Context and Observation Driven Latent Variable Model for Human Pose Estimation

Abhinav Gupta[1], Trista Chen[2], Francine Chen[2], Don Kimber[2] and Larry S Davis[1]
[1] University of Maryland, College Park, MD
[2] FX Palo Alto Research Center, Palo Alto, CA
agupta@cs.umd.edu, {tchen, chen, kimber}@fxpal.com, lsd@cs.umd.edu

Abstract

Current approaches to pose estimation and tracking can be classified into two categories: generative and discriminative. While generative approaches can accurately determine human pose from image observations, they are computationally expensive due to search in the high dimensional human pose space. On the other hand, discriminative approaches do not generalize well, but are computationally efficient. We present a hybrid model that combines the strengths of the two in an integrated learning and inference framework. We extend the Gaussian process latent variable model (GPLVM) to include an embedding from observation space (the space of image features) to the latent space. GPLVM is a generative model, but the inclusion of this mapping provides a discriminative component, making the model observation driven. Observation Driven GPLVM (OD-GPLVM) not only provides a faster inference approach, but also more accurate estimates (compared to GPLVM) in cases where dynamics are not sufficient for the initialization of search in the latent space.

We also extend OD-GPLVM to learn and estimate poses from parameterized actions/gestures. Parameterized gestures are actions which exhibit large systematic variation in joint angle space for different instances due to difference in contextual variables. For example, the joint angles in a forehand tennis shot are function of the height of the ball (Figure 2). We learn these systematic variations as a function of the contextual variables. We then present an approach to use information from scene/objects to provide context for human pose estimation for such parameterized actions.

1. Introduction

Human pose tracking is a challenging problem because of occlusion, a high dimensional search space and high variability in people's appearance due to shape and clothing variations. There is a wide range of approaches to human pose tracking which can be broadly divided into two categories:

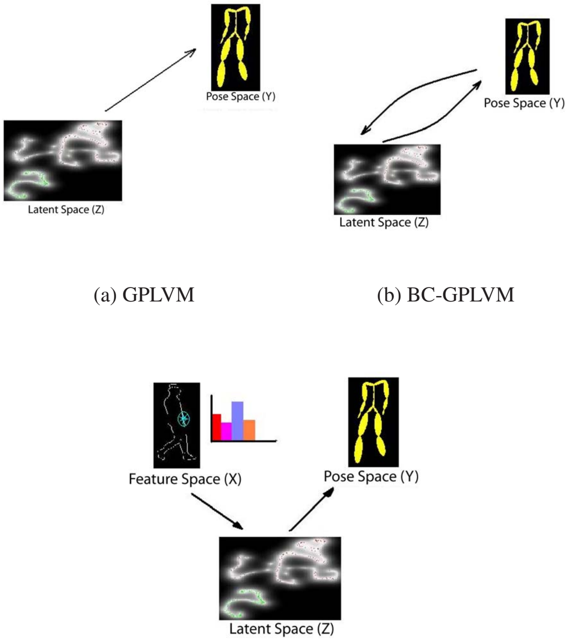
- **Discriminative Approaches**: Discriminative methods employ a parametric model of posterior probabilities of pose and learn the parameters from the training data. The parametric model is generally an ambiguous mapping from observation space to pose space.
- **Generative Approaches**: Generative methods model the joint probability distribution of hypothesis and observation using class conditional densities (image likelihoods $P(I|Y)$) and class prior probabilities ($P(Y)$). Such approaches search the pose-space to find the pose that best explains the image observations.

Discriminative approaches involve learning the mapping from feature/observation space ($\mathcal{X}$) to the pose space ($\mathcal{Y}$). This mapping ($\phi : \mathcal{X} \rightarrow \mathcal{Y}$) may not be simple because it is generally ambiguous (two different poses can look similar in some views). Due to this inherent ambiguity, multiple functions or a mixture of experts model have been used for representing the mapping from $\mathcal{X}$ to $\mathcal{Y}$. On the other hand, the inverse problem of generating image observations given a pose vector is a well defined problem. One can easily build a mapping from pose space to observation space which can be used as the likelihood model in the generative approach. Discriminative approaches are, however, faster compared to generative approaches, which require search in the high-dimensional pose space.

While either searching or learning a prior model in a high dimensional space is expensive, dimensionality reduction techniques can be used to embed the high-dimensional pose space in a lower dimensional manifold. The Gaussian process latent variable model (GPLVM) [13] is a generative approach which models the pose-configuration space ($\mathcal{Y}$) as low dimensional manifold and the search for the best configuration is performed in this low-dimensional latent space ($\mathcal{Z}$). GPLVM is a smooth[1] mapping from the latent space to the pose space. It keeps latent points far apart if their corresponding poses lie far apart. An extension to GPLVM,

[1] the points in latent space which are 'close' will be mapped to points in pose space which are 'close'.

978-1-4244-2243-2/08/$25.00 ©2008 IEEE



(a) GPLVM

(b) BC-GPLVM

(c) OD-GPLVM

Figure 1. Comparison of mappings in the three gaussian models.

called Back Constrained GPLVM (BC-GPLVM), was introduced in [14]. By having an additional inverse mapping from the pose space to the latent space, BC-GPLVM also preserves local distances in the pose space.

Both GPLVM and BC-GPLVM determine the low dimensional embedding of the pose space regardless of the distances between poses in the observation/feature space. It is important to consider distances in observation space since the cost function that drives the search for the pose is based on distances and gradients in the observation space. We introduce observation driven GPLVM (OD-GPLVM), which has a smooth mapping from the observation space to the latent space in addition to the mapping from the latent space to the pose space (See Figure 1). OD-GPLVM is a hybrid model that combines the strengths of both generative and discriminative models. The mapping from observation space to latent space allows us to estimate the latent positions directly from observations. The best pose can then be searched for in the neighborhood of the estimated point in latent space. Thus, OD-GPLVM has better initialization based on observations and is not limited to motion dynamics within the training data. We also extend the Gaussian Process Dynamical Model (GPDM) [31] in a similar manner to include an embedding from joint space ($\mathcal{X} \times \mathcal{Z}$) to the latent space.

While approaches such as GPLVM and OD-GPLVM can be used to find a low-dimensional embedding of pose space for an action, it has been observed that such embeddings often model multiple instances of the same action as very different trajectories in the latent space. Such a variation in latent/joint-angle spaces is either due to differences in styles or environmental conditions (See Figure 2). We describe how to extend our approach to model systematic variations in pose-space for parameterized actions. In addition to using features from human silhouettes, our model also uses contextual information from the scene and objects to estimate human pose.

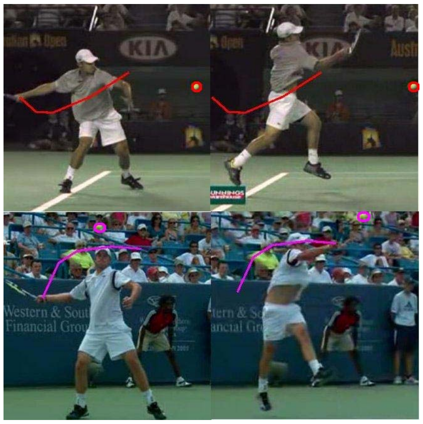


Figure 2. Parameterized Actions: A tennis forehand shot is an example of a parameterized action. The trajectory in pose space is a function of ball height(as shown in the example) and the direction the ball is to be hit. The parameter can be determined not only using the pose observations, but also the ball position and opponent's position (Contextual Features)

2. Related Work

Human pose estimation has been studied extensively in computer vision. Generative approaches [8, 23] search in the high dimensional pose space to determine the pose which best explains image observations. This is generally posed as a non-linear optimization problem. Given an initial estimate, approaches such as gradient descent can be used for optimization. However, such approaches are easily trapped in local minima. Approaches such as particle filtering [11] have been used to overcome this problem. However, particle filtering fails to scale well in high dimensional spaces, such as human pose, because of the large number of particles required for effective representation.

A few attempts have been made to reduce the high-dimensionality of pose space using principal component analysis [22]. Linear subspace models are, however, inappropriate for modeling the space of human poses due to its underlying non-linearity. Other approaches, such as [10], either tend to overfit the data or require large amounts of data for training. One can, instead, use non-linear dimensionality reduction approaches such as Isomaps [26] or LLE (local linear embedding) [20, 4]. These approaches, however, lack mappings from the embedded space to the data space, which is important for a generative search framework.

Lawrence et al. [13] introduced GPLVM, which not only determines a low dimensional embedding but also a mapping from this embedding (latent space) to pose space. Urtasun et al. [29] proposed an approach to estimate human pose using SGPLVM [6], where each input dimension is scaled independently to account for different variances of different data dimensions. Other approaches such as GPDM [28], BC-GPLVM [9], LL-GPLVM [30], SLVM [12] and LELVM [17] have also been used for human body tracking. All these approaches use either deterministic optimization [29] or particle filtering to search for the best pose [16]. While the initialization approach based on search in latent space proposed in [29] is very expensive, other initialization approaches such as in [28] rely too heavily on learned dynamics. Our approach provides an effective, more computationally efficient method for pose estimation and balances the utilization of image features and dynamics. It computes the embedding by considering image observations in conjunction with pose data. This is achieved by adding a mapping [2] from observation space to latent space. This mapping provides natural initialization points where features from observations are used to obtain the starting point for search in the latent space. Thus, our approach avoids expensive initialization as well as unreliable dynamics.

Some approaches such as [3, 21] use a shared latent space for observation and pose. The mapping in such a case is from latent space to observation space. The mapping used in our approach, from observation space to latent space, is significant for two reasons: (1) Such a mapping is a prime requirement for the discriminative flavor which provides faster speeds and has been used in [12]. (2) Our mapping ensures that two points close in observation space will be close in latent space whereas in [3] the other mapping ensures two points far in observation space will be far in latent space(which was already true since they were far in pose and hence already far in latent space).

The joint angle trajectories in many actions show systematic variations with respect to environmental variables. Wilson et al. [33] introduced an approach to represent and recognize parameterized actions that exhibit systematic spatial variations. We present an approach to human pose tracking by modeling the variation in dynamics with respect to location of an object being acted on and other environmental variables. Such variations cannot be modeled as stylistic variations [5, 32], since they are dependent on external contextual variables and their variational magnitudes are larger. Urtasun et al. [27] use a golf club tracker to provide cues for human hand tracking. Their approach is complementary to ours; they use the golf club as a source of discriminative features to track the hand and estimate its 3D locations. Our approach, on the other hand, models the variations in human pose with respect to scene and object features. While contextual information has been used to improve object and action recognition [7, 18, 19], to the best of our knowledge, this is the first attempt to apply contextual information to human pose estimation.

3. Observation Driven GPLVM

GPLVM is a probabilistic, non-linear, latent variable model. It constructs a smooth mapping from latent space to pose space; hence, pose configuration can be recovered if the corresponding latent position is known. While GPLVM has been used for pose-tracking, it suffers from the drawback that two points may be far from each other in latent space even though the observations/poses are very similar. Preservation of local distance in observation space is important for gradient-descent based approaches as it leads to smoother cost functions. It is also important for sampling based approaches as it brings two points similar in observations within sampling range of each other.

Our proposed model, OD-GPLVM, overcomes this by creating two smooth mappings, one from observation space to latent space and the other from latent space to pose space. Such a mapping pair offers two benefits: (a) It provides a better and natural initialization for search in the latent space. The mapping from observation space to latent space provides the starting point for search in latent space. This initialization approach is more effective than the one employed in GPLVM or BC-GPLVM because it is fast and based on observation, rather than on smoothness or a constraint of "small" motion between frames. (b) Such a mapping not only preserves local distances in pose space but also preserves local distances in observation space. Therefore, two latent points which generate similar observations tend to lie close to each other.

Let $Y = [y_1, .., y_N]^T$ be the poses of the training dataset. Similarly, let $X = [x_1, .., x_N]$ represent the observations in feature space and $Z = [z_1, .., z_N]$ be the corresponding positions in the latent space. Given, a training dataset (X, Y) we want to compute the model $M = \{\{z_i\}, \Phi_{L\mapsto P}, \Phi_{O\mapsto L}\}$, where $\Phi_{L\mapsto P}$ and $\Phi_{O\mapsto L}$ are the parameters of the two mappings from latent space to pose space and observation space to latent space, respectively. The posterior of M, $P(M|Y, X)$, can be decomposed using Bayes rule as

$$\begin{aligned} P(M|Y,X) &\propto P(Y|M,X)P(M|X) \\ &= P(Y|M)P(M|X) \\ &= P(Y|Z,\Phi_{L\mapsto P})P(Z|X,\Phi_{O\rightarrow L})P(\Phi_{O\mapsto L}|X) \end{aligned}$$

Under the Gaussian process model, the conditional density for the data is multivariate Gaussian and can be written

[2] While approaches such as [12] also learn a mapping from observation space to latent space after learning the embedding, their mapping is generally discontinuous because the embedding is learned independent of distances in observation space.

as

$$P(Y|Z,\Phi_{L\mapsto P}) = \frac{1}{\sqrt{2\pi^{ND}|K_Z|^D}} exp(-\frac{1}{2}tr(K_Z^{-1}YY^T)) \tag{1}$$

where K_Z is the kernel matrix and D is the dimensionality of the pose space. The elements of the kernel matrix are given by a kernel function, $K_{Z_{ij}} = k(z_i, z_j)$. We use a Radial Basis Function (RBF) based kernel function of the form:

$$k(z_i,z_j) = \alpha_\Phi exp(\frac{-\gamma_\Phi}{2}(z_i - z_j)w_\Phi(z_i - z_j)^T) + \beta_\Phi \delta_{z_i,z_j} \tag{2}$$

where δ is the Kronecker delta function. Similarly, the conditional density $P(Z|X,\Theta)$ can also be broken down as

$$P(Z|X,\Phi_{O\mapsto L}) = \frac{1}{\sqrt{2\pi^{NQ}|K_X|^Q}} exp(-\frac{1}{2}tr(K_X^{-1}ZZ^T)) \tag{3}$$

where K_X is the kernel matrix and Q is the dimensionality of the latent space. The elements of the kernel matrix are given by a kernel function, $K_{X_{ij}} = k(x_i, x_j)$. We again use RBF kernel given by:

$$k(x_i,x_j) = \tilde{\alpha_\Phi} exp(\frac{-\tilde{\gamma_\Phi}}{2}(x_i - x_j)\tilde{w_\Phi}(x_i - x_j)^T) + \tilde{\beta_\Phi} \delta_{x_i,x_j} \tag{4}$$

We assume a uniform prior on the parameters of the mapping from $\mathcal{X} \rightarrow \mathcal{Z}$. Therefore, the log posterior of M, L, is given by

$$\begin{aligned} L &= \frac{-(D+Q)(N)}{2}ln(2\pi) - \frac{D}{2}ln|K_Z| - \frac{Q}{2}ln|K_X|.. \\ .. & -\frac{1}{2}tr(K_Z^{-1}YY^T) - \frac{1}{2}tr(K_X^{-1}ZZ^T) \end{aligned} \tag{5}$$

We need to optimise the likelihood with respect to the latent positions and various parameters. We compute the gradients of (5) with respect to Z using the chain rule

$$\frac{\partial L}{\partial Z} = -K_x^{-1}Z + (\frac{1}{2}K_Z^{-1}Y^TYK_Z^{-1} - \frac{DK_Z^{-1}}{2})\frac{\partial K_Z}{\partial Z} \tag{6}$$

We optimize (5) using a non-linear optimizer such as scaled conjugate gradient(SCG). The optimization is performed similarly to the optimization in [14]. For initialization, we obtain Z using principal component analysis (PCA). We then use an iterative approach where the parameters and latent positions are updated using the gradients.

3.1. Inference Process

GPLVM is a generative model, while the mapping from observation space to latent space provides a discriminative flavor to the model. To infer a pose in a frame, we first extract image features. The features are based on shape context histograms and are similar to those used in [1].

Based on the features, we use the discriminative mapping to obtain the proposal distribution $q(z|x)$. This proposal distribution is used to obtain the samples in the latent space. Sampling is done using the importance sampling procedure. Samples are evaluated based on posterior probabilities defined by:

$$\begin{aligned} P(y,z|I,\mathcal{M}) &\propto P(I|y,z,\mathcal{M})P(y,z|\mathcal{M}) \\ &= P(I|y)P(y|z,\mathcal{M})P(z|\mathcal{M}) \end{aligned} \tag{7}$$

The first term in the equation is the image likelihood given a hypothesized pose. We use an edge based likelihood model which uses a distance transform, similar to one proposed in [15]. The second term represents the probability of the hypothesized pose given a hypothesized latent position. From [13], we know $P(y|z,\mathcal{M})$ is given by $\mathcal{N}(y, f(z), \sigma(z))$ where:

$$\begin{aligned} f(z) &= \mu + Y^T K_Z^{-1} k(z) \\ \sigma^2(z) &= k(z,z) - k(z)^T K_Z^{-1} k(z) \end{aligned}$$

3.2. Using Multiple Regressors

The mapping from observation space to latent space is generally ambiguous. Many pose configurations lead to similar observations and hence the inherent ambiguity. Such ambiguity generally disappears in the tracking framework due to temporal consistency constraints (Section 3.3). A mixture of experts regressors [24] can be used to overcome this problem for static image analysis. In this modified model, the training process is modified to an EM-based approach similar to [25].

3.3. Extension to Tracking

GPDM is a latent variable model which consists of a low dimensional latent space, a mapping from latent space to data space and a dynamical model in the latent space. Observation driven GPLVM also provides a natural extension to GPDM. Instead of only having a mapping from the observation space $\mathcal{X}$ to pose space, we also include a mapping, $\psi : \mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{Z}$. In a tracking framework, the latent position at time t is given by

$$z^t = \psi(x^t, z^{t-1}) + noise \tag{8}$$

Using such a mapping, we can again regress to the current latent position using current observations and the previous frame's latent position. The new log-posterior function, L^*, is similar to L except that K_X is replaced by K_{XZ} and each element is given by

$$\begin{aligned} k(x_i,z_i,x_j,z_j) &= \alpha exp(\frac{-\gamma}{2}((z_i - z_j)w(z_i - z_j)^T.. \\ .. & +(x_i - x_j)w'(x_i - x_j)^T)) + \beta\delta_{XZ} \end{aligned}$$

The new gradient with respect to Z can be computed as:

$$\frac{\partial L^*}{\partial Z} = (\frac{1}{2}K_{XZ}^{-1}Z^TZK_{XZ}^{-1} - \frac{QK_{XZ}^{-1}}{2})\frac{\partial K_{XZ}}{\partial Z} + \frac{\partial L}{\partial Z} \tag{9}$$

The inference procedure in the tracking framework is similar to the inference process explained previously. We obtain the proposal distribution using the current observations x^t and previous frame latent position z^{t-1}. Based on this proposal distribution, the samples which are evaluated are constructed using importance sampling.

3.4. Comparison With Back-constrained GPLVM

Lawrence et. al [14] introduced BC-GPLVM as a variant of GPLVM which preserves local distances of pose space under dimensionality reduction. While GPLVM tries to preserve dissimilarity (no two points 'far apart' in pose space can lie 'close together' in latent space), there is nothing that prevents two points lying close in the pose space from being far apart in the latent space. BC-GPLVM tackles this problem by having another smooth mapping from pose space to latent space. Therefore, by creating two smooth mappings local distances are preserved in BC-GPLVM.

On the other hand, by taking into consideration the observation space during the dimensionality reduction and having a smooth mapping from observation-space to latent-space, OD-GPLVM preserves local distances implicitly. Two points which are close in the pose space should lie close in the observation space as well, and by having a smooth mapping from observation space to latent space, it is ensured that the two points lie close in latent space as well. Thus, while BC-GPLVM preserves local distances of pose space, OD-GPLVM preserves local distances of both pose space and observation space.

4. Using Context for Pose Estimation

OD-GPLVM can be used to learn an activity manifold for the pose estimation problem. Consider an activity like sitting (See Figure 3). The execution of such an activity and the trajectory in joint angle space is determined by a few contextual variables (the height of the surface to sit on, in this case). Many activities show a systematic variation in their execution with respect to external variables such as surface height. Using non-linear dimensionality reduction techniques is not appropriate without modeling these variations. We extend our approach to model these variations and use observations/features from the scene and objects to estimate the contextual variables, followed by human pose estimation. For example, in the case shown in Figure 3, using the features from the chair/stool can be used to provide strong cues on the height parameter. Using the estimated height and current pose observations, one can predict the possible latent point in the latent space.

4.1. The Model

We need to model the variations in pose-space as a function of a contextual variable. While one can learn multiple models for different values of the contextual variables, we use a single latent space to represent all the possible poses for different values of contextual variables. We use OD-GPDM with multiple mappings from observation space to latent space for modeling the variations in parameterized activity. A mapping from the observation space to latent space is learned from an instance of the activity for a certain value of the variable from the training dataset. For example, if we have a training dataset of three possible sitting heights with the sitting objects being chair, step-stool and ground, we will learn three mappings from observation space to latent space, one for each height. Only a single mapping from latent space to the pose space is used.

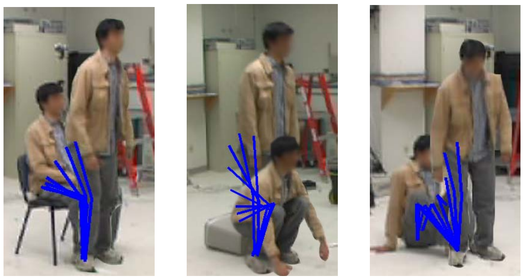

(a)Chair (b) StepStool (c) Ground

Figure 3. Joint angle variations for different parameter values(heights of sitting surfaces).

Figure 4 shows the graphical representation of the model used for inference. Let x_c represent contextual features and x represent shape-context features from the silhouette. We want to obtain an estimate of the probability distribution $P(z|x, x_c, M)$. This distribution can then be used for importance sampling and to evaluate the samples using the equations described in section 3.1. Let θ represent the contextual variables which are used to parameterize the activity (for example, in case of sitting θ corresponds to the height of the sitting surface). We can then compute $P(z|x, x_c, M)$ as

$$P(z|x, x_c, M) = \sum_{\theta} P(z|\theta, x, M)P(\theta|x, x_c) \quad (10)$$

$$= \sum_{\theta} P(z|x, M_\theta)P(\theta|x, x_c) \quad (11)$$

where M_θ corresponds to the mapping for a particular value of θ. We use a discrete representation of the variable θ based on the instances used to learn the activity.

Contextual features x_c are extracted from regions where the objects are present. Human pose provides a prior on the location of an object being interacted with. For example, in the case of sitting, the location of the hip and knee joints provide priors on the location of the surface on which the person will sit. So, this leads us to a chicken-egg problem, where the pose of a person can be used to extract features x_c and these features can be used to estimate the pose. We use an iterative approach, where we re-compute the distribution $P(z|x, x_c, M)$ at every iteration to update the possible pose.

We use the same SCG method for learning the model as before. However, since there are multiple mappings from observation space to latent space, the log-posterior function has terms for all mappings.

5. Experimental Results

We performed a series of experiments to evaluate our algorithms. In the first set of experiments, we compared OD-

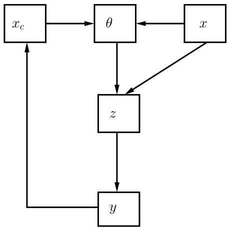


Figure 4. The Graphical Model for Inference

GPDM to GPLVM and GPDM. In the second set of experiments, we trained our model for sitting, a parameterized activity, and compare the performance of our algorithm with and without the use of contextual information.

5.1. Observation Driven Models

We used the CMU-Mocap datset [2] for evaluating OD-GPDM. Experiments were performed to evaluate the algorithm's performance on three activities: jumping-jack, walking and climbing a ladder. Training requires both joint-angles and the silhouette observations. In a few cases where the observations were not provided in the dataset, animation software was used to obtain the silhouettes.

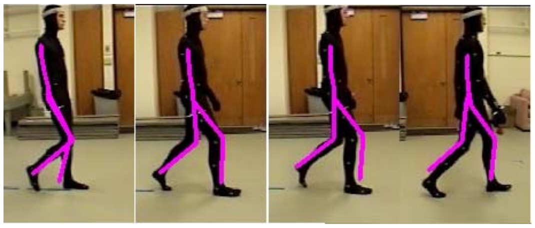
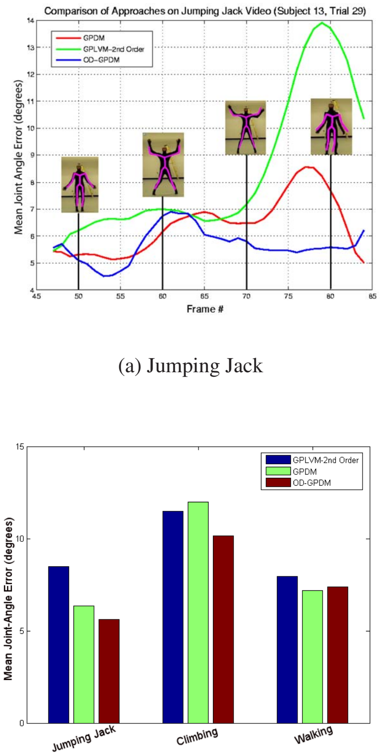


Figure 6. Pose Tracking Results on Walking Activity using OD-GPDM (Subject=35, Instance=05).

Figures 5 and 6 show the performance of OD-GPDM on the jumping jack and walking activities. For the walking activity, only the joint angles corresponding to the torso and lower body are estimated. In all experiments, the tracking algorithm was initialized using the closest observation in the training dataset.

Quantitative: We compared the performance of OD-GPDM to two tracking approaches: GPLVM with second order dynamics [29] and GPDM [28]. The mean joint angle error was calculated using the ground truth data. Figure 7(a) compares the performance of OD-GPDM for the jumping jack activity. While GPLVM and GPDM suffer from an accumulation of tracking errors, OD-GPDM does not have that problem due to less reliance on dynamics. Figure 7(b) shows the mean error for three different activities. While OD-GPDM outperforms GPLVM and GPDM in the jumping jack and climbing activities, the performance is similar for all three in the walking activity. OD-GPDM is computationally fast (upto 5fps on a Pentium 4) since the initialization of search is obtained using the mapping from observation space to latent space.



(a) Jumping Jack

(b) Different Action Classes

Figure 7. Quantitative Evaluation: Comparison of OD-GPDM with GPLVM (2nd Order Dynamics) and GPDM. (a) Frame-by-frame comparison (b) Comparison for three activities. OD-GPDM outperforms both the algorithms in the jumping jack activity.

5.2. Context based GP Models

We trained our context driven model for the sitting activity. As shown in the example of Figure 3, there are systematic variations in trajectories in joint-angles and latent space for different heights of the sitting surfaces. The training dataset for sitting was taken from the CMU-Mocap data and included instances with four different seat heights. Figure 8 shows the latent space after training our model. The four trajectories, shown by different colored points, correspond to four different instances of sitting.

For testing, videos were obtained of subjects sitting on chair, stepstool and the ground. Figure 9 shows the performance of context driven OD-GPDM for subject 1. Ground truth was manually hand-labeled to compare the performance of OD-GPDM with and without using contextual information(Figure 10). It can be seen that use of contextual information improves the performance of the algorithm.

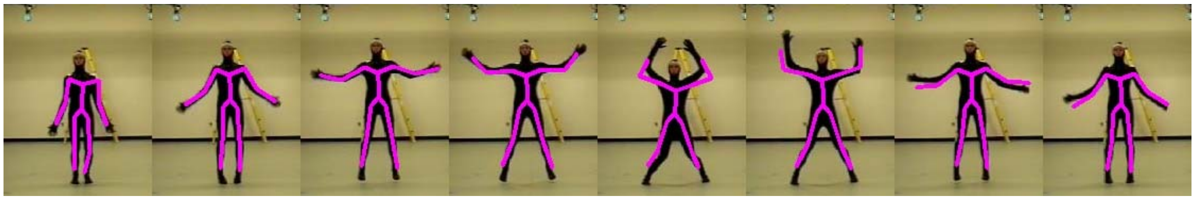


Figure 5. Pose Tracking Results on Jumping Jack activity using OD-GPDM (Subject=13, Instance=29)

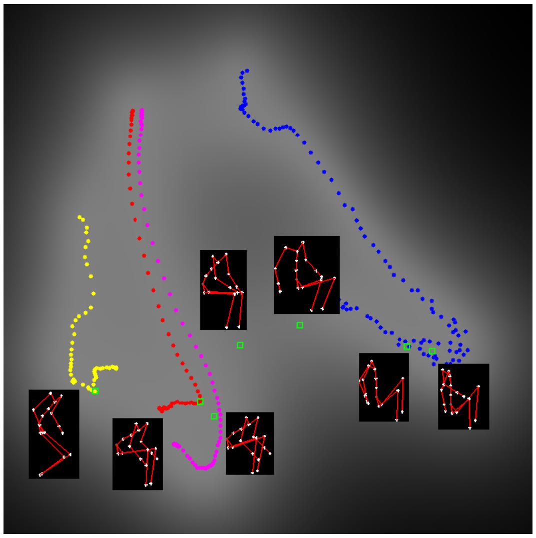
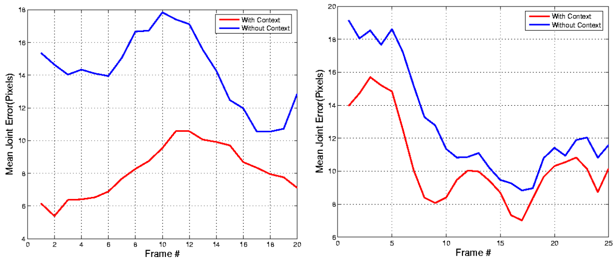


Figure 8. Parameterized Actions: Latent Space for Sitting Action. The four trajectories correspond to sitting on surfaces of different heights. Yellow corresponds to sitting on a bar stool, Red corresponds to sitting on a chair, Magenta corresponds to a sitting on a stepstool and Blue corresponds to sitting on the ground. Our model was able to generalize pose variations over different surfaces, the poses corresponding to higher sitting surfaces occur on the left and the poses for lower sitting surfaces on the right.

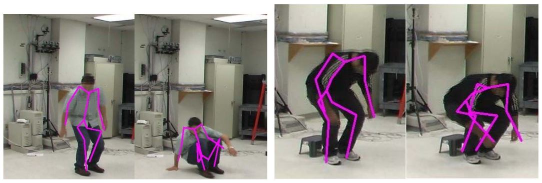
Figure 11(a) and (b) shows the performance on other subjects with sitting surfaces being the ground and step stool respectively.



(a) Step Stool

(b) Chair

Figure 10. Quantitative Evaluation: Comparison of OD-GPDM with and without contextual information on Subject 1.



(a) Ground

(b) Step Stool

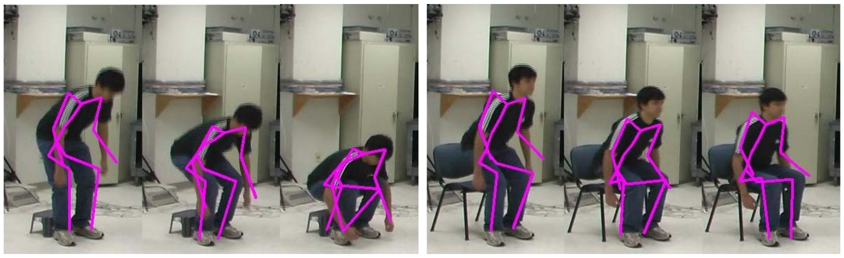
Figure 11. Tracking Results on other subjects.

6. Conclusion

We presented an approach to extend GPLVM and GPDM by including an embedding from observation space to latent space. Such an embedding preserves local distances in both the observation space and the pose space. Our approach provides an effective and computationally efficient approach for pose estimation. Unlike previous approaches, it emphasizes the importance of image observation in prediction of latent positions and tries to optimally balance reliance on image features and dynamics. We then introduced an extension to our model, OD-GPDM, to include contextual information. The joint angle trajectories in many actions show variations with respect to environmental and contextual variables. Instead of learning a separate model for different (quantized) values of the contextual variables, we presented an approach that models these variations and uses a single latent space to embed all pose variations due to differences in contextual variables. We also demonstrated the importance of contextual information in prediction of poses in such parameterized actions.

Acknowledgement

The authors would like to thank FXPAL for supporting the research described in the paper. Most of the work has been done while the first author was visiting FXPAL. This



(a) Step Stool

(b) Chair

Figure 9. Results of Context and Observation Driven GPDM on sitting action of Subject 1.

research was funded in part by the U.S Government's VACE program.

References

[1] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevance vector regression. *CVPR*, 2004. 4

[2] CMU-Mocap. http://mocap.cs.cmu.edu/. 6

[3] C. Ek, P. Torr, and N. Lawrence. Gaussian process latent variable model for human pose estimation. *MLMI*, 2007. 3

[4] A. Elgammal and C. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. *CVPR*, 2004. 2

[5] A. Elgammal and C. Lee. Separating style and content on a nonlinear manifold. *CVPR*, 2004. 3

[6] K. Grochow, S. L. Martin, A. Hertzmann, and Z. Popovic. Style-based inverse kinematics. *SIGGRAPH*, 2004. 3

[7] A. Gupta and L. Davis. Objects in action: An approach for combining action understanding and object perception. *CVPR*, 2007. 3

[8] A. Gupta, A. Mittal, and L. Davis. Constraint integration for efficient multiview pose estimation with self-occlusions. *PAMI*, 30(3), 2008. 2

[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. 3

[10] N. Howe, M. Leventon, and W. Freeman. Bayesian reconstruction of 3d human motion from single camera video. *NIPS*, 1999. 2

[11] A. B. J. Deutscher and I. Reid. Articulated body motion capture by annealed particle filtering. *CVPR*, 2000. 2

[12] A. Kanaujia, C. Sminchisescu, and D. Metaxas. Spectral latent variable models for perceptual inference. *ICCV*, 2007. 3

[13] N. Lawrence. Gaussian process models for visualisation of high dimensional data. *NIPS*, 2004. 1, 3, 4

[14] N. Lawrence and J. Candela. Local distance preservation in the gp-lvm through back constraints. *ICML*, 2006. 2, 4, 5

[15] M. Lee and R. Nevatia. Body part detection for human pose estimation and tracking. *WMVC*, 2007. 4

[16] R. Li, M. H. Yang, S. Scarloff, and T. Tian. Monocular tracking of 3d human motion with a coordinated mixture of factor analyzers. *ECCV*, 2006. 3

[17] Z. Lu, M. Carreira-Perpinan, and C. Sminchisescu. People tracking with the laplacian eigenmaps latent variable model. *NIPS*, 2007. 3

[18] D. Moore, I. Essa, and M. Hayes. Exploiting human action and object context for recognition tasks. *ICCV*, 1999. 3

[19] K. Murphy, A. Torralba, and W. Freeman. Graphical model for scenes and objects. *NIPS*, 2003. 3

[20] S. Roweis and L. Saul. Non linear dimensionality reduction by locally linear embedding. *Science*, 2000. 2

[21] A. Shon, K. Grochow, A. Hertzmann, and R. Rao. Learning shared latent structure for image synthesis and robotic imitation. *NIPS*, 2006. 3

[22] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3d human figures using 2d motion. *ECCV*, 2000. 2

[23] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking loose limbed people. *CVPR*, 2004. 2

[24] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Learning joint top-down and bottom-up processes for 3d visual inference. *CVPR*, 2006. 4

[25] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Bm3e: Discriminative density propagation for visual tracking. *PAMI*, 2007. 4

[26] J. Tenenbaum, V. DeSilva, and J. Langford. A global geometric framework for non-linear dimesionality reduction. *Science*, 2000. 2

[27] R. Urtasun, D. Fleet, and P. Fua. Monocular 3d tracking of the golf swing. *CVPR*, 2005. 3

[28] R. Urtasun, D. Fleet, and P. Fua. 3d people tracking with gaussian process dynamical models. *CVPR*, 2006. 3, 6

[29] R. Urtasun, D. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. *ICCV*, 2005. 3, 6

[30] R. Urtasun, D. Fleet, and N. Lawrence. Modeling human locomotion with topologically constrained latent variable models. *Human Motion Workshop*, 2007. 3

[31] J. Wang, D. Fleet, and A. Hertzmann. Gaussian process dynamical models. *NIPS*, 2005. 2

[32] J. Wang, D. Fleet, and A. Hertzmann. Multifactor gaussian process models for style content separation. *ICML*, 2007. 3

[33] A. Wilson and A. Bobick. Parametric hidden markov models for gesture recognition. *PAMI*, 21(9):884–900, 1999. 3