Learning Shape Priors for Single View Reconstruction

Yu Chen and Roberto Cipolla
Department of Engineering, University of Cambridge
{yc301 and rc10001}@cam.ac.uk

Abstract

In this paper, we aim to reconstruct free-from 3D models from a single view by learning the prior knowledge of a specific class of objects. Instead of heuristically proposing specific regularities and defining parametric models as previous research, our shape prior is learned directly from existing 3D models under a framework based on the Gaussian Process Latent Variable Model (GPLVM). The major contributions of the paper include: 1) a probabilistic framework for prior-based reconstruction we propose, which requires no heuristic of the object, and can be easily generalized to handle various categories of 3D objects, and 2) an attempt at automatic reconstruction of more complex 3D shapes, like human bodies, from 2D silhouettes only. Qualitative and quantitative experimental results on both synthetic and real data demonstrate the efficacy of our new approach.

1. Introduction

Reconstructing 3D shapes from 2D images is a popular topic in computer vision. Pure geometrical methods are the main streams of current research on single view reconstruction (SVR). Since SVR is a severely under-constrained problem, available geometrical clues, such as silhouettes, depth maps and normal maps, are usually far from enough to obtain an unambiguous reconstruction of the model. In view of this problem, previous research proposes additional geometrical or topological priors which are inherent in all 3D objects or in just a specific class of objects, e.g., minimizing overall smoothness [16], to further constrain the problem. A common major drawback of these methods is that those priors, however, are mainly defined by heuristics and suitable only for particular scenes, e.g., planar scenes, ground-vertical scenes, or under the orthogonal view. Such limitation prevents these methods from reconstructing those models with more complex and curved geometry like human faces or human bodies.

In this paper, we make an attempt to incorporate learning techniques into the 3D reconstruction problem. Instead of proposing specific reconstruction rules from heuristics, we learn the prior shape knowledge from existing 3D models. We observe that the shapes of the same class of objects are actually controlled by a small number of factors notwithstanding the complex geometrical or topological structures. Through manipulating these factors, we can then model the 3D shape of the whole class of objects with much fewer values and the detailed reconstruction from the 2D image can be estimated with much less difficulty.

Previous research manually define parametric models for each specific category, such as 3D faces [3] and human bodies [1], [20], to characterize the variation or movement of the 3D shape. The model parameters can thus be learned through training on the dataset. The main drawback of these approaches is that parametric models are only suitable for describing the shape of limited categories of objects, and it is hard to generalized one model for reconstructing objects in other categories.

In our framework, we assume no prior knowledge from the shape, and the latent factors that control the shape of the objects are treated unknown in advance. Hence our task is more generalized, that is, to extract these factors automatically in the learning process. The Gaussian Process Latent Variable Model (GPLVM) [11] and its variants can be suitable for completing this task as it automatically extracts the unknown low dimensional embedded information of the object from high dimensional observations given a relatively small amount of training samples. The information of 2D positions and depths are both used as observed data to train the shape model, which can then be used to predict the 3D shapes from new 2D instances.

The framework we propose requires no interaction and it can be easily generalized to reconstruct various categories of 3D objects which have more complex structures. Experiments performed on the synthetic and real examples show that our new approach is plausible even though only the 2D silhouettes are given as inputs.

The rest of this paper is organized as follows. In Section 2, we will give a brief review on previous methods on single view reconstruction. The framework of our learning-based reconstruction and the detailed techniques involved will be presented in Section 3. Section 4 will provide the experi-

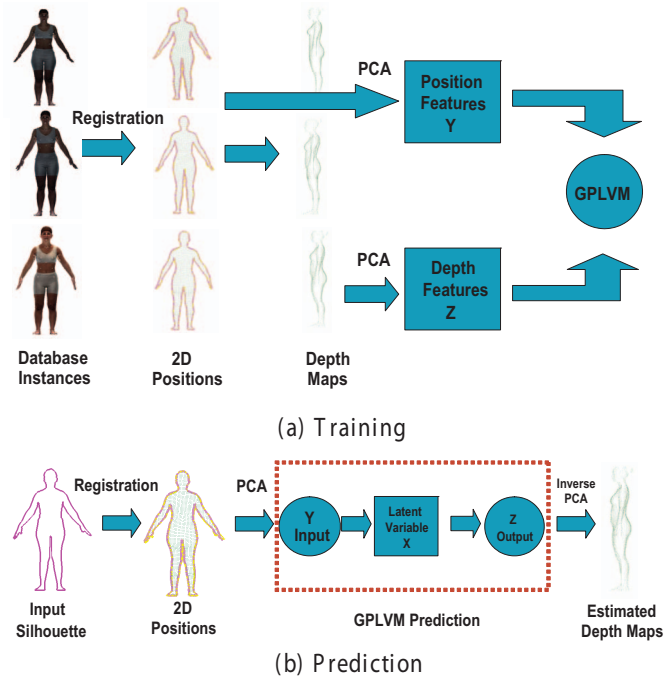

Figure 1. The process of both training and prediction stages in our approach.

ment results, and related discussions are given in Section 5. Finally, a brief conclusion is drawn in Section 6.

2. Related Work

Single view reconstruction is an under-constrained problem, and hence most previous methods tend to make strong assumptions about the input scene. Much research has been done on planar outdoor architecture scenes. Criminisi et al. [5] recover the 3D affine geometry from a single perspective image based on vanishing points information and projective geometry constraints. All the measurements in the scene can be accurately obtained once the scale factor is determined. Similar constraints are used by [2], [8]. Hoiem et al [8] first segment the image into 3 categories: ground, sky, verticals. A coarse pop-up reconstruction is then obtained based on the segmentation results. Barinova et al [2] further use a Conditional Random Field (CRF) model to infer the ground-vertical boundary parameters, and they claim that their algorithm can generate visually pleasant 3D models and be more computationally efficient. Similar projective geometry-based approach is also generalized by Delage et al. [6] to reconstruct indoor scenes. All the research above, however, only focus on planar scenes and also assume a ground verticality for the scene reconstruction. In contrast, Saxena et al [18] investigate the relation between scene depths and image features using a Markov Random Field (MRF) model. The depth estimate can be done efficiently with linear programming in their model. No explicit assumption about the geometrical structure of the scene is made in the approach.

Some efforts have been made on the problem of reconstructing curved or free-form objects [23], [16]. Zhang et al [23] introduce several user constraints such as normal map, depth discontinuities, creases, etc, and finally formulated the reconstruction problem into linearly constrained quadratic optimization problem which has a closed-form solution. However, their method is only limited to generate 2.5D Monge patches. Prasad et al. [16] generalize the scope of the problem to full 3D surfaces. They basically adopt the framework of constraints-based optimization in [23]. Contour Generators (CGs) are used for creating curved patches and objects with more complicated topology are studied. However, their approach still requires subtle interaction to mark up the parametric space when the topology of the 3D object becomes complex. Due to the ambiguous nature of SVR problems, all these approaches above often require considerable amount of interaction from users and they are usually based on some heuristic regularities such as minimizing the overall smoothness.

Our approach is different from the previous ones in the respect that we aim to learn the prior knowledge and rules of reconstruction from existing 3D models using statistical methods instead of proposing them heuristically. Some research infers the 3D shape of the object using learning methods. In [22], Torresani et al. try to learn the time-varying shape of non-rigid 3D object from uncalibrated 2D tracking data, usually monocular video sequences which record the motion of a specific object. The shape distribution is assumed, and the motion and deformation of the model are estimated through a generalized EM algorithm. The methods give satisfying results on synthetic data and are robust to missing data. However, these algorithms are very slow even for relatively simple models. The difference between the goal of their approaches and ours is that we focus more on learning the shape of a class of objects instead of tracking the motion of a specific object.

3. Methods

The framework of our reconstruction method is illustrated in Fig. 1, which includes both procedures of training the shape prior model and predicting the 3D shape based on the model obtained. 2D silhouettes and the corresponding depth scans of 3D objects are used as training data in our approach. In the training stage, we first use a common shape template to encode the 2D position and depth information for each instance in the database (see Section 3.1). After that, Principal Component Analysis (PCA) (see Section 3.2) is applied to reduce the dimension of input data before training the shape prior (see Section 3.3). In the prediction (reconstruction) stage, only the 2D silhouette is used. And

Figure 2. Template matching

the same preprocessing steps of registration and dimension reduction as the training stage are performed. Finally, the depths of the object are inferred by the GPLVM, which is trained from the combinational inputs of both 2D position and depth information.

3.1. Preprocessing and Registration

Registration of the 2D input and vectorizing the 2D position and depth information of the objects are necessary steps before the model training. For this purpose, a template matching scheme is applied in our approach. For each category of objects, the shape template with a deformable silhouette and a uniform grid inside the silhouette is generated, as shown in Fig. 2. The motivation of using a template is that we find that the position information encoded by the grid is effective and less susceptible to the imperfectness and local distortion of the input silhouettes.

Simply, we generate the template from an arbitrary instance in the category. In both training and testing stages, the template is deformed to fit the 2D shape of each instance by matching the silhouette and warping the internal grid points accordingly. A method based on the modified Iterative Closest Points (ICP) is adopted to perform the silhouette matching in our approach. Then, the 2D position and depth information of that object are encoded by the displacements of these grid points and the depth values extracted at each grid point. For the details of silhouette matching and grid warping, please see Appendix A.

3.2. Dimension Reduction

Considering the fact that the raw position and depth data obtained from the template can be of enormous dimension (around 5000), which can result in a huge memory consumption during the training stage of the GPLVM. We hence compress the data using principal component analysis (PCA). Given the new position vector $\mathbf{P}$ and depth vector $\mathbf{D}$ which are obtained by doing the template matching, we can approximately represent them into the linear combinations of mean vectors $\bar{\mathbf{P}}$ and $\bar{\mathbf{D}}$, and the first m eigenvectors $\mathbf{p}_i$ and $\mathbf{d}_i$ given by PCA.

$$\mathbf{P} = \bar{\mathbf{P}} + \sum_{i=1}^{m} \alpha_i \mathbf{p}_i, \tag{1}$$

$$\mathbf{D} = \bar{\mathbf{D}} + \sum_{i=1}^{m} \beta_i \mathbf{d}_i, \tag{2}$$

where the linear coefficients α_i and β_i can be used to characterize the 3D shape of the new instance. For our experiment, we use the first $m = 30$ principal components of both the 2D positions and the depth maps, respectively, and they usually account for around 90% variance of the data sets we use. For each instance, m-D feature vectors $\mathbf{y} = \{\alpha_i\}_{i=1}^{m}$ and $\mathbf{z} = \{\beta_i\}_{i=1}^{m}$ which consists of linear coefficients are then used as the input data pair for training the GPLVM.

3.3. Model Training and Prediction

The GPLVM [11] is shown to be an effective approach for probabilistically modeling high dimensional data that lies on a low dimensional non-linear manifold. The GPLVM and its variants have been applied to solve computer vision problems, mainly in the context of human pose estimation [7], [9], [15], [19] and tracking the deformable surface [17]. In the setting of the reconstruction problem, the training data includes 2D position features and depth features, which are given in pairs. We aim to recover underlying low-dimensional sub-manifold structure that can model such pair-wise relationship. Shared-GPLVM [19], a variant of GPLVM which handles multiple observations that share the same latent structure, is suitable to model this relationship.

3.3.1 Training a Shape Model

In our problem, N pairs of position and depth features: $(\mathbf{Y}, \mathbf{Z}) = [(\mathbf{y}_1, \mathbf{z}_1), (\mathbf{y}_2, \mathbf{z}_2), \cdots, (\mathbf{y}_N, \mathbf{z}_N)]$ which are obtained from the previous subsection, are given as the training data of the model. In the shared GPLVM, such a manifold structure is described by q-dimensional latent variables $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$ and each $\mathbf{x}_i$ simultaneously controls the corresponding observed feature pair $(\mathbf{y}_i, \mathbf{z}_i)$, $i = 1, 2, \cdots, N$. Observations $\mathbf{Y}$ and $\mathbf{Z}$ are conditional independent given the latent structure $\mathbf{X}$.

The likelihood of observations can be formulated as the following product of Gaussian process:

$$P(\mathbf{Y}|\mathbf{X}, \theta_{\mathbf{Y}}) = \prod_{i=1}^{m} \mathcal{N}(\mathbf{Y}_{:,i}|0, \mathbf{K}_{\mathbf{Y}}), \tag{3}$$

$$P(\mathbf{Z}|\mathbf{X}, \theta_{\mathbf{Z}}) = \prod_{i=1}^{m} \mathcal{N}(\mathbf{Z}_{:,i}|0, \mathbf{K}_{\mathbf{Z}}), \tag{4}$$

where $\mathbf{Y}_{:,i}$ and $\mathbf{Z}_{:,i}$ denotes the $N \times 1$ column vector constructed from the i-th dimension of $\mathbf{Y}$ and $\mathbf{Z}$, respectively, $\mathbf{K}_{\mathbf{Y}} = [k_Y(\mathbf{x}_i, \mathbf{x}_j)]_{1 \le i \le N, 1 \le j \le N}$ and $\mathbf{K}_{\mathbf{Z}} =$

$[k_Z(\mathbf{x}_i, \mathbf{x}_j)]_{1\leq i\leq N, 1\leq j\leq N}$ are non-linear kernel matrices, and $\theta_\mathbf{Y} = \{\theta_{Y,i}\}_{i=1}^4$ and $\theta_\mathbf{Z} = \{\theta_{Z,i}\}_{i=1}^4$ refer to the hyper-parameters in the nonlinear kernels $\mathbf{K_Y}$ and $\mathbf{K_Z}$, respectively. $\mathbf{K_Y}$ and $\mathbf{K_Z}$ are defined as the compound kernels (the combination of RBF kernels and linear kernels) in this paper:

$$\begin{aligned} k_{Y(Z)}(\mathbf{x}_i, \mathbf{x}_j) &= \theta_{Y(Z),1} e^{-\frac{\theta_{Y(Z),2}}{2}(\mathbf{x}_i-\mathbf{x}_j)^T(\mathbf{x}_i-\mathbf{x}_j)} \\ &\quad +\theta_{Y(Z),3}^{-1}\delta_{ij} + \theta_{Y(Z),4}\mathbf{x}_i^T\mathbf{x}_j, \end{aligned} \quad (5)$$

And we assume the prior of the latent variables $\mathbf{X}$ to be the product of independent Gaussian distributions:

$$P(\mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|0, I). \quad (6)$$

The model minimizes the log joint marginal posterior L with respect to both the latent points $\mathbf{X}$ and the hyper-parameters θ_Y and θ_Z of the kernels $\mathbf{K_Y}$ and $\mathbf{K_Z}$, where

$$\begin{aligned} L &= -\log P(\mathbf{X}|\mathbf{Y}, \mathbf{Z}, \theta_\mathbf{Y}, \theta_\mathbf{Z}) \\ &= -\log P(\mathbf{Y}|\mathbf{X}, \theta_\mathbf{Y}) P(\mathbf{Z}|\mathbf{X}, \theta_\mathbf{Z}) P(\mathbf{X}) + const \\ &= \frac{1}{2} tr(\mathbf{K_Y}^{-1}\mathbf{Y}\mathbf{Y}^T) + \frac{1}{2} tr(\mathbf{K_Z}^{-1}\mathbf{Z}\mathbf{Z}^T) + \frac{1}{2}\|\mathbf{X}\|^2 \\ &\quad + \frac{m}{2}\log|\mathbf{K_Y}| + \frac{m}{2}\log|\mathbf{K_Z}| + const. \end{aligned} \quad (7)$$

We use scaled conjugate gradient (SCG) method [13] to minimize (7) with an Isomap [21] initialization on the latent positions.

3.3.2 Prediction of Depths

Inferring the depth feature $\tilde{\mathbf{z}}$ from a new 2D-position feature $\tilde{\mathbf{y}}$ in our model is theoretically equivalent to maximizing the following conditional distribution.

$$\begin{aligned} &P(\tilde{\mathbf{z}}|\tilde{\mathbf{y}}, \mathbf{Z}, \mathbf{Y}, \mathbf{X}, \theta_Y, \theta_Z) = \\ &\int P(\mathbf{x}|\tilde{\mathbf{y}}, \mathbf{Y}, \mathbf{X}, \theta_Y) P(\tilde{\mathbf{z}}|\mathbf{x}, \mathbf{Z}, \mathbf{X}, \theta_Z) \mathrm{d}\mathbf{x}, \end{aligned} \quad (8)$$

Since there is no closed-form solution to maximize the integral in (8), we approximate the prediction with a two-stage process. In the first stage, we shall find the position $\tilde{\mathbf{x}}$ in the latent space which is most likely to generate the observed 2D-position feature. Unfortunately, GPLVM does not give a simple functional representation for this inverse mapping. Hence, here the latent position $\tilde{\mathbf{x}}$ is found by maximizing the predictive posterior using gradient-based method.

$$\begin{aligned} \tilde{\mathbf{x}} &= \mathrm{argmax}_x P(\mathbf{x}|\tilde{\mathbf{y}}, \mathbf{Y}, \mathbf{X}, \theta_Y) \\ &= \mathrm{argmax}_x P(\tilde{\mathbf{y}}|\mathbf{x}, \mathbf{Y}, \mathbf{X}, \theta_Y) P(\mathbf{x}). \end{aligned} \quad (9)$$

Concerning the fact that Eqn. (9) can usually be multi-modal, which means that the same 2D-position feature input can correspond to several solutions in the latent space. We hence adopt a multiple-initialization scheme to search those multiple peaks.

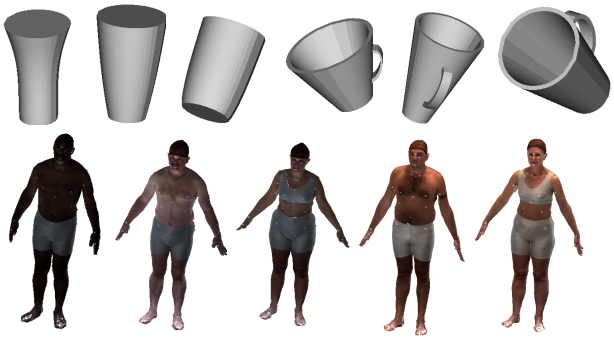

Figure 3. Instances of the datasets we use.

In the second stage, the depth feature $\tilde{\mathbf{z}}$ is to be estimated based on the latent positions we have found.

$$\tilde{\mathbf{z}} = \mathrm{argmax}_z P(\mathbf{z}|\tilde{\mathbf{x}}, \mathbf{Z}, \mathbf{X}, \theta_Z). \quad (10)$$

The second stage is the forward mapping and it has a Gaussian closed-form representation as follows.

$$P(\tilde{\mathbf{z}}|\tilde{\mathbf{x}}, \mathbf{Z}, \mathbf{X}, \theta_Z) = \mathcal{N}(\tilde{\mathbf{x}}|\mathbf{k_Z}(\tilde{\mathbf{x}}, \mathbf{X})\mathbf{K_Z}^{-1}\mathbf{Z}, \Sigma(\tilde{\mathbf{x}}, \tilde{\mathbf{x}})), \quad (11)$$

And hence the most probable depth features that correspond to all optimized latent points found in the first stage can be simply predicted by the mean $\tilde{\mathbf{z}} = \mathbf{k_Z}(\tilde{\mathbf{x}}, \mathbf{X})\mathbf{K_Z}^{-1}\mathbf{Z}$.

Finally, the complete depth maps of the 3D object can be fully reconstructed from the estimated depth feature $\tilde{\mathbf{z}}$ with the mean vector $\bar{\mathbf{D}}$ and the PCA eigenvectors $\mathbf{d}_i$ ($i = 1, 2, \cdots, m$) which are stored in memory.

4. Results

In order to verify the efficacy of our approach, we first train the shape models on 3 datasets: a dataset of 2000 synthetic vases, a dataset of 2000 synthetic mugs, as well as a real human body dataset which contains over 2000 body scans of North American and European adults (Civilian American and European Surface Anthropometry Resource (CAESAR), SAE International). Some instances of each dataset are illustrated in Fig. 3. We examine the reconstruction results estimated by these shape models with the 2D silhouettes of objects in the corresponding category provided. In this paper, we assume that the silhouettes of the objects in the frontal view are the only data given as the input for reconstruction.

4.1. Synthetic 3D Shape Reconstruction

We use parametric models to generate two 3D synthetic shape datasets: a vase dataset and a mug dataset. The ranges of parameters of both datasets are listed in Table 1.

Table 1. The range of parameters for the synthetic objects.

Dataset	Parameters (mm)						
	Top Radius r_1	Bottom Radius r_2	Height H	Handle Size r_h	Curvature Parameters s_1	s_2	Side Silhouette $t \in [0,1]$
Mugs	25–40	15–25	50–110	$\frac{r_1}{4} - (\frac{r_1}{4} + 3)$	N/A	N/A	$r_1 t + r_2(1-t)$
Vases	25–40	15–25	70–130	N/A	-5 – 5	0.5–1.5	$r_1 t + r_2(1-t) + s_1 \sin(\pi t^{s_2})$

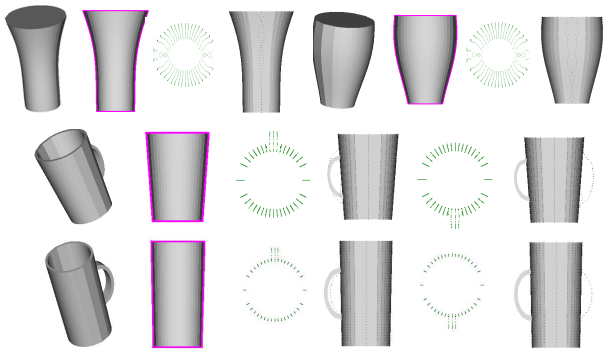

Figure 4. Qualitative results on synthetic mug and vase datasets. For each testing instance, we give its one arbitrary view, the input frontal view with the silhouette, and the reconstructed result (in dots) in both top view and side view along with the ground truth. For the mug instances (row 3 and 4), two modes with the highest predictive posterior values (given in green dots) are given.

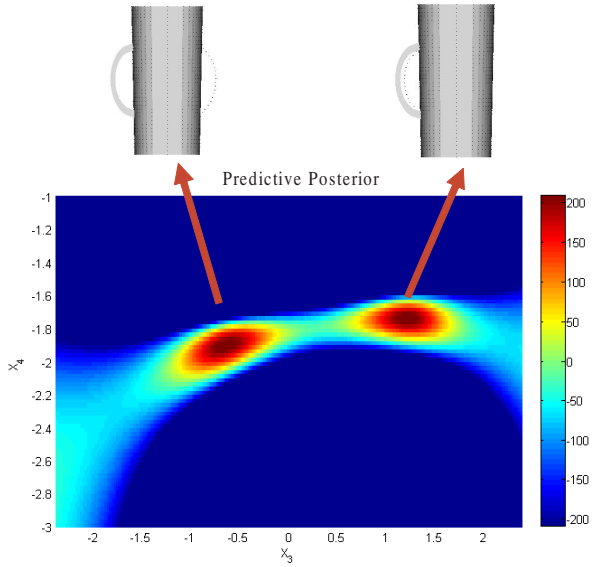

Figure 5. The bimodal predictive posterior of the last mug sample in Fig. 4. Here we fix the first two dimensions of the latent space and plot the distribution of the posterior with respect to the 3rd and 4th dimensions. Each peak of the predictive posterior corresponds to the case of two different handle directions.

We apply our approach to reconstruct new objects based on a training set of 800 instances for each type of shapes, and some good qualitative results are given in Fig. 4. In general, the shape model we learn can generate satisfactory reconstruction results.

It is worthy to mention that in the mug dataset, we deliberately change the direction of mug handle randomly, which can be either in front or behind. Since the change of handle position does not affect the external silhouette of the mug, we expect to see a bimodal structure in the GPLVM prior model which is learned from such a mug dataset. As expected, our model is capable of generating both feasible reconstruction results. As shown in Fig. 5, given an ambiguous frontal silhouette of the mug (a trapezoid), our approach is able to find both possible solutions from the two peaks in the predictive posterior (Eqn. (9)), each of which corresponding to the two different handle directions.

4.2. Human Body Reconstruction

Compared with the main stream approaches are based on multi-view reconstruction and parametric deformable body model [4], [14], [20], our approach learns the human body prior in the single-view setting.

We train the GPLVM body shape prior from the 2D projections and depth maps of 800 instances in standing pose. We train two shape models for both males and females as well as a model for mixed-gender. In the testing stage, we provide only the 2D standing pose silhouettes in frontal view as inputs to the shape model we obtain. Some of the reconstruction results are given in Fig. 6. For each testing instance, the model automatically returns several depth distributions with the highest predictive posterior values, which can be compared with the ground truth.

The model training takes around 3 hours (including the time for registration and dimension reduction), while the prediction (reconstruction) for each instance takes 25 minutes on average when we run the code on a 2.5GHz processor, which includes 30 restarts for searching the multiple peaks in the predictive posterior.

4.3. Quantitative Evaluations

We also evaluate the accuracy of the reconstruction method quantitatively. For the purpose of comparison, we also implement another reconstruction method which is based on searching the nearest-neighbors (NN) in the database and returning the corresponding depth maps. We adopt Hausdorff Fraction [10] among the silhouettes as the distance measurement for instances in the implementation.

We compare the performance of both algorithms on all

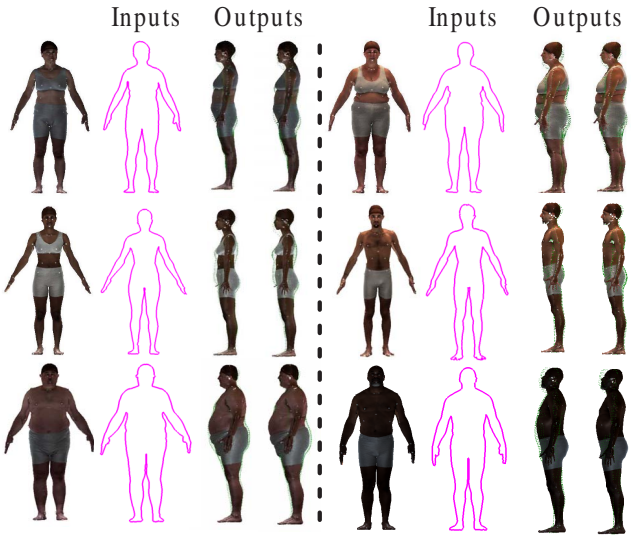

Figure 6. Experimental results on CAESAR human dataset. The prior model is trained from mixed gender instances and the testing inputs are the silhouettes (column 2) extracted from the frontal views of standing human instances (column 1). In column 3 and 4, two candidate results with the highest predictive posterior values (given in green dots) are given in contrast with the ground truth.

synthetic data sets and the real human body dataset we apply. For each dataset, we use 800 instances randomly selected from the datasets to train the GPLVM shape prior and another 800 as the testing set. The exact same training sets and testing sets are used in the nearest-neighbor search. The evaluation is based on the error between the ground truth and the candidate reconstructed results around the certain locations. In this paper, the error measurement of the reconstruction is defined as:

$$Err = \frac{1}{N}\sum_{i=1}^{N}|d_i - d_{i,0}|, \tag{12}$$

where d_i and $d_{i,0}$ are the reconstructed and ground truth thickness values at sampling position i, respectively, and N is the total number of sampling positions. For the synthetic data, we sample the thickness values at all the grid points, while for the human body data, the thickness values are sampled around both chest and waist parts. From the results tabulated in Table 2, we can see that our approach always outperforms the nearest-neighbor-based approach in all the contexts listed.

Choosing the proper complexity of the model can be a problem. In GPLVM, a low dimensional latent space may not be enough to characterize the inherent manifold structure of the data, while a latent dimension which is too high could lead to over-fitting and poor generalization. We test several different latent dimension settings in our experiments to find out the most appropriate latent dimension for each category of objects, and we finally set the latent dimension of the model to be 5 for the vase data, 4 for the mug data, and 6 for the human-body data, respectively.

We also investigate how the size of training set influences the speed and the precision of our approach. The computational complexity of GPLVM training is determined by the inversion of the kernel matrices, i.e., $O(N^3)$, where N is the size of the training set; while in the depth prediction stage, the complexity is $O(N)$ for each iteration, which is determined by the matrices multiplications in Eqn. (9) and (10), where $\mathbf{K_Z}^{-1}\mathbf{Z}$ can be pre-computed and stored. For each dataset, we also train another GPLVM based on a small training set with only 200 instances which are selected from the original training set based on a sparse method, named Informative Vector Machine (IVM) [12]. In the training process, data points are added to the model one at a time, and at each step the point with highest construction variances will be selected. In this way, the selected straining set tends to be made up of those data points that are reasonably well spaced throughout the latent space. The approximate time for model training (excluding the preprocessing and dimension reduction) is then shortened to 3 minutes, while the average time for depth prediction on each instance drops sharply to 40 seconds. From Table 2, we can see that this significant reduction in the number of training samples only leads to a slight drop in the performance of our GPLVM, while greatly enhancing the training and prediction speed.

5. Discussion and Future Work

Experiments show that our approach works well on most of the data in the previous section. In this section, we discuss several issues regarding the reconstruction framework presented in this paper.

In our single view reconstruction problem, the frontal silhouettes do not convey enough information for precisely inferring the depth distribution, which usually results in ambiguity in the 3D structure. For most categories of 3D objects, multi-modality is common in the predictive posterior, as the mug example that Fig. 5 shows, i.e., the instances with a similar silhouette may have strikingly different depth distributions, which correspond to multiple local maximum of the predictive posterior. A failure result from our approach is given in Fig. 7. In that example, the lady with a relative wider waist is predicted to have a thicker belly according our model learned from the training set, although it is not the truth. The solution to this problem can be incorporating more visual information, such as texture and color, or introducing another input view (multi-views) in order to remove the ambiguity and achieve more precise reconstruction, which will be our next step.

At the current stage, we assume all the silhouettes are obtained from the frontal view, and for the human body data, only the standing pose is considered. We have not taken into account the pose changes and view changes of the in-

Table 2. The quantitative comparison between our method (GPLVM and GPLVM-IVM (which is based on a smaller training set of 200 instances selected by IVM) and nearest neighbor reconstruction (NN). Errors and standard deviations are given in millimeters.

Data Set		The 1st Candidate			Best among the First 3		
		GPLVM	GPLVM-IVM	NN	GPLVM	GPLVM-IVM	NN
Synthetic Vases		**1.65±0.92**	1.83±1.40	4.42±2.16	**1.57± 0.98**	1.68± 1.28	2.87±1.55
Synthetic Mugs		1.08±0.48	**1.00±0.55**	1.92±1.49	**0.90± 0.48**	0.90± 0.57	1.32±0.81
Human Bodies (mixed-gender)	Chest	**20.8±17.3**	22.0±17.8	34.4±27.3	**14.9±15.0**	15.9±14.2	19.1±17.9
	Waist	**21.7±18.7**	23.1±18.4	39.7±33.2	**18.8±17.1**	19.6±16.9	23.0±20.9
Human Bodies (female)	Chest	**21.9±20.5**	22.1±18.8	39.8±30.3	15.4±14.1	**15.0±13.8**	20.3±19.5
	Waist	25.4±24.0	**25.0±21.9**	42.5±31.5	18.3±13.7	**18.1±16.0**	24.0±21.0

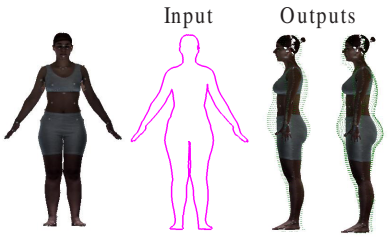

Figure 7. A failure example. The model generates the wrong depth maps in the first two candidates.

put. Presumably, a complete shape model in our framework is expected to learn these variations also as a part of latent factors and thus be adapted to input changes. Therefore, another important research issue for us in the near future is to investigate how these variations in the input data influence our reconstruction approach and fit our framework to various types of inputs.

6. Conclusion

In this paper, we propose a novel single-view reconstruction framework on the basis of learning the shape prior with a Gaussian Process Latent Variable Model. Compared with previous methods, our approach is fully automatic and it does not depend on any predefined parametrical model and heuristic regularities. A significant advantage of the framework of learning-based reconstruction we propose in this paper is that it can be easily generalized to deal with various categories of 3D objects that may have more complex geometrical and topological structures just by simply adjusting the dimension of latent space in the model. Besides, the framework we propose is able to reconstruct 3D shapes from limited input information, such as 2D silhouette, without any interaction. To verify the approach, some preliminary qualitative and quantitative experiments under standard settings are conducted.

The extension of the current research may include: 1) expanding the current framework to incorporate multiple visual cues to achieve more accurate reconstruction; 2) using the propose framework to solve multi-view reconstruction problem; 3) to further cope with articulation pose changes and viewpoint changes of the input data; 4) conducting a more thorough experiment over a wider range of objects, such as fruit, cars, 3D faces, etc.; 5) speeding up the implementation and recovering the 3D shape from monocular video sequences.

References

[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, and J. Rodgers. SCAPE: Shape completion and animation of people. *SIGGRAPH*, 24:408–416, 2005.

[2] O. Barinova, V. Konushin, A. Yakubenko, K. Lee, H. Lim, and A. Konushin. Fast automatic single-view 3-d reconstruction of urban scenes. *ECCV*, 2008.

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

[4] A. Bălan and M. Black. The naked truth: estimating body shape under clothing. *ECCV*, 2008.

[5] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. *IJCV*, 40(2):123–148, 2000.

[6] E. Delage, H. Lee, and A. Ng. A dynamic bayesian network model for autonomous 3D reconstruction from a single indoor image. *CVPR*, 2:2418–2428, 2006.

[7] A. Gupta, T. Chen, F. Chen, D. Kimber, and L. Davis. Context and observation driven latent variable model for human pose estimation. *CVPR*, 2008.

[8] D. Hoiem, A. Efros, and M. Hebert. Automatic photo pop-up. *SIGGRAPH*, pages 577–584, 2005.

[9] S. Hou, A. Galata, F. Caillette, N. Thacker, and P. Bromiley. Real-time body tracking using a gaussian process latent variable model. *ICCV*, 2007.

[10] D. Huttenlocher, J. Noh, and W. Rucklidge. Tracking non-rigid objects in complex scenes. *ICCV*, pages 93–101, 1993.

[11] N. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *NIPS*, 16:329–336, 2004.

[12] N. Lawrence, M. Seeger, and R. Herbrich. Fast sparse gaussian process methods: The informative vector machine. *NIPS*, 15:625–632, 2003.

[13] M. Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6:525–533, 1993.

[14] L. Mündermann, S. Corazza, and T. Andriacchi. Accurately measuring human movement using articulated ICP with soft-joint constraints and a repository of articulated model. *CVPR*, 2007.

[15] R. Navaratnam, A. Fitzgibbon, and R. Cipolla. Semi-supervised joint manifold learning for multi-valued regression. *ICCV*, 2007.

[16] M. Prasad, A. Zisserman, and A. Fitzgibbon. Single view reconstruction of curved surfaces. *CVPR*, 2:1345–1354, 2006.

[17] M. Salzmann, R. Urtasun, and P. Fua. Local deformation models for monocular 3D shape recovery. *CVPR*, 2008.

[18] A. Saxena, S. Chung, and A. Ng. 3-D depth reconstruction from a single still image. *IJCV*, 76(1):53–69, 2008.

[19] A. Shon, K. Grochow, A. Hertzmann, and R. Rao. Learning shared latent structure for image synthesis and robotic imitation. *NIPS*, 16:1233–1240, 2006.

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

[21] J. Tenenbaum. Mapping a manifold of perceptual observations. *NIPS*, 10:682–688, 2004.

[22] L. Torresani, A. Hertzmann, and C. Bregier. Nonrigid structure-from-motion: Estimating shape and motion with hierachical priors. *IEEE Trans. PAMI*, 30(5):878–892, 2008.

[23] L. Zhang, G. Dugas-Phocion, and J. Samson. Single-view modelling of free-form scenes. *The Journal of Visualization and Computer Animation*, 13:225–235, 2002.

A. Appendix: Silhouette Matching and Grid Warping

We denote the deformable silhouette template in the t-th iteration as the point set $\{\mathbf{X}_{i,t}\}_{i=1}^{N}$ and the target silhouette as the point set $\{\mathbf{Y}_j\}_{j=1}^{N}$. The update equation at the t-th iteration can be written as:

$$\begin{aligned}\mathbf{X}_{i,t+1} &= \mathbf{X}_{i,t} + \eta\boldsymbol{\Delta}\mathbf{X}_{i,t} \\ &= \mathbf{X}_{i,t} + \eta(\boldsymbol{\Delta}\mathbf{X}_{\mathbf{b}i,t} + \lambda_1\boldsymbol{\Delta}\mathbf{X}_{\mathbf{d}i,t} + \lambda_2\boldsymbol{\Delta}\mathbf{X}_{\mathbf{l}i,t}) \quad (13)\end{aligned}$$

In (13), the position update $\boldsymbol{\Delta}\mathbf{X}_{i,t}$ consists of the following three terms in our algorithm.

First, the boundary term $\boldsymbol{\Delta}\mathbf{X}_{\mathbf{b}}$, simply defined by the point-wise distance of two silhouettes, enforces the good matching between template and the target silhouette:

$$\boldsymbol{\Delta}\mathbf{X}_{\mathbf{b}i,t} = \frac{\sum_{j=1}^{M}\alpha_j(\mathbf{X}_{i,t} - \mathbf{Y}_j)}{\sum_{j=1}^{M}\alpha_j}, \quad (14)$$

where $\alpha_j = \exp(-\|\mathbf{X}_{i,t} - \mathbf{Y}_j\|^2/\sigma^2)$, $j = 1, 2, \cdots, M$.

Second, the deformation term $\boldsymbol{\Delta}\mathbf{X}_{\mathbf{d}}$ regulates that the neighboring points on the silhouette should maintain their relative positions from each other during the deformation. Ideally, they should neither cluster together nor stay far away from each other after the template deformation. The term is given in the following equation:

$$\boldsymbol{\Delta}\mathbf{X}_{\mathbf{d}i,t} = \frac{\sum_{j=1}^{N}\beta_j(\mathbf{X}_{i,t} - \mathbf{X}_{j,t} - \mathbf{X}_{i,0} + \mathbf{X}_{j,0})}{\sum_{j=1}^{N}\beta_j}, \quad (15)$$

where $\beta_j = \exp(-\|\mathbf{X}_{i,0} - \mathbf{X}_{j,0}\|^2/\sigma^2)$, $j = 1, 2, \cdots, N$.

Third, the landmark term $\boldsymbol{\Delta}\mathbf{X}_{\mathbf{l}}$ ensures that the silhouette deformation should coincident with the landmark registration of the template, as given in the equation below.

$$\boldsymbol{\Delta}\mathbf{X}_{\mathbf{l}i,t} = \frac{\sum_{j=1}^{L}\gamma_j(\mathbf{X}_{i,t} - \mathbf{L}_j^* - \mathbf{X}_{i,0} + \mathbf{L}_j^*)}{\sum_{j=1}^{L}\gamma_j}, \quad (16)$$

where $\{\mathbf{L}_j\}_{j=1}^{J}$ are the default positions of the landmarks in the template silhouette, $\{\mathbf{L}_j^*\}_{j=1}^{J}$ are the corresponding positions of those landmarks in the target silhouette, and the weighting factors $\gamma_j = \exp(-\|\mathbf{X}_{i,0} - \mathbf{L}_j\|^2/\sigma^2)$, $j = 1, 2, \cdots, L$. The landmark term is removed if no landmark information is provided. However, we find it can be quite useful for guiding a quick and accurate silhouette matching.

The silhouette matching algorithm is run for several iterations and the free parameters are given as follows: $\lambda_1 = \lambda_2 = 1.0$, $\eta = 0.5$, $\sigma = 0.02/4^t$, which are fixed through the experiments.

With the deformation of the silhouette, we also hope to establish a one-to-one mapping between the deformed template and the original one for all the grid points lying inside the silhouette. Let $\mathbf{P}_0$ be an arbitrary point lying inside the template silhouette and $\mathbf{P}_t$ be its correspondence after t iterations of silhouette warping. The local warping can be formulated as the following displacement equation:

$$\begin{aligned}\mathbf{P}_t &= \mathbf{P}_0 + \boldsymbol{\Delta}\mathbf{P} \\ &= \mathbf{P}_0 + \lambda'\boldsymbol{\Delta}\mathbf{P}_{\mathbf{s}} + (1-\lambda')\boldsymbol{\Delta}\mathbf{P}_{\mathbf{l}}, \quad (17)\end{aligned}$$

In our method, we initialize the local displacement $\boldsymbol{\Delta}\mathbf{P}$ with the deformed silhouette and the landmark displacement (if landmark is given), which correspond to two terms: the silhouette-driven displacement $\Delta\mathbf{P}_s$ and the landmark-driven displacement $\Delta\mathbf{P}_l$:

$$\boldsymbol{\Delta}\mathbf{P}_{\mathbf{s}} = \frac{\sum_{i=1}^{N} w_i(\mathbf{X}_{i,t} - \mathbf{X}_{i,0})}{\sum_{i=1}^{N} w_i}, \quad (18)$$

$$\boldsymbol{\Delta}\mathbf{P}_{\mathbf{l}} = \frac{\sum_{j=1}^{L} v_j(\mathbf{L}_j^* - \mathbf{L}_j)}{\sum_{j=1}^{L} v_j}, \quad (19)$$

where $w_i = \frac{\sigma'}{\|\mathbf{P}_0 - \mathbf{X}_{i,0}\|}\exp(-\|\mathbf{P}_0 - \mathbf{X}_{i,0}\|^2/\sigma'^2)$, $i = 1, 2, \cdots, N$, $\{\mathbf{L}_j\}_{j=1}^{J}$ are the original positions of the landmarks in the template silhouette, $\{\mathbf{L}_j^*\}_{j=1}^{J}$ are the corresponding positions of those landmarks in the target silhouette, and the weighting factors $v_j = \frac{\sigma'}{\|\mathbf{P}_0 - \mathbf{L}_j\|}\exp(-\|\mathbf{P}_0 - \mathbf{L}_j\|^2/\sigma'^2)$, $j = 1, 2, \cdots, L$. We fix the parameters to be $\lambda' = 0.2$ and $\sigma' = 0.1$.

In order to generate a smoother and more homogeneous warping, we further apply a simple iterative averaging scheme on each grid point on the basis of its 4-neighborhood.