to name a few, Prasad et al. [19] reconstructed curved objects from wireframes; Han et al. [10] applied Bayesian reconstruction for polyhedral objects, trees, or grass. Black et al. [6, 28] estimated detailed human body shapes using parametric morphable models. In [28], a discriminative+generative method was proposed to help initialise the body parameters for reconstruction. In [6], shading cues are incorporated for single view reconstruction. Although they showed detailed shape recovery, it does not seem easy, in general, to solve the regression problem of the huge parametric space of joint angles, and to extend the approach to model other object categories. Chen et al. [3] tackled more general deformable object categories. The shape and pose generators need only a small number of latent variables to estimate, yet are able to capture complex 3D object shapes.

One close work to ours is [23], where an unified method to segment, infer 3D shapes and recognise object categories is proposed. They used a crude voxel representation for the shape prior and apply it to object categories such as cups, mugs, plates, etc. However, they are limited to simple and rigid objects. In [15, 34], 3D geometrical models are learnt to detect objects in images, but similarly, no articulation or deformation is considered.

The following branches of studies have conceptual differences from our work. Studies for human gait recognition [18] perform human identification from video sequences (instead of images) in a fixed camera view-point. Image-based face recognition across pose is an intensively studied area [20, 36]. Representative methods exploit active face shape models for view-point invariance [36] or expression invariance [21], however, these models are specifically designed for faces, involving many control points manually defined. Studies for 3D object retrieval are quite different, as they match one 3D model with another.

2. Phenotype Recognition and Shape Reconstruction Based on Classifiers

In the paper, the phenotype recognition problem is formulated as follows. We assume that a set of 2D phenotype galleries $\mathcal{G} = \{\mathbf{S}_\mathbf{c}^\mathbf{G}\}_{c=1}^{N_c}$ of N_c instances, which contains one sample silhouette $\mathbf{S}_\mathbf{c}^\mathbf{G}$ for each phenotype class c (see Fig. 1 for examples), is provided as the reference, and all the silhouettes in $\mathcal{G}$ are in a common canonical pose. We hope to find the phenotype label $c^* \in \{1, 2, \cdots, N_c\}$ for a query silhouette $\mathbf{S}^\mathbf{q}$ in an arbitrary pose and camera viewpoint.

To handle the difficulties caused by poses and camera view changes, our approach learns 3D shape priors $\mathcal{M}$ on available 3D data. Gaussian Process latent variable models (GPLVMs) [14] have been shown powerful in pose estimation and shape modeling [3, 17, 35]. We implement the framework in [3], in which two GPLVMs, the shape generator $\mathcal{M}_\mathcal{S}$ and the pose generator $\mathcal{M}_\mathcal{A}$, are learned to sepa-

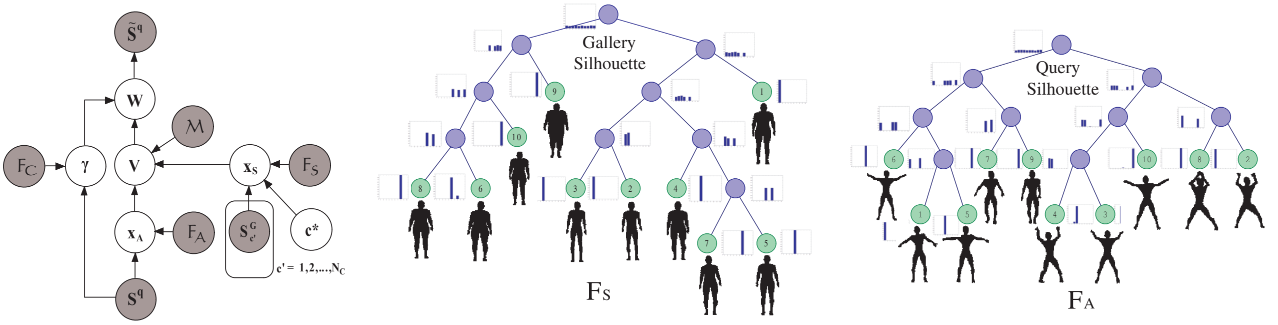


Figure 2. The graphical model for the 3D shape recognition and reconstruction problem (left). Example trees of Random Forests for the phenotype $\mathcal{F}_\mathcal{S}$ (middle) and pose $\mathcal{F}_\mathcal{A}$ (right): they show the class histogram at each split node and the phenotype/pose class at each leaf node. Note that the trees shown here are grown tiny for the visualisation purpose.

rately capture 3D shape and pose variations and jointly used to synthesize new 3D shapes. Each 3D shape $\mathbf{V}$ is then parametrized by a phenotype latent variable $\mathbf{x_S}$ and a pose latent variable $\mathbf{x_A}$, as Fig. 2(left) shows.

Given a silhouette image $\mathbf{S^q}$, we infer its embedding pose latent parameters $\mathbf{x_A}$, and the camera parameters γ so that we can neutralise the influence of pose and camera viewpoint changes in the recognition. Inferring these parameters can be done through the optimisation process of the generative model in [3]. However, the back-projection from 2D to 3D is usually multi-modal, and this results in a non-convex objective function with multiple local optima, which is usually difficult to solve. To avoid this non-convex optimisation, some previous studies have tried combining generative and discriminative approaches [25, 28]. We here propose an approach for fast hypothesizing shape (phenotype), pose, and camera parameters based on random forest (RF) classifiers, which are shown to have exceptional performance in solving multi-modal mapping problems [22, 27]. In our approach, three RFs $\{\mathcal{F}_\mathcal{S}, \mathcal{F}_\mathcal{A}, \mathcal{F}_\mathcal{C}\}$, are learned on a large number of silhouettes synthesized by $\mathcal{M}_\mathcal{S}$ and $\mathcal{M}_\mathcal{A}$ with different camera parameters γ (see Section 3 for details of learning these RFs). $\mathcal{F}_\mathcal{S}$ predicts the shape parameter $\mathbf{x_S}$ from a gallery silhouette $\mathbf{S_c^G}$, while $\mathcal{F}_\mathcal{A}$ and $\mathcal{F}_\mathcal{C}$ predict the pose and camera parameters $\{\mathbf{x_A}, \gamma\}$ from the query silhouette $\mathbf{S^q}$. $\mathbf{S^q}$ or $\mathbf{S_c^G}$ is passed down each tree in the forest, and the leaf nodes of these trees quickly provide multiple candidates of its corresponding shape, pose or camera parameters (see Fig. 2 for examples).

Finally, the 3D shape $\mathbf{V}$ of the query $\mathbf{S^q}$ is recovered by the estimated pose latent values $\mathbf{x_A}$ of $\mathbf{S^q}$ and the shape latent values $\mathbf{x_S}$ of each gallery instance $\mathbf{S_c^G}$ (see Section 2.2), and the recognition is achieved by a model-selection, i.e., assigning the phenotype class c^* that yields the best matching between the query $\mathbf{S^q}$ and the projection of the reconstructed shape $\mathbf{V}$ in camera viewpoint γ (see Section 2.1).

2.1. Phenotype Recognition

Phenotype recognition is formulated as a model-selection problem. Based on the graphical model in Fig. 2(left), we infer the label c^* of the query instance by maximizing a posteriori probability given pre-learned shape priors $\mathcal{M} = \{\mathcal{M}_\mathcal{S}, \mathcal{M}_\mathcal{A}\}$ and classifiers $\mathcal{F} = \{\mathcal{F}_\mathcal{S}, \mathcal{F}_\mathcal{A}, \mathcal{F}_\mathcal{C}\}$ as

$$\begin{aligned}
&P(c^*|\mathbf{S^q}, \tilde{\mathbf{S}}^\mathbf{q}, \{\mathbf{S_c^G}\}_{c=1}^{N_c}, \mathcal{M}, \mathcal{F}) \\
\propto &P(\tilde{\mathbf{S}}^\mathbf{q}|c^*, \mathbf{S^q}, \{\mathbf{S_c^G}\}_{c=1}^{N_c}, \mathcal{M}, \mathcal{F})P(c^*) \\
= &P(c^*)\int_{\mathbf{x_S},\mathbf{x_A},\gamma} P(\mathbf{x_S}|\mathbf{S_{c^*}^G}, \mathcal{F}_\mathcal{S})P(\mathbf{x_A}|\mathbf{S^q}, \mathcal{F}_\mathcal{A}) \\
&P(\gamma|\mathbf{S^q}, \mathcal{F}_\mathcal{C})P(\tilde{\mathbf{S}}^\mathbf{q}|\mathbf{x_S}, \mathbf{x_A}, \gamma, \mathcal{M})\mathbf{dx_S dx_A d}\gamma, \qquad (1)
\end{aligned}$$

where $\tilde{\mathbf{S}}^\mathbf{q}$ denotes the mirror node of $\mathbf{S^q}$. Here, we assume that the class prior $P(c^*)$ is subject to a uniform distribution, i.e., $P(c^*) = 1/N_c$.

In (1), the first three terms describe the prior of shape and pose latent parameters $(\mathbf{x_S}, \mathbf{x_A})$ and camera parameters γ from the random forest classifiers $\mathcal{F}_\mathcal{S}$, $\mathcal{F}_\mathcal{A}$, and $\mathcal{F}_\mathcal{C}$, respectively. The shape classifier $\mathcal{F}_\mathcal{S}$ predicts N_S candidate phenotype shapes $\{\mathbf{x_{S,i}^{c^*}}\}_{i=1}^{N_S}$ for the canonical posed gallery silhouette $\mathbf{S_{c^*}^G}$ of each class c^*; while the pose classifier $\mathcal{F}_\mathcal{A}$ and the camera viewpoint classifier $\mathcal{F}_\mathcal{C}$ predict N_A candidate poses $\{\mathbf{x_{A,j}}\}_{j=1}^{N_A}$ and N_K candidate camera parameters $\{\gamma_\mathbf{k}\}_{k=1}^{N_K}$ for the query silhouette input $\mathbf{S^q}$. Mathematically, these three terms can be written as delta impulses.

$$P(\mathbf{x_S}|\mathbf{S_{c^*}^G}, \mathcal{F}_\mathcal{S}) = \sum_{i=1}^{N_S} h_{S,i}^{c^*}\delta(\mathbf{x_S} - \mathbf{x_{S,i}^{c^*}}), \qquad (2)$$

$$P(\mathbf{x_A}|\mathbf{S^q}, \mathcal{F}_\mathcal{A}) = \sum_{j=1}^{N_A} h_{A,j}\delta(\mathbf{x_A} - \mathbf{x_{A,j}}), \qquad (3)$$

$$P(\gamma|\mathbf{S^q}, \mathcal{F}_\mathcal{C}) = \sum_{k=1}^{N_K} h_{C,k}\delta(\gamma - \gamma_\mathbf{k}), \qquad (4)$$

where $h_{S,i}^{c^*}$, $h_{A,j}$, and $h_{C,k}$ are class histogram values voted by every tree in $\mathcal{F}_\mathcal{S}$, $\mathcal{F}_\mathcal{A}$, and $\mathcal{F}_\mathcal{C}$, respectively, and they satisfy $\sum_{i=1}^{N_S} h_{S,i} = \sum_{j=1}^{N_A} h_{A,j} = \sum_{k=1}^{N_K} h_{C,k} = 1$.[1]

[1] For the purpose of robustness and acceleration, we discard all the small-weighted candidates under the thresholds $h_{S,i}^{c^*} < 0.05$, $h_{A,j} < 0.05$, and $h_{C,k} < 0.05$ in the experiments.

In the last term of the model, each combination of shape and pose latent parameters $(\mathbf{x_S}, \mathbf{x_A})$, and the camera pose γ are verified by the silhouette likelihood of the query image $\tilde{\mathbf{S}}^{\mathbf{q}}$. It can be formulated as the following equation:

$$P(\tilde{\mathbf{S}}^{\mathbf{q}}|\mathbf{x_S}, \mathbf{x_A}, \gamma, \mathcal{M}) \approx \frac{1}{Z_S\sqrt{\det\left(\mathbf{I} + \frac{1}{\sigma_s^2}\mathbf{\Sigma_W}\right)}} e^{-\mathbf{OCM}\left(\mu_\mathbf{W}, \tilde{\mathbf{S}}^{\mathbf{q}}\right)/2\sigma_s^2}, \tag{5}$$

where $\mathbf{W}$ is referred to the projected silhouette of the latent 3D shape $\mathbf{V}$ in the camera viewpoint γ; $\mu_\mathbf{W}$ and $\mathbf{\Sigma_W}$ refer to the mean and the covariance matrix of $\mathbf{W}$, respectively (refer to [3] for detail formulations); σ_s^2 and Z_S are normalisation factors. We use oriented Chamfer matching (OCM) distance [31] to measure the similarity between the mean projected silhouette $\mu_\mathbf{W}$ and the silhouette of the query image $\tilde{\mathbf{S}}^{\mathbf{q}}$. Detailed formulations of OCM are described in Section 4.2. Given all the probability terms, the final posterior probability in (1) can be computed as:

$$P(c^*|\mathbf{S^q}, \tilde{\mathbf{S}}^{\mathbf{q}}, \{\mathbf{S_c^G}\}_{c=1}^{N_c}, \mathcal{M}, \mathcal{F}) \approx \frac{1}{N_C}\sum_{i=1}^{N_S}\sum_{j=1}^{N_A}\sum_{k=1}^{N_K} h_{S,i}^{c^*} h_{A,j} h_{C,k} P(\tilde{\mathbf{S}}^{\mathbf{q}}|\mathbf{x}_{\mathbf{S,i}}^{\mathbf{c^*}}, \mathbf{x_{A,j}}, \gamma_\mathbf{k}, \mathcal{M}). \tag{6}$$

2.2. Single View 3D Shape Reconstruction

As a by-product, our framework can also be used to quickly predict an approximate 3D shape $\mathbf{V}$ from the query silhouette $\mathbf{S^q}$. This shape reconstruction problem can be formulated probabilistically as follows:

$$\begin{aligned} & P(\mathbf{V}|\mathbf{S^q}, \{\mathbf{S_c^G}\}_{c=1}^{N_c}, \mathcal{M}, \mathcal{F}) \\ = & \sum_{c=1}^{N_c}\left[\int_{\mathbf{x_S},\mathbf{x_A}} P(\mathbf{V}, \mathbf{x_S}, \mathbf{x_A}, c|\mathbf{S^q}, \mathbf{S_c^G}, \mathcal{M}, \mathcal{F})\mathbf{dx_S dx_A}\right] \\ = & \sum_{c=1}^{N_c} P(c)\left[\int_{\mathbf{x_S},\mathbf{x_A}} P(\mathbf{x_S}|\mathbf{S_c^G}, \mathcal{F}_S)P(\mathbf{x_A}|\mathbf{S^q}, \mathcal{F}_\mathcal{A})\right. \\ & \left. P(\mathbf{V}|\mathbf{x_S}, \mathbf{x_A}, \mathcal{M})\mathbf{dx_S dx_A}\right] \\ \approx & \frac{1}{N_C}\sum_{c=1}^{N_c}\sum_{i=1}^{N_S}\sum_{j=1}^{N_A} h_{S,i}^{c^*} h_{A,j} N\left(\mathbf{V}|\mu_\mathbf{V}, \mathbf{\Sigma_V}\right). \end{aligned} \tag{7}$$

where $\mu_\mathbf{V} = \mu_\mathbf{V}(\mathbf{x_{A,j}}, \mathbf{x_{S,i}^c})$ and $\mathbf{\Sigma_V} = \mathbf{\Sigma_V}(\mathbf{x_{A,j}}, \mathbf{x_{S,i}^c})$ are referred to the mean and variance function of the 3D shape distribution $\mathbf{V}$, respectively, and their detailed formulations can be found in [3]. Compared with the optimisation approach in [3], the classifiers-based approach in the paper provides fairly good qualitative results and is much more efficient in computation time (See Section 4.4).

3. Training Random Forest Classifiers

In order to learn the random forest classifiers $\mathcal{F} = \{\mathcal{F}_S, \mathcal{F}_\mathcal{A}, \mathcal{F}_\mathcal{C}\}$, we use the shape and pose generators $\{\mathcal{M}_S, \mathcal{M}_\mathcal{A}\}$ to synthesize a large number of silhouettes with different latent parameters $\{\mathbf{x_S}, \mathbf{x_A}\}$ and camera viewpoints γ.

The shape classifier $\mathcal{F}_S$, an ensemble of randomised decision trees, is used to encode the phenotype information of each gallery silhouette $\mathbf{S_c^G}$ in the canonical pose. It is trained on a dataset $\mathcal{D}_1$ consisting of canonical-posed silhouettes of $N = 50$ phenotype samples $\{\mathbf{x_{S,i}}\}_{i=1}^N$ which are uniformly sampled from the latent space of the shape generator $\mathcal{M}_S$. For each sample of phenotype label $i \in \{1, 2, \cdots, N\}$, we generate $R = 250$ sample silhouettes from the 3D mesh model with minor pose perturbations and camera parameter changes, e.g., slight camera rotations and focal length changes. All $N \times R = 12500$ binary images are aligned and normalised to have the same size.

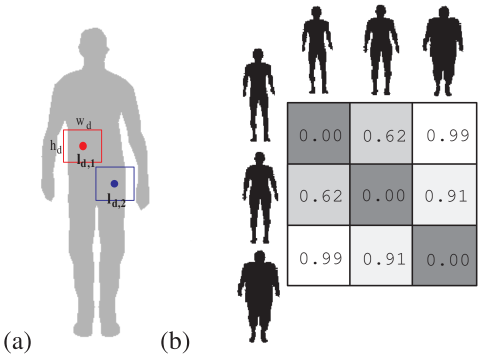
On the other hand, the pose classifier $\mathcal{F}_\mathcal{A}$ and the camera classifier $\mathcal{F}_\mathcal{C}$ are used to predict the pose and camera viewpoint of the query silhouette $\mathbf{S^q}$. We train them on another dataset $\mathcal{D}_2$ with large pose and camera viewpoint variations as well as phenotype variations. We uniformly sample $M = 50$ pose samples $\{\mathbf{x_{A,j}}\}_{j=1}^M$ from the latent space of the pose generator $\mathcal{M}_\mathcal{A}$, and $K = 50$ camera viewpoint samples $\{\gamma_\mathbf{k}\}_{k=1}^K$ uniformly distributed in the 3D space, and generate 3D shapes along with the same $N = 50$ phenotype samples $\{\mathbf{x_{S,i}}\}_{i=1}^N$ used in the shape classifier training stage. This generates $N \times M \times K = 125,000$ silhouette instances, and each of them is labeled by (i, j, k) representing phenotype, pose and camera viewpoint, respectively. An ensemble of decision trees for $\mathcal{F}_\mathcal{A}$ and $\mathcal{F}_\mathcal{C}$ are grown by the pose label j and camera label k, respectively[2]. See below for the random features and split criteria used.

3.1. Error Tolerant Features

We generate $D = 12000$ random rectangle pairs $\{(R_{d,1}, R_{d,2})\}_{d=1}^D$ with arbitrary centroid locations $(\mathbf{l_{d,1}}, \mathbf{l_{d,2}})$, heights h_d, and widths w_d (see Fig 3(a) for example). For each binary silhouette image I for training, the difference of mean image intensity values within each pair of rectangles is then computed as the split feature $f_d = \frac{1}{w_d h_d}\left(\sum_{\mathbf{l}\in R_{d,1}} I(\mathbf{l}) - \sum_{\mathbf{l'}\in R_{d,2}} I(\mathbf{l'})\right)$. In this way, each training instance I is hence converted to a 12000-D feature vector $\mathbf{f} = [f_d]_{d=1}^D$. These features are efficient to compute and capture spatial context [26].

When training the phenotype classifier $\mathcal{F}_S$, we also introduce a feature-correction scheme. Since $\mathcal{F}_S$ is trained on synthetic silhouettes generated by the shape priors $\mathcal{M}$, which are clean and unclothed, its discriminative power is usually reduced when working on noisy gallery silhouettes segmented from real images. To model the systematic errors between the synthetic and real silhouettes, we use the approach in [3] to create an extra silhouette set which con-

[2] In our implementation, we set the tree number N_T of all the forests $\mathcal{F}_S$, $\mathcal{F}_\mathcal{A}$, and $\mathcal{F}_\mathcal{C}$ to be 30, and the maximum depth d_{max} of a single tree to be 30.



0.00	0.62	0.99
0.62	0.00	0.91
0.99	0.91	0.00

Figure 3. (a) Random paired-rectangle features. (b) A 3×3 example of dissimilarity matrix $\mathbf{\Pi}$ for human phenotype classes.

sists of N_e pairs of real and synthetic silhouettes describing different clothing and segmentation errors and capturing different phenotypes. We then extract the features from images using the same set of random rectangle pairs. Here, $\tilde{\mathbf{f}}^{\mathbf{e}}_{\mathbf{m}}$ $(m = 1, 2, \cdots, N_e)$ denotes the features extracted from the real silhouette images, and $\mathbf{f}^{\mathbf{e}}_{\mathbf{m}}$ denotes those from the corresponding synthetic silhouette images. The feature errors can be thus modeled by $\mathbf{e_m} = \tilde{\mathbf{f}}^{\mathbf{e}}_{\mathbf{m}} - \mathbf{f}^{\mathbf{e}}_{\mathbf{m}}$.

To compensate for the systematic silhouette errors when training $\mathcal{F_S}$, we correct all those synthetic training data with these error vectors $\{\mathbf{e_m}\}_{m=1}^{N_e}$. For each feature vector $\mathbf{f}$ of instance I, we find its T nearest neighbor synthetic features in $\mathcal{E}$ (we choose $T = 3$), and use the corresponding error vectors $\mathbf{e_t}$ to correct $\mathbf{f}$ as $\tilde{\mathbf{f}}_{\mathbf{t}} = \mathbf{e_t} + \mathbf{f}$, $(t = 1, 2, \cdots, T)$. Finally, all $N \times R \times T$ corrected features vectors $\tilde{\mathbf{f}}_{\mathbf{t}}$ of $N \times R$ training instances are used as training samples for $\mathcal{F_S}$.

3.2. Similarity-Aware Criteria Functions

When training a random forest classifier $\mathcal{F}^* \in \{\mathcal{F_S}, \mathcal{F_A}, \mathcal{F_C}\}$, the instance I is pushed through each tree in $\mathcal{F}^*$ starting from the root node. At each split node, the error-corrected random feature $\tilde{\mathbf{f}} = [\tilde{f}_d]_{d=1}^{D}$ is evaluated for every single training instance (see Section 3.1). Then, based on the result of the binary test $\tilde{f}_d > \tau_{th}$, I is sent to the left or the right child node. The feature dimension index d and the split value τ_{th} at a split node n are chosen to maximize $\Delta C(n) = C(n) - \frac{|n_L|C(n_L)+|n_R|C(n_R)}{|n_L|+|n_R|}$, where C measures the distribution purity of the node, and n_L and n_R denote the left and right children of node n. For the criteria function C, we generalise Gibbs and Martin's diversity index [7] and take the class similarity into account:

$$C(n) = \mathbf{p}_n^T \mathbf{\Pi} \mathbf{p}_n, \tag{8}$$

where $\mathbf{p}_n = [p_{n,1}, p_{n,2}, \cdots, p_{n,N_B}]$ is referred to the class distribution of node n; N_B denotes the number of class labels of the random forest $\mathcal{F}^*$; the weighting matrix $\mathbf{\Pi} = \{\pi_{ij} = 1 - e^{-\|\mathbf{\Delta V_{i,j}}\|^2/\sigma^2}\}_{N_B \times N_B}$, which is defined by the average spatial mesh distance $\|\mathbf{\Delta V_{i,j}}\|^2$ between classes i and j (see Fig. 3(b) for an example of phenotype classes). When $\mathbf{\Pi} = \mathbf{1} - \mathbf{I}$, equation (8) is reduced to the standard diversity index $C(n) = 1 - \sum_{c=1}^{N_B} p_{n,c}^2$. Intuitively, a lower misclassification penalty is assigned between two visually similar classes in (8). The experiment shows that such a similarity weighting scheme notably improves the recognition rate (see Section 4.3).

4. Experimental Results

4.1. Datasets

We have verified the efficacy of our approach on two shape categories: humans and sharks. For the human data, we train the shape model $\mathcal{M_S}$ on the CAESAR database, which contains over 2000 different body shapes of North American and European adults in a common standing pose, and train the pose model $\mathcal{M_A}$ on 42 walking and jumping-jack sequences in CMU Mocap dataset. For the shark data, we learn $\mathcal{M_S}$ on a shape data set that contains eleven 3D shark models of different shark species available from the Internet, and $\mathcal{M_A}$ on an animatable 3D MEX shark model to generate an 11-frame sequence of shark tail-waving motion [3]. The mesh resolutions are: 3678 vertices/7356 faces for the human data, and 1840 vertices/3676 faces for shark data, respectively. We empirically set the latent space dimension of the shape model $\mathcal{M_S}$ to be 6 for human data and 3 for shark data, while for the pose model $\mathcal{M_A}$, we set the latent dimension to be 2 for both, similarly to [3].

As there is no suitable public datasets to evaluate the proposed approach, we have collected two new silhouette datasets which capture a wide span of phenotype, pose, and camera viewpoint changes(see Fig. 4 for examples). **Human motion dataset** mainly captures two different human motions: walking (184 images of 16 human instances) and jumping-jack motion (170 images of 13 human instances). The images are cropped from video sequences on YouTube and public available human motion datasets, e.g., HumanEva [29]. For each instance, a canonical standing pose image is provided (see Fig. 1 and Row 1, 2 of Fig. 4). All the instances are in tightly-fitting clothing. **Shark motion dataset** includes 168 images of 13 shark instances of 5 sub-species. These images are cropped from underwater swimming sequences downloaded from Internet. For each instance, a profile-view image is provided as the canonical-pose gallery image.

The silhouettes are manually segmented from the images and all of them are normalised by their height and resized to the resolution 121×155. For both datasets, $N_e = 20$ additional images are collected for modeling the feature errors (in Section 3.1).

4.2. Comparative Methods

For the purpose of comparison, we also implemented three state-of-the-art methods based on 2D shape matching: 1) Shape contexts (SC) [1], 2) Inner-Distance Shape Context (IDSC) [16], and 3) the oriented chamfer matching

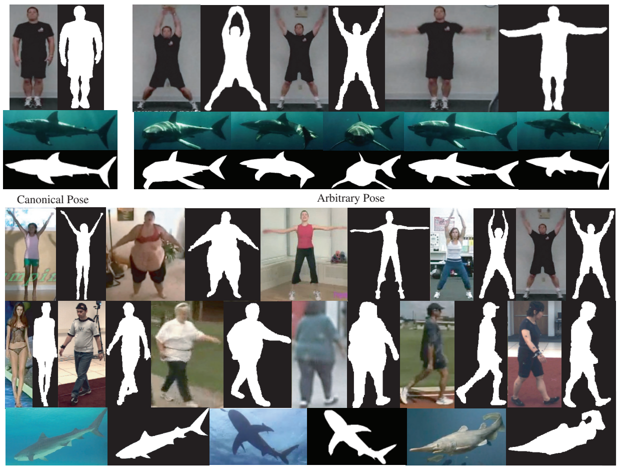


Figure 4. Examples of images and their corresponding silhouettes in our phenotype-class recognition datasets. Row 1,2: dataset structure: one canonical-posed instance and several arbitrary-posed instances; Row 3: human jumping-jack motion; Row 4: human walking; Row 5: shark underwater swimming motion.

(OCM) [31], and two methods using the 3D shape priors: 4) the single-view 3D shape reconstruction method by Mixture of Experts [28] and 5) the RF implementation directly using the shape class labels. Nearest Neighbor classification is performed in terms of the similarity provided by the compared methods.

Histogram of Shape Context (HoSC). Shape contexts (SC) are rich local shape-based histograms encoding contour information and they have been widely used for shape matching and pose recognition. Since SCs are defined locally on every single silhouette point, representing the whole shape can be expensive. To reduce the dimensionality of shape contexts, Agarwal and Triggs introduce a bag-of-features scheme called histogram of shape context (HoSC) [1] for human pose estimation. In HoSC, k-means clustering is used to yield a L-dimensional codebook of the cluster means ($L = 100$ in the paper), and all its shape contexts are then softly binned to a quantized L-dimensional histograms. We implemented a 2D approach HoSC-χ^2, which compares the χ^2-distances of HoSC features extracted from the query and each gallery silhouette.

Inner-Distance Shape Context (IDSC). Recent research on shape matching has addressed the problem of finding articulation invariant distance measurement for 2D shapes. Among them, a representative recent work is Inner-Distance Shape Context (IDSC) by Ling and Jacobs [16], which has been proved successful in 2D shape classification problems. The authors' own code is used.

2D Oriented Chamfer matching (OCM). Chamfer matching and its variants have been widely used for shape matching and pose recognition.Among them, oriented Chamfer matching has been proved to be an effective method for shape-based template matching [31]. The query silhouette $\mathbf{S^q} = \{\mathbf{s^q_k}\}_{k=1}^{N_q}$ and gallery silhouettes $\mathbf{S^G_c} = \{\mathbf{s^G_{c,j}}\}_{j=1}^{N_c^G}$ ($c = 1, 2, \cdots, N_c$), where $\mathbf{s^q_k}$ and $\mathbf{s^G_{c,j}}$ denote edge points, are divided into N_{ch} orientation channels: $\{\mathbf{S^q_t}\}_{t=1}^{N_{ch}}$ and $\{\mathbf{S^G_{c,t}}\}_{t=1}^{N_{ch}}$, respectively. In our implementation, we set $N_{ch} = 8$. To minimise the allocation error of image edges in orientation, an edge point $\mathbf{s^G_{c,j}}$ is assigned to both adjacent channels when its orientation is around the border region. The OCM distance between $\mathbf{S^G_c}$ and $\mathbf{S^q}$ is calculated as the sum of independent chamfer distance with each independent orientation channel, as the following equation shows:

$$OCM(\mathbf{S^G_c}, \mathbf{S^q}) = \frac{1}{N_q} \sum_{t=1}^{N_{ch}} \sum_{\mathbf{s^G_{c,j}} \in \mathbf{S^G_{c,t}}} \min_{\mathbf{s^q_k} \in \mathbf{S^q_t}} \|\mathbf{s^q_k} - \mathbf{s^G_{c,j}}\|^2, \quad (9)$$

Mixture of Experts for the shape reconstruction. We implemented a 3D shape recognition approach, called HoSC-MoE-Chamfer, based on the shape reconstruction framework proposed in [28], in which mappings from HoSC features to shape and pose parameters are learned using a Mixture of Experts (MoE) model. Weighted linear regressors are used as mixture components. For a fair comparison, the same training sets $\mathcal{D}_1$ and $\mathcal{D}_2$ and shape priors $\mathcal{M}$ are used, and the recognition is also based on the OCM distance between the predicted shape and the query silhouette.

Single Random Forest Shape Verification. We also compare our framework with a straightforward classification approach based on a single shape random forest, in which $\mathcal{F}_\mathcal{S}$ is directly learned on the large pose and camera viewpoint variation dataset $\mathcal{D}_2$ according to the phenotype label i (see Section 3). For an arbitrary input silhouette, the phenotype prediction from the forest $\mathcal{F}_\mathcal{S}$ is given by a histogram which summaries the phenotype vote from each tree. The phenotype similarity between the query silhouette and an gallery silhouette can be measured by the χ^2-distance between their random forest prediction histograms.

4.3. Numerical Results of Phenotype Recognition

We perform cross validations by randomly selecting 5 different instances, where we use their canonical posed images as the galleries and any other poses as the query. The results of the proposed approach (G+D) and its variants are reported to show the effect of components and internal parameters. To evaluate the benefit of using the feature correction (Section 3.1) and similarity-based criteria function (Section 3.2), we present the results of our approach without error modeling (G+D-E) and using standard diversity index [7] as the criteria function (G+D-S) in Fig. 5(a). It shows that both schemes help improve the recognition performance of our approach to some extent in all three datasets. We also investigate how the maximum tree depth d_{max} and the tree number N_T of random forests $\mathcal{F}_\mathcal{S}$, $\mathcal{F}_\mathcal{A}$, and $\mathcal{F}_\mathcal{C}$ affect the performance. As shown in Fig. 5(b) and 5(c), the accuracy does not vary much at over 25 depths, but increasing the number of trees of each forest gradually

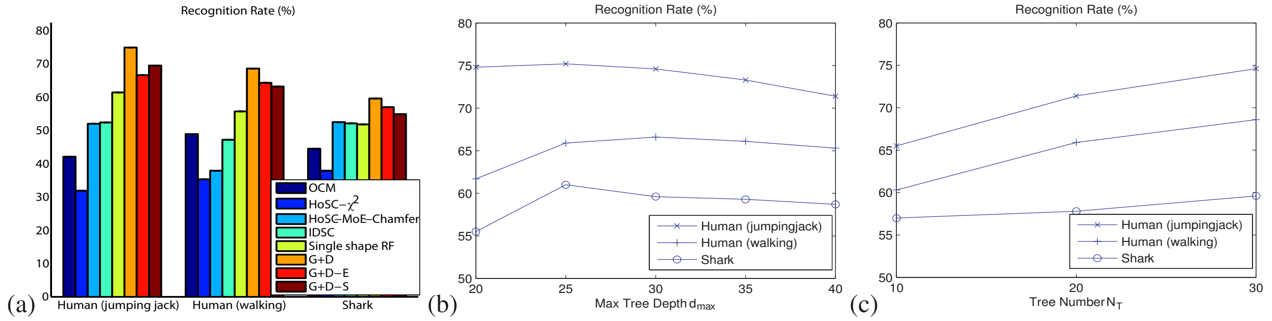


Figure 5. Phenotype recognition accuracy on human and shark datasets. (a) Comparison over 8 different approaches; (b) performance under different maximum tree depths d_{max}; and (c) different tree numbers N_T of the random forests $\mathcal{F}_\mathcal{S}$, $\mathcal{F}_\mathcal{A}$, and $\mathcal{F}_\mathcal{C}$.

improves the recognition rate.

Fig. 5(a) provides the recognition rates of different approaches. In general, the 3D-based approaches (single RF, HoSC-MoE-Chamfer and the proposed method G+D) outperform those 2D-based ones (OCM, HoSC-χ^2, and IDSC) in the phenotype recognition tasks. The best 2D shape measurement IDSC achieves a close performance to that of 3D approaches. This indicates the benefit of using 3D shape priors to handle pose deformations and camera viewpoint changes. On the other hand, given the same training data, our approach (G+D) performs best among three 3D approaches under all contexts. Compared to the single shape RF, our framework that factorizes three types of variations in the training stage, better captures subtle shape variations. In most cases, object pose and camera viewpoint changes are more dominant factors that affect the silhouette appearance than phenotype variations, and hence they greatly distract the discriminative power of the single RF which is directly learnt on the mixed variation data set $\mathcal{D}_2$ with the shape labels. Instead, we learn the phenotype classifier $\mathcal{F}_\mathcal{S}$ on a canonical-posed dataset $\mathcal{D}_1$, which does not include large pose and camera viewpoint changes. For the pose and camera classifiers, we use the the mixed variation data set $\mathcal{D}_2$ but with the pose and camera labels respectively. The pose and camera parameters are much more reliably estimated than the shape parameter for given the same training data. The comparison between our approach and HoSC-MoE-Chamfer shows that given the same training data, the random forests and rectangle features we used also outperform the combination of MoE and HoSC features in the setting of phenotype discrimination. This could partially be owing to the feature selection process during the RF training stage and the scheme of generating multiple hypotheses for a single input in the RF prediction stage.

4.4. Approximate Single View Reconstruction

In our framework, these intermediate 3D shape candidates $\mathbf{V}$ obtained during the recognition process can be used for approximate 3D reconstruction from a single silhouette input, as mentioned in Section 2.2. In Fig. 6, we show some qualitative 3D outputs of different phenotypes using our framework in contrast with those generated using the approach in [3]. In general, these highest-weight shape candidates generated by random forest classifiers often include meaningful shapes which can be used as fairly good approximate reconstruction results, albeit relatively lower silhouette coherency and less accurate pose estimation. However, we also notice that some results may still be in wrong phenotype (e.g., instance 5) or in a wrong pose or camera viewpoint (e.g., instance 9). This is mainly due to the silhouette ambiguity or a limitation on the discriminative power of random forest classifiers given our training set. We also compute the running time of both approaches under a 2.8GHz CPU. The average time for generating a 3D shape using our new generative+discriminative framework is less than 10 seconds using unoptimised Matlab codes, while using the approach in [3] takes about 10 to 15 minutes for generating 10 candidates. This improvement in computational efficiency owes much to using RFs for hypothesizing $\mathbf{x_S}$, $\mathbf{x_A}$, and γ, which greatly narrows down the search space of the algorithm.

5. Conclusions

The paper presents a probabilistic framework which combines both generative and discriminative cues for recognizing the phenotype class of an object from a single silhouette input and reconstructing its approximate 3D shape. We learn 3D probabilistic shape priors of the object category by GPLVM to handle the difficulties in the camera viewpoint changes and pose deformation, and use random forests for efficient inference of phenotype, pose, and camera parameters. Experiments on human and shark silhouettes have shown the advantage of our approach against both standard 2D-based methods and relevant 3D-based methods.

The present accuracy on the datasets we provide, especially on the shark dataset, is limited due to the descriptive power of the shape and pose generators we used to synthesize silhouettes and insufficient number of 3D shapes and motion data used for training. Using more extensive 3D training data would improve the accuracy. Another major problem which limits the application of the current framework is in the requirement of silhouette segmentation. This

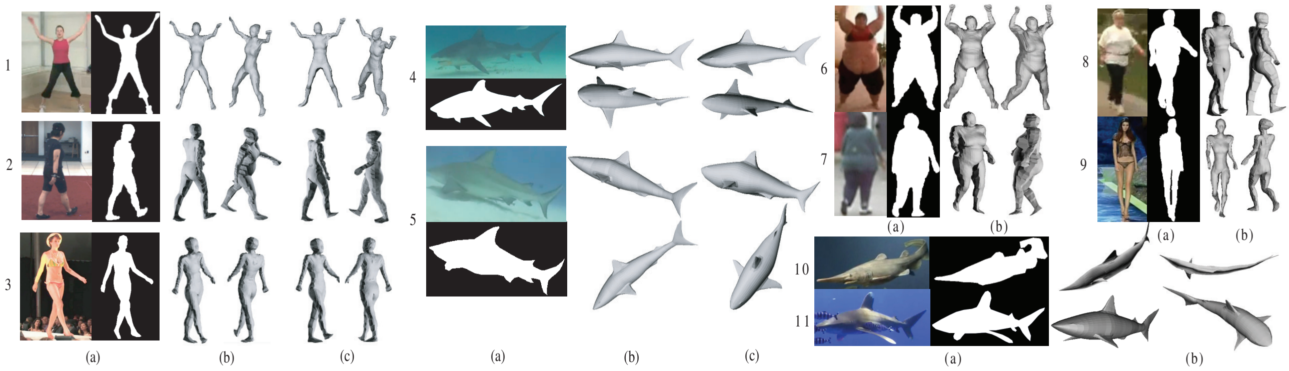


Figure 6. Approximate single view reconstruction using the shape candidates from the random forest classifiers. (a) Input query images and silhouettes; (b) the highest-weight 3D shape candidates from $\mathcal{F}_S$ and $\mathcal{F}_\mathcal{A}$ for each query silhouette (in two different views); (c) results generated by the approach in [3] (in two different views).

could be helped by e.g. Kinect camera which yields reliable foreground-background segmentation in real time. Also, as our future work, we plan to build up a larger-scale phenotype recognition dataset of different categories of objects and make it available to public. It would help evaluate our approach and do comparative studies.

References

[1] A. Agarwal, B. Triggs, Recovering 3D human pose from monucular images, PAMI 28 (1) (2006) 44–58

[2] M. Andriluka, S. Roth, B. Schiele, Monocular 3D Pose Estimation and Tracking by Detection, CVPR (2010).

[3] Y. Chen, T-K. Kim, R. Cipolla, Inferring 3D shapes and deformations from single views, ECCV (2010).

[4] C. Cyr, B. Kimia, 3D object recognition using shape similarity-based aspect graph, ICCV (2001).

[5] V. Ferrari, F. Jurie, C. Schmid, From images to shape models for object detection, IJCV 87 (3) (2010) 284–303.

[6] P. Guan, A. Weiss, A. Balan, M. Black, Estimating human shape and pose from a single image, ICCV (2009).

[7] J.P. Gibbs, W.T. Martin, Urbanization, technology and the division of labor, American Sociological Review 27 (1962) 66–77.

[8] R. Gopalan, P. Turaga, R. Chellappa, Articulation-Invariant Representation of Non-planar Shapes, ECCV (2010).

[9] C. Gu, X. Ren, Discriminative mixture-of-templates for viewpoint classification, ECCV (2010).

[10] F. Han, S. Zhu, Bayesian reconstruction of 3D shapes and scenes from a single image, HLK'03: Proc. the 1st IEEE Int. Workshop on Higher-Level Knowledge (2003).

[11] D. Hoiem, C. Rother, J. Winn, 3D Layout CRF for Multi-View Object Class Recognition and Segmentation, CVPR (2007).

[12] W. Hu, S. Zhu, Learning a Probabilistic Model Mixing 3D and 2D Primitives for View Invariant Object Recognition, CVPR (2010).

[13] S. Johnson, M. Everingham, Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation, BMVC (2010).

[14] N. Lawrence, Gaussian process latent variable models for visualisation of high dimensional data, NIPS 16 (2004) 329–336

[15] J. Liebelt, C. Schmid, Multi-View Object Class Detection with a 3D Geometric Model, CVPR (2010).

[16] H. Ling, D.W. Jacob, Shape classification using the inner-distance, PAMI, 29 (2) (2007) 286–299.

[17] R. Navaratnam, A. Fitzgibbon, R. Cipolla. Semi-supervised Joint Manifold Learning for Multi-valued Regression, ICCV (2007).

[18] M.S. Nixon, T.N. Tan, and R. Chellappa, Human Identification based on Gait. International Series on Biometrics, Springer (2005)

[19] M. Prasad, A. Fitzgibbon, A. Zisserman, L. Gool, Finding Nemo: Deformable Object Class Modelling using Curve Matching, CVPR (2010).

[20] S. Prince, J. Elder, J. Warrell, F. Felisberti, Tied factor analysis for face recognition across large pose differences, PAMI, 30 (6) (2008) 970–984.

[21] Z. Riaz, C. Mayer, M. Wimmer, B. Radig, Model Based Face Recognition Across Facial Expressions, Journal of Information and Communication Technology (2008).

[22] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, P.H.S. Torr, Randomized Trees for Human Pose Detection, CVPR (2008).

[23] D. Rother, G. Sapiro, Seeing 3D objects in a single 2D image, ICCV (2009).

[24] F. Rothganger, S. Lazebnik, C. Schmid, J. Ponce, 3D Object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints, IJCV, 66 (3) (2006) 231–259.

[25] M. Salzmann and R. Urtasun. Combining discriminative and generative methos for 3D deformable surface and articulated pose reconstruction. CVPR (2010).

[26] J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, IJCV, 81 (1) (2009) 2–23.

[27] J. Shotton, A. Fitzgibbon, M. Cook, A. Blake, Real-time human pose recognition in parts from single depth images, CVPR (2011).

[28] L. Sigal, A. Bălan, M. Black, Combined discriminative and generative articulated pose and non-rigid shape estimation, NIPS (2007).

[29] L. Sigal, A. Bălan, M. Black, HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion, IJCV, 87 (1) (2010) 4–27.

[30] P. Srinivasan, Q. Zhu, J. Shi, Many-to-one Contour Matching for Describing and Discriminating Object Shape, CVPR (2010).

[31] B. Stenger, A. Thayananthan, P.H.S. Torr, R. Cipolla, Model-Based Hand Tracking Using a Hierarchical Bayesian Filter, PAMI 28 (9) (2006) 1372–1384.

[32] M. Sun, H. Su, S. Savarese, L. Fei-Fei, A multi-view probabilistic model for 3D object classes, CVPR (2009).

[33] G.W. Taylor, L. Sigal, D.J. Fleet, G.E. Hinton, Dynamical Binary Latent Variable Models for 3D Human Pose Tracking, CVPR (2010).

[34] A. Toshev, A. Makadia, K. Daniilidis, Shape-based object recognition in videos using 3D synthetic object models, CVPR (2009).

[35] R. Urtasun, D. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. CVPR (2010).

[36] X. Zhang, Y. Gao, Face recognition across pose: A review, Pattern Recognition 42 (2009), 2876–2896.