Cloth Dynamics Modeling in Latent Spaces and its Application to Robotic Clothing Assistance

Conference Paper · September 2015

DOI: 10.1109/IROS.2015.7353860

Nishanth Koganti

Nara Institute of Science and Technology

7 PUBLICATIONS **7** CITATIONS

SEE PROFILE

Kazushi Ikeda

Nara Institute of Science and Technology

145 PUBLICATIONS **389** CITATIONS

SEE PROFILE

Tomohiro Shibata

Kyushu Institute of Technology

110 PUBLICATIONS **1,065** CITATIONS

SEE PROFILE

Cloth Dynamics Modeling in Latent Spaces and its Application to Robotic Clothing Assistance

Nishanth Koganti[1], Jimson Gelbolingo Ngeo[1,2], Tamei Tomoya[1], Kazushi Ikeda[1] and Tomohiro Shibata[1,2]

Abstract— **Real-time estimation of human-cloth relationship is crucial for efficient learning of motor skills in robotic clothing assistance. However, cloth state estimation using a depth sensor is a challenging problem with inherent ambiguity. To address this problem, we propose the offline learning of a cloth dynamics model by incorporating reliable motion capture data and applying this model for the online tracking of human-cloth relationship using a depth sensor. In this study, we evaluate the performance of using a shared Gaussian Process Latent Variable Model in learning the dynamics of clothing articles. The experimental results demonstrate the effectiveness of shared GP-LVM in capturing cloth dynamics using few data samples and the ability to generalize to unseen settings. We further demonstrate three key factors that affect the predictive performance of the trained dynamics model.**

I. Introduction

Robotics for elderly care and rehabilitation is a growing social need due to recent demographic trends such as the aging population combined with shortage of trained professionals. Clothing Assistance is a basic assistance activity in the daily life of the elderly and disabled. However, robotic clothing assistance is still considered an open problem as it involves a tightly coupled interaction between the human subject, robot and the non-rigid clothing articles. Robust and real-time estimation of the human-cloth relationship is crucial for the efficient learning of motor-skills by the robot. The challenge in this problem mostly lies with cloth state estimation due to their inherent non-rigidity and occlusion by the human subject along with self-occlusion.

There have been several studies in the recent years that handle the challenge of robotic cloth handling. Towner et al. [16] proposed a method to identify a clothing article and bring it to a desired configuration using a dual-arm robot. They have used a Hidden Markov Model (HMM) for tracking the configuration of clothing articles and simulated the clothing articles using a triangulated mesh model. Ramisa et al. [17] detected the best grasping point in clothes lying on a horizontal surface to avoid multiple regraspings by using Bag of Features based detector to handle large variations in the shape of clothing material. Miller et al. [18] proposed the use of parameterized shape models and energy functions to recognize the configuration of clothing articles when spread out on a flat surface. Balaguer et al. [19] proposed a motor-skills learning framework for a dual-arm robot to perform towel folding task where the robot learned the optimal trajectory using Imitation Learning and Reinforcement Learning. In the work by Kita et al. [20], cloth state recognition was done by simulating physical deformation of cloth and adjusting the predicted shape to observed 3D data. The cloth was modeled using a deformable mesh model and the predicted shape was used to perform relevant robotic tasks. These previous studies mentioned, however, are not suitable for our objective as they do not model the interaction of clothing article with human and also do not handle high dynamics motion followed by a clothing article during a clothing task.

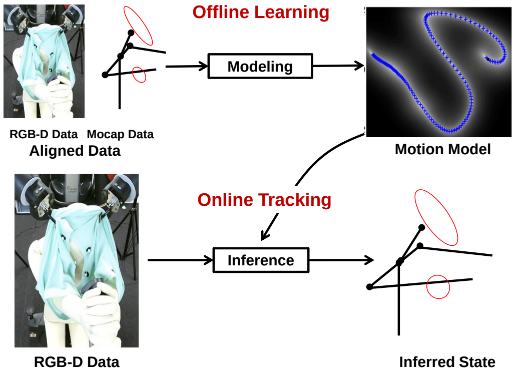

Fig. 1. Framework for cloth dynamics modeling and estimation of human-cloth relationship

In our previous study [1], we have developed a method for the real-time estimation of the human-cloth relationship using a low cost depth sensor. However, the solution relied upon the use of color markers on the clothing article and the accuracy was reduced under severe occlusion. In this study, we propose a more generic solution for the tracking of human-cloth relationship through marker-less estimation. We tackle the problem by assuming that the clothing articles follow consistent dynamics while the clothing task is being performed and by learning the underlying cloth dynamics using the nonlinear dimensionality reduction technique shared Gaussian Process Latent Variable Model (GP-LVM). This dynamics model learned in an offline fashion can be used to infer the human-cloth relationship from noisy depth sensor observations as shown in Figure 1.

Shared GP-LVM can be used to learn a joint low-dimensional latent manifold/ dynamics model for data ob-

[1] Nishanth Koganti, Jimson Gelbolingo Ngeo, Tamei Tomoya and Kazushi Ikeda are with the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara, Japan nishanth-k, jimson-n, tomo-tam, kazushi@is.naist.jp

[2] Tomohiro Shibata is with the Graduate School of Life Sciences and Systems Engineering, Kyushu Institute of Technology, Kitakyushu, Fukuoka, Japan tom@brain.kyutech.ac.jp

tained from simultaneous observation of the clothing article using a motion capture system (pose space) and a depth sensor (feature space). The motivation behind this sensor fusion is that both systems have complimentary capabilities, when combined provide the most informative dynamics model. The motion capture system can provide accurate location information of discrete markers in the environment, however, it is an expensive and complex system that requires precise calibration. On the other hand, depth sensors are low-cost and calibration free, however, they provide noisy point cloud information of the whole environment. The shared GP-LVM model provides a principled probabilistic framework for inferring the accurate motion capture state when only the noisy depth sensor feature state is available in real-time using the learned dynamics model. In this study, we further investigate the effect of factors such as feature representations and inference techniques on the predictive performance of the trained cloth dynamics model.

The rest of the paper is organized as follows. Section II, introduces the shared latent variable model. In Section III, we describe our approach for learning cloth dynamics. Section IV shows the experimental results. Finally we conclude in Section V with some future directions.

II. Non-linear Latent Variable Models

In this study, we assume that the clothing article follows consistent dynamics during clothing assistance and we propose the use of shared Gaussian Process Latent Variable Model [6] to learn the underlying dynamics. In this section, we describe the working of shared GP-LVM.

A. GP-LVM

Lawrence et al. [2] proposed a non-linear dimensionality reduction technique based on Gaussian Processes (GP) called the Gaussian Process Latent Variable Model (GP-LVM). GP-LVM is a generative model where the observations, $\mathbf{y}_i \in \mathbb{R}^D$, are assumed to be generated through a noisy process from a latent variable $\mathbf{x}_i \in \mathbb{R}^q$,

$$\mathbf{y}_i = f(\mathbf{x}_i) + \epsilon, \;\; \epsilon \sim \mathbb{N}(\mathbf{0}, \beta^{-1}\mathbf{I}) \tag{1}$$

In contrast to Probabilistic PCA [3], GP-LVM takes the approach of marginalizing over the mapping f and optimizing the latent positions $\{\mathbf{x}_i\}$. The use of a non-linear covariance function with a GP mapping f leads to non-linear dimensionality reduction. The marginal likelihood for the GP mapping is given by:

$$p(\mathbf{Y}|X, \Phi) = \frac{1}{\sqrt{(2\pi)^{ND}|\mathbf{K}|^{D}}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T)\right) \tag{2}$$

where N is the number of samples, Φ are the unknown hyper parameters and $\mathbf{K}$ is the kernel matrix. The unknowns here are estimated by maximizing the marginal likelihood with respect to both the latent positions and the hyper parameters:

$$\{\hat{\mathbf{X}}, \hat{\Phi}\} = \mathrm{argmax}_{\mathbf{X},\Phi}\, p(\mathbf{Y}|\mathbf{X}, \Phi) \tag{3}$$

There is no closed form solution for this maximization and so an iterative approach is used. The only parameter that needs to be explicitly set is the dimensionality of the latent space q.

B. Extensions to GP-LVM

Several extensions to GP-LVM have been proposed in the recent years:

- **Gaussian Process Dynamics Model**: Wang et al. [4] proposed an extension to find a latent space that includes the ordering of the observed sequential data. This is done by specifying a predictive function over the sequence in the latent space,

$$\mathbf{x}_t = h(\mathbf{x}_{t-1}) + \epsilon_{dyn} \tag{4}$$

  where $\epsilon_{dyn} \sim \mathbb{N}(\mathbf{0}, \beta_{dyn}^{-1}\mathbf{I})$ and a GP prior is placed over the function $h(\mathbf{x})$. Incorporating this dynamics model ensures learning of a mapping that considers the ordering of data.
- **Back-Constrained GP-LVM**: The use of a smooth covariance function for f in GP-LVM ensures that points close in the latent space will be close in the observation space. However, the inverse mapping need not be smooth. Lawrence et al. [5] have proposed a constrained extension of GP-LVM where each training point in the latent space is given by a smooth mapping from its corresponding observed data point, $\mathbf{x}_i = g(\mathbf{y}_i, \Phi_X)$. This ensures a bijective mapping between the latent space and the corresponding observation space. The marginal likelihood is modified as follows:

$$\{\hat{\Phi}_Y, \hat{\Phi}_X\} = \mathrm{argmax}_{\Phi_Y,\Phi_X}\, p(\mathbf{Y}|\Phi_Y, \Phi_X) \tag{5}$$

C. Shared GP-LVM

Ek et al. [6] proposed a latent variable model capable of learning a shared manifold that can relate multiple observation spaces of the same underlying phenomenon. Given two datasets, $\mathbf{Y} = [\mathbf{y}_1 \dots \mathbf{y}_N] \in \mathbb{R}^{N\times D_Y}$ and $\mathbf{Z} = [\mathbf{z}_1 \dots \mathbf{z}_N] \in \mathbb{R}^{N\times D_Z}$, the goal is to learn a latent manifold that relates the two observation spaces. An important aspect of the model is that the latent space $\mathbf{X}$ is partitioned into three orthogonal sub spaces $\mathbf{X}_S, \mathbf{X}_Y, \mathbf{X}_Z$, to capture the shared latent features as well as *the features specific to each observation space*. The datasets are assumed to be generated through noise corrupted smooth mappings from the latent spaces to the observation spaces,

$$\mathbf{y}_n = f^Y(\mathbf{x}_n^s, \mathbf{x}_n^y) + \epsilon_n^y, \;\;\; \mathbf{z}_n = f^Z(\mathbf{x}_n^s, \mathbf{x}_n^z) + \epsilon_n^z \tag{6}$$

The latent manifold is learned by maximizing the joint marginal likelihood of the two observation spaces given by,

$$p(\mathbf{Y}, \mathbf{Z}|\mathbf{X}, \Phi) = p(\mathbf{Y}|\mathbf{X}_{S,Y}, \Phi_{S,Y})p(\mathbf{Z}|\mathbf{X}_{S,Z}, \Phi_{S,Z}) \tag{7}$$

This model was used to infer the 3D human pose $\mathbf{Z}$ i.e. pose space from ambiguous 2D silhouette information $\mathbf{Y}$ i.e. feature space. By assuming a *one-to-one* mapping between the pose space and the shared latent space, they applied a back-constraint described in Section II-B on the pose space-latent space mapping, allowing for robust inference of the pose state. The robustness was further improved by adding dynamics to the latent space representation as described in Section II-B. The training for this model involves the learning of the hyper parameters for the GP mapping from each observation space to the shared latent space along with the parameters for the back-constraint and the GP dynamics in the latent space.

We follow a similar modeling approach in this study. The latent space is partitioned into three subspaces as we assume that the depth sensor and motion capture system capture both shared latent features as well as features specific to each sensor. The addition of GP dynamics to the latent space leads to learning of a latent space that captures the sequential state transitions of clothing articles hence capturing the cloth dynamics.

III. Learning Cloth Dynamics

In this section, we present our proposed method for the modeling of cloth dynamics and its application to the estimation of human-cloth relationship. Firstly we explain the formulation of the cloth dynamics model and its motivation. We then describe the representations used for the feature space and pose space in the cloth dynamics model. Finally the different techniques used for the inference of pose state from feature state is explained.

A. Cloth Dynamics Model

The aim is to learn a latent representation $\mathbf{X}_{Y,S,Z} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$ corresponding to an aligned data set of clothing article observations using the depth sensor $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^T$ and motion capture system $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]^T$. The motivation behind this modeling approach as described in Section I is that the motion capture system can provide precise location information of markers placed on the cloth whereas the depth sensor can provide a generalized shape description. By learning a shared latent structure, we are achieving two tasks. Firstly, we are able to learn a latent dynamics model for the clothing article that represents the changes in cloth state during a clothing assistance task. We are also able to learn an indirect mapping from the generic depth sensor information to the specific motion capture marker information, which can be used for constrained cloth state estimation in the absence of motion capture system. We incorporate the model proposed by Ek et al. [6] as described in Section II-C. The dimensionality of the latent space was set as follows: shared space $\mathbf{X}_S \in \mathbb{R}^2$, Y private space $\mathbf{X}_Y \in \mathbb{R}^2$ and Z private space $\mathbf{X}_Z \in \mathbb{R}^2$. The predictive performance of the learned latent structure depends on several factors i.e. i) Pose and Feature space representations, ii) Inference technique used. In the following subsections we describe the approach we used in handling these factors.

B. Pose Space Representation

In this study, we consider the clothing task where the robot has to cloth a mannequin with a T-shirt which is initially on the mannequin's hands. We assume that the details of clothes such as wrinkles are not important to achieve clothing tasks and hence used the low-dimensional topology coordinates [7] to represent the human-cloth relationship. Topology coordinates [7] were formulated for synthesizing human-like motions that involve close interactions. This representation compactly defines the relationship between two curves given in the Cartesian space using three different attributes, i.e. writhe w, center $\mathbf{c} = [c_1 \; c_2]$ and density d. Writhe w measures the total twisting between two curves γ_1, γ_2 by using an approximation of the Gauss Linking Integral (GLI) [8]:

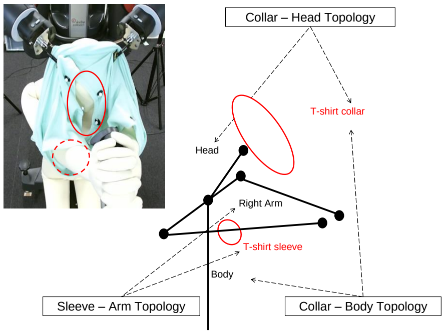

Fig. 2. Representation of human-cloth relationship using topology coordinates

$$GLI(\gamma_1, \gamma_2) = \frac{1}{4\pi} \int_{\gamma_1} \int_{\gamma_2} \frac{(\gamma_1 - \gamma_2) \cdot (d\gamma_1 \times d\gamma_2)}{\|\gamma_1 - \gamma_2\|^3} \quad (8)$$

The center $\mathbf{c}$, composed of two scalars explains the center of twist with respect to each of these lines. The density d represents the relative twisting between the two lines, i.e. which line is twisting around the other. These parameters can be analytically computed by dividing the given curves into chains of small line segments. Further details on the computation of topology coordinates can be found in [7].

The necessary motor skills required for the robot to complete the clothing task are i) to pull the T-shirt collar over the mannequin's head and onto the mannequin's body, ii) to pull the T-shirt sleeves along the mannequin's arm towards its shoulder. The pose space representation is given by considering the writhe and center parameters for four different topologies as shown in Figure 2: i) T-shirt Collar - Mannequin's Head Topology, ii) Collar - Body, iii) Left Sleeve - Left Arm, iv) Right Sleeve - Right Arm thereby forming a twelve dimensional pose space representation $\mathbf{Z} \in \mathbb{R}^{12}$. Each topology represents an interaction between a pair of curves. For example, the Collar - Head topology represents the interaction between the collar curve and the mannequin's head to neck line. To compute the four topologies the following needs to be estimated: T-shirt collar, T-shirt sleeves and the mannequin's posture.

The topology coordinate values were estimated using the Optitrack motion capture system. The setup had six infrared (IR) cameras placed carefully around the experimental setting to avoid occlusion of markers. Six IR markers were attached on the T-shirt collar, three markers on each T-shirt sleeve and five markers on the mannequin respectively to estimate the human-cloth topological relationship. In this

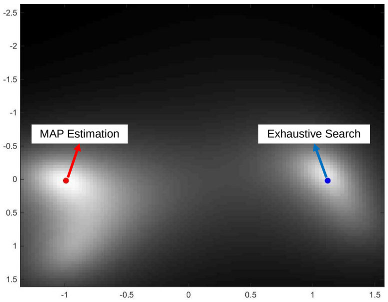

Fig. 3. Estimation of x^*_Z using MAP and Exhaustive search for multi-modal posterior

study, we compared the predictive performance of shared GP-LVM for two pose state representations, either using the 12-dimensional topological coordinate representation or using the full 36-dimensional marker-based representation of the clothing article.

C. Feature Space Representation

In this study, we used the Kinect for Windows V2 to obtain the cloth state which obtains reasonably high resolution depth information using the time-of-flight sensing technology. The depth sensor can be used to obtain point cloud data which can in turn provide a generic shape description of clothing articles. In this section, we describe the estimation of feature space representation from the raw RGB-D data. This can be divided into two stages:

1) Point Cloud Estimation: The first stage is to obtain T-shirt point cloud data from the raw RGB-D data. For this, prior to the actual tracking we perform a HSV-based color calibration to obtain a histogram of hue distribution on the T-shirt. For each input frame, we perform the following steps:

- Color histogram is applied to the input frame to obtain a back-projection image.
- Back-projection image along with a seed T-shirt bounding box is provided as input to the standard CAMshift algorithm [9] where the shift and scaling of the bounding box between frames is estimated.
- Back-projection image within the bounding box corresponds to T-shirt and is applied as a mask to the depth image using which we obtain the raw point cloud of the T-shirt.
- The raw-point point cloud has outliers due to measurement noise which are removed using a statistical based outliers removal technique.
- The point cloud is downsampled using the voxel grid filter and is made zero mean to remove global location information. We set a voxel resolution of $1cm^3$ which downsamples point clouds by a factor of 10 for Kinect V2 resulting in point clouds with 500 to 2000 points.

The image processing functions were implemented using the OpenCV library [10] and the point cloud processing was done using the PCL library [11].

2) Feature Estimation: The extracted T-shirt point cloud can not directly be used as a feature description as it is varying in size and is very high dimensional. Therefore we consider the use of feature descriptors to represent the T-shirt shape information from the point cloud. Ideally we would need a global feature description that is scale, rotation and translation invariant. Alexandre et al. [12] have performed a comparative study on some of the popular feature descriptors to represent 3D point clouds. Based on this study and our criterion, we have shortlisted two suitable descriptors for evaluation:

- **Viewpoint Feature Histogram**: is a feature descriptor proposed by Rusu et al. [13] used for the recognition and pose estimation of rigid objects. VFH is a fixed 308 dimensional feature histogram, consisting of two parts. The first part consists of a histogram over angles made by each point with respect to a fixed viewpoint and the second part consists of histogram over local features of each point with respect to the point cloud centroid.
- **Ensemble of Shape Functions**: is a feature descriptor proposed by Wohlkinger et al. [14] used to represent the underlying shape of a 3D point cloud. ESF is a fixed 640 dimensional feature histogram, consisting of a concatenation of 10 histograms with 64 bins each in them. These histograms are generated by repeated random sampling of 3 points from the point cloud and computing various parameters of the resultant triangles.

D. Inference of Pose State

The latent space in the shared GP-LVM model is dividing into 3 subspaces $\mathbf{X}_{Y,S,Z}$ and so the inference of pose state in shared GP-LVM involves several steps. Given a test depth sensor observation y^*, we first optimize the latent state $x^*_{Y,S}$ corresponding to this observation, then predict corresponding x^*_Z state and finally compute the output pose state z^*. The shared GP-LVM framework provides methods for all these steps except for the prediction of x^*_Z. The problem is that X_Z subspace is orthogonal to other subspaces and so y^* can not be used to provide any further information for inferring x^*_Z. However, we can obtain a likelihood for generating z^* over X_Z subspace given the position $[x^*_Z, x^*_S]$. We consider two techniques for estimating x^*_Z from the likelihood estimates:

- **MAP Estimation**: In this method x^*_Z is given by maximum-a-posteriori estimation. This method provides a good estimate for z^* given that the likelihood is unimodal. However, this model is not effective for multi-modal likelihoods which arises from ambiguity in the pose state inference.
- **Exhaustive Search**: In this method x^*_z is given by iterating through the $\mathbf{X}_Z$ space and compute z^* to find the point that minimizes the error with respect to the ground truth z^T. This inference method is only used as a metric to evaluate the best possible performance of the trained model and can not be used in real-time as it relies on the knowledge of knowing the ground truth pose-state.

Figure 3 illustrates the difference between these inference techniques where the posterior over $\mathbf{X}_Z$ is multi-modal and

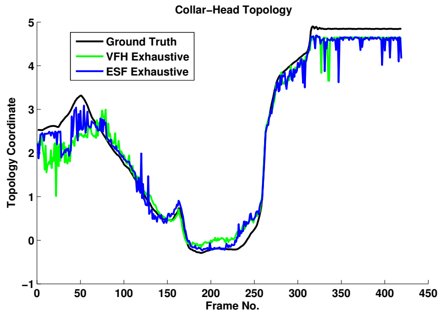

Fig. 4. Collar-Head Topology Coordinates estimated by ESF, VFH models using the exhaustive inference technique.

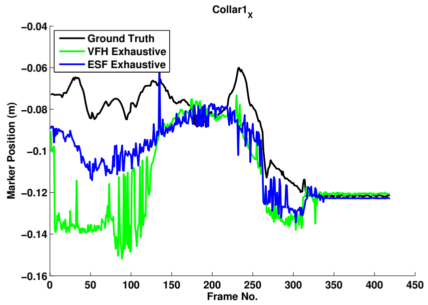

Fig. 5. Collar marker X-coordinate estimated by ESF, VFH models using the exhaustive inference technique.

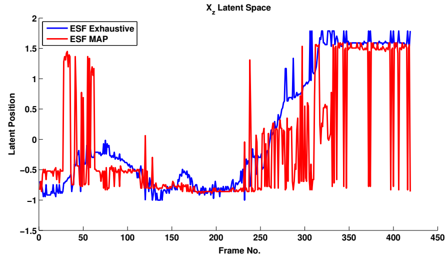

Fig. 6. Latent positions estimated for X_Z latent space using MAP and Exhaustive search inference techniques

TABLE I
ACCURACY OF SHARED GP-LVM MODELS IN TERMS OF RMS ERRORS

Model Type	Train		Test	
	MAP	Exhaustive	MAP	Exhaustive
VFH → Marker (mm)	39.908	2.694	111.260	47.475
ESF → Marker (mm)	25.353	2.301	95.014	39.420
VFH → TopCoord	0.271	0.017	1.201	0.225
ESF → TopCoord	**0.1229**	**0.018**	**0.778**	**0.235**

the MAP estimate and exhaustive search provide estimates that lie on two different modes of the posterior. The MAP and Exhaustive methods can be considered as upper and lower bounds for the performance error of the trained dynamics model as shown by the results in Table I.

IV. EVALUATION

In this section, we describe the experiments conducted to evaluate the performance of our proposed framework. Section IV-A shows the results of applying shared GP-LVM to the clothing assistance framework. Section IV-B includes discussion about the obtained results.

A. Experimental Results

The performance of the proposed method was evaluated by collecting clothing trials with simultaneous observation of the T-shirt state using both the depth sensor and motion capture system. The observations were temporally aligned through socket interfacing which ensures point-to-point correspondences in the training phase. The observations were also spatially aligned by performing an absolute orientation calibration between the motion capture system and the depth sensor. The method proposed by Umeyama [15] was used to compute the transformation between the two reference frames.

Six clothing trials were collected for six different postures of the mannequin obtained by varying the head inclination ($\{30^o, 45^o\}$) and the shoulder inclination ($\{70^o, 75^o, 80^o\}$) angles. These angles were measured with respect to the positive Z-axis which is normal to the ground plane. The motivation behind creating such a dataset was that the force applied by the robot changes largely between different postures thereby imparting a variation in dynamics between the clothing trials.

For the evaluation, we choose one of the clothing trials (Head Angle:30^o, Shoulder Angle:75^o) and trained four separate shared GP-LVM models with the Kinect feature space representations being Ensemble of Shape Functions (ESF), Viewpoint Feature Histogram (VFH) and the Motion capture pose space representations being Topology Coordinates (TopCoord), Marker Cartesian Positions (Marker). For each of these trained models, the remaining five clothing trials were taken as test data to evaluate the adaptability of the learned dynamics model. For the test data, the Kinect feature state were provided as input to the model and the pose state was inferred. The performance was evaluated by computing the RMS error between the inferred state and the ground truth pose state values.

We evaluated the performance for both the inference techniques discussed in Section III-D which can be considered as the upper and lower bounds for the performance error. The results for this comparison are shown in Table I. It can be seen that under the exhaustive search, both the feature descriptors have almost similar performance. However for the MAP estimation, the ESF descriptor provides better estimates compared to the VFH descriptor. Figure 4 shows the estimated collar-head topology coordinate values for one of the test clothing trials when exhaustive inference technique is used, which demonstrates that potentially the learned shared latent structure can provide accurate topology coordinate estimates. Figure 5 shows the estimated X-coordinate for one of the T-shirt collar markers. This result shows that even under exhaustive search, the learned model is not able to accurately predict marker positions in comparison to the estimation of topology coordinates. Figure 6 shows the latent positions x_z^* estimated for a test clothing trial. It can be seen that under exhaustive search the latent coordinates have smooth transitions compared to MAP estimation where the latent position jumps between different modes.

B. Discussion

The experimental results shown in the previous section shows that shared GP-LVM is a good candidate for learning the dynamics of clothing articles. However, it can also be seen that the performance depends on three key factors:

- Performance difference between ESF and VFH descriptors shows that the feature space representation plays an important role. We need to choose a descriptor that describes the overall shape of the clothing article and needs to be robust to outliers.
- Performance difference between topology coordinate and marker position representations show that the pose space representation also greatly affects the performance. Our interpretation for the difference is that topology coordinates capture the motion profile of cloth extremities which is consistent across different settings and can be learned by shared GP-LVM. On the other hand, there can be larger variations in a single marker motion across different settings and the learned model can not generalize to such motion.
- The exhaustive inference technique shows that *potentially* the shared latent structure can provide accurate topology coordinate estimates. However, there is a need for a robust inference technique. The results in Figure 6 indicate that a possible inference candidate needs to provide latent space estimates that rely on temporal smoothness i.e. current latent space estimates also depend on previous latent space estimates along with current feature space observation.

V. Conclusion

Robust estimation of human-cloth relationship plays a crucial role for the implementation of robotic clothing assistance. An approach to this estimation is to learn a dynamics model of the human-cloth relationship and use this model as a prior for robust tracking in real-time. In this study, we learn the underlying cloth dynamics using the shared Gaussian Process Latent Variable Model and by incorporating accurate state information obtained from the motion capture system into the dynamics model. Shared GP-LVM provides a principled probabilistic framework to infer the accurate cloth state from the noisy depth sensor readings. The experimental results show that shared GP-LVM is able to learn reliable motion models of the T-shirt state for robotic clothing assistance tasks. We also demonstrated three key factors that contribute to the performance of the trained dynamics model.

The advantage of using GP-LVM is that a corresponding latent space manifold can be learned for any representation used in the observation spaces. Based on this flexibility, our future work will be to learn dynamics models for pose space that incorporates T-shirt state as well as mannequin's posture, proprioceptive information of the robot. The dynamics model can be used as a prior for robust tracking of the human-cloth relationship with depth sensors and thereby providing a feasible framework for efficient learning of motor-skills.

Acknowledgement

This work was partly supported by the Grant-in-Aid for Scientific Research from Japan Society for the Promotion of Science (No. 23240028).

References

[1] K. Nishanth, T. Tamei, T. Matsubara and T. Shibata, "Real-time Estimation of Human-Cloth Topological Relationship using Depth Sensor for Robotic Clothing Assistance," in *Proc. of the 23rd IEEE Intl. Symp. on Human Robot Interactive Communication,* 2014.

[2] N. D. Lawrence, "Gaussian Process Latent Variable Models for Visualization of High Dimensional Data," in *Advances in Neural Information Processing Systems 16,* 2004.

[3] M. E. Tipping, and C. M. Bishop, "Probabilistic principal component analysis," in *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 1999.

[4] J. Wang, A. Hertzmann, and D. M. Blei, "Gaussian Process Dynamical Models" in *Advances in Neural Information Processing Systems,* 2005.

[5] N. D. Lawrence, and J. Quinonero-Candela. "Local distance preservation in the GP-LVM through back constraints," in *Proc. of the 23rd ACM Intl. Conf. on Machine learning,* 2006.

[6] C. H. Ek, J. Rihan, P. H. Torr, G. Rogez, and N. D. Lawrence, "Ambiguity modeling in latent spaces," in *Machine Learning for Multimodal Interaction, Springer Berlin Heidelberg,* 2008.

[7] E. S. L. Ho, T. Komura, "Character motion synthesis by topology coordinates," in *Proc. of EUROGRAPHICS2009,* 2009.

[8] W. Pohl, "The self-linking number of a closed space curve,", in *Journal of Mathematics and Mechanics,* 1968.

[9] G. R. Bradski, "Computer vision face tracking for use in a perceptual user interface" 1998.

[10] G. Bradski, "The opencv library," in *Doctor Dobbs Journal 25.11,* 2000.

[11] R. B. Rusu, and S. Cousins, "3d is here: Point cloud library (pcl)," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA),* 2011.

[12] A. L. Alexandre, "3D descriptors for object and category recognition: a comparative evaluation," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS),* 2012.

[13] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu, in "Fast 3d recognition and pose using the viewpoint feature histogram," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS),* 2010.

[14] W. Wohlkinger, and M. Vincze, "Ensemble of shape functions for 3d object classification," in *Proc. of the IEEE Int. Conf. on Robotics and Biomimetics (ROBIO),* 2011.

[15] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," in *IEEE Transactions on pattern analysis and machine intelligence,* 1991.

[16] M. Cusumano-Towner, A. Singh, S. Miller, J.F. O'Brien, P. Abbeel,"Bringing clothing into desired configurations with limited perception," in *Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA),* 2011.

[17] A. Ramisa, G. Alenya, F. Moreno-Noguer, C. Torras, "Using depth and appearance features for informed robot grasping of highly wrinkled clothes," in *Proc of. IEEE Int. Conf. on Robotics and Automation (ICRA),* 2012.

[18] S. Miller, M. Fritz, T. Darrell, P. Abbeel, "Parametrized shape models for clothing," in *Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA),* 2011.

[19] B. Balaguer, S. Carpin, "Combining imitation and reinforcement learning to fold deformable planar objects," in *Proc. of IEEE/RSJ Int. Conf. on Robots and Systems (IROS),* 2011.

[20] Y. Kita, T. Ueshiba, E. S. Neo, N. Kita, "Clothes state recognition using 3D observed data," in *Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA),* 2009.