3D HUMAN MOTION TRACKING USING MANIFOLD LEARNING

Feng Guo‡, *Gang Qian*†‡

Arts, Media and Engineering Program†,
Department of Electrical Engineering‡
Arizona State University, USA
Email:{Feng.Guo,Gang.Qian@asu.edu}

ABSTRACT

This paper introduces a framework to track 3D human movement using Gaussian process dynamic model (GPDM) and particle filter. The framework combines the particle filter and discriminative learning approaches so that the 3D human model is not needed and optimal proposal distribution can be used. The structure of the joint motion and appearance are modelled using GPDM in a low dimensional space. Relevance vector machine (RVM) is used to construct the regression mapping between image latent space and joint angle latent space using the small training data set. Backward mapping from appearance to motion latent space makes the samples better drawn according to the most recent observation. Forward mapping from joint angle to silhouettes makes computation fast without generating synthetic images in tracking for particle weight evaluation. The experimental results show that our approach can track 3D people movement accurately given noisy image and different subjects' movements.

Index Terms— 3D Human Motion Tracking, GPDM, Particle filter, RVM

1. INTRODUCTION

Robust tracking of body kinematics of human movement from videos has been a big challenge in computer vision area for over a decade. There are two general solutions proposed to tackle the problem. One is model-based and another is view-based.

For model-based approaches, the explicitly specifying articulated models of the body parts, joint angles and dynamics are exploited in high dimensional space by searching the optimal solutions. The posterior probability of the human pose is obtained from Bayes' rule [9]. As an application of human pose tracking, particle filter is widely used as temporal inference. However, they usually lack an ability to directly use the data structure existing in the observation and direct map from observation to the state space. In addition, it's challenging to evaluate the weights of a large number of particles in the generative frame. The view-based approaches infer the human pose directly from the image observation. It estimates the pose using the examples obtained from the training images. Different learning approaches can be applied, including nearest-neighbor [10], relevance vector machine(RVM) [11]. But this method does not explore the rich dynamics represented in the state space. Also this approach requires a sufficient number of training examples to learn the good mapping.

For both approaches, the high dimensional data of the 3D poses and image observations make the problem harder. One way to tackle the problem is to reduce the dimensionality through nonlinear dimensionality reduction. In [3], Elgammal used manifold recovery method such as LLE to recover the underlying structure in the silhouette space and conduct regression of joint angles. In [2], Grauman inferred the 3D body structure from probabilistic PCA of combination of input silhouettes and 3D poses. In [1], Sminchisescu used kernel PCA over image inputs and joint angles and learned the underlying multimodal mapping from inputs to outputs using Bayesian mixture expert. Both [2] and [1] approaches do not use any dynamic information for the 3D poses as the tracking application. Recently, sophisticated generative models such as Gaussian Process Dynamic Model (GPDM)[6] have shown to be able to capture the underlying dynamics of movement and at the same time reduce the dimensionality of movement data. Such models have been used as priors for kinematic tracking of walking successfully.

By combining the view-based and model-based approaches in embedding space[4], the tracking system can better draw samples in particle filter framework. Still their approach modeled the dynamics using basic first order Markov model and a complex 3D model was needed to obtain the observation likelihood. In this paper, we propose a 3D human motion tracking system in a particle filter framework using GPDM. The system explores the underlying structures of the joint movement and appearance modeled by GPDM as the prior information. The mapping from silhouette to kinematics is utilized to better draw particles according to the most recent observation. For the likelihood calculation, the system is able to conduct fast computation by mapping from latent space of joint motion to visual latent space without need of generating synthetic images.

A sketch of our system is given in figure 1. It includes the learning part and the tracking part. The learning process will obtain the parameters of the system and the tracking part implements particle filtering from learned models.

For the learning part, firstly we use GPDM to obtain the low dimensional manifold Lj, the latent space of the local joint angles and Ls, the latent space of silhouettes for different views. Given such two embeddings, the nonlinear mapping between these two low dimensional spaces are learned. Forward mapping from Lj to Ls and the backward mapping from Ls to Lj are established using sparse Bayesian regression such as relevance vector machine. In tracking, the latent points of the input silhouettes are localized. Then based on this, we determine "observations" in Lj using the backward mapping. By using these latent observations, the dynamics can be better used to produce particles. Instead of using 3D human model to generate images, we forwardly map those particles to the corresponding points in Ls, then to visual space. The likelihood is evaluated by computing the distance between observed image data and new generated data in visual space.

The details of the learning and tracking steps are described in the following part.

1-4244-1437-7/07/$20.00 ©2007 IEEE

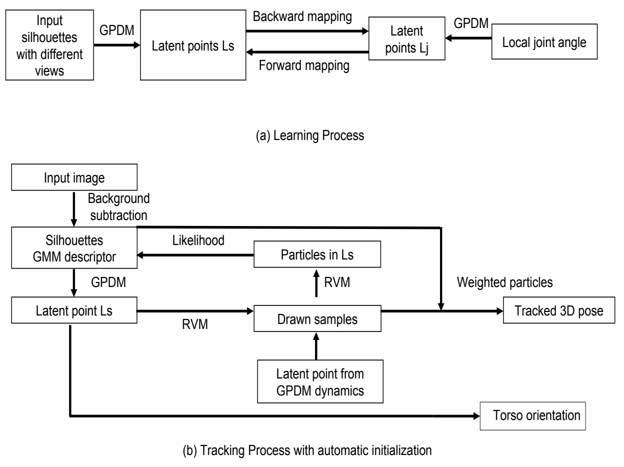


Fig. 1. Overview of our framework.(a): Learning process. (b): Tracking process.

2. IMAGE FEATURES REPRESENTATION

Among image features, silhouettes have been commonly adopted due to the simplicity of silhouette extraction. A number of silhouette representations have been applied to kinematic recovery, including Hu-movement, shape context. Studies [12] shows that shape context representation is better than Hu-moments. Here we use a Gaussian mixture model (GMM)-based silhouette representation [13]. GMM model considers that the silhouette is represented as a set of coherent regions in 2D space. It's critical to measure the silhouette similarities. Based on the GMM descriptor, the Kullback-Leibler divergence (KLD) is used to measure the distance between two silhouettes (pixel spatial distributions). Although GMM representation with KLD measurement can capture the similarity of the image silhouettes well, it is not a simple way in computation as the vector representation. To obtain the vector descriptor for each GMM descriptor, we use the relative distances of one silhouette to several key silhouettes to locate this point. The distance between the measured silhouette and each key silhouette is one element in the vector. So the key frames are needed as the bases. From the silhouettes training data, the distance matrix of all silhouettes is computed firstly. The multidimensional scaling(MDS) is performed on the distance matrix. The eigenvalues are analyzed to decide the key frame number. In 3000 training samples, 46 dimension can keep 99.5% variance of the eigenvalues. This is the dimensionality of the silhouette vector. To obtain the key silhouette, *k*-means is used to get the cluster centers of the training samples. Because no clear cluster center exists, we pick up the sample which is the closest to the cluster center as one key silhouette. The distance matrices in Figure 2 show that GMM vector representation measures the relative distance of the silhouettes better than shape context vectorization.

3. SYSTEM LEARNING

3.1. Gaussian Process Dynamic Model

Gaussian Process Dynamic Model (GPDM) [6] provides a low dimensional embedding of the data and the latent dynamics simultaneously. [8] extended the GPDM to balanced GPDM to handle multiple subjects' stylistic variation by raising the dynamic density function. Given a set of data $Y = [y_1, ..., y_t, ..., y_N]^T$ and denote the latent variable associated with each data point as $X = [x_1, ...x_t, ..., x_N]^T$. y_t are D-dimension data points and we assume that they have subtracted the mean μ and are zero mean. x_t are d dimension data points. Here $d < D$. As a regression function, GPDM defines two Gaussian processes to relate y_t with x_t and x_t with x_{t-1} at time t. The model is defined as:

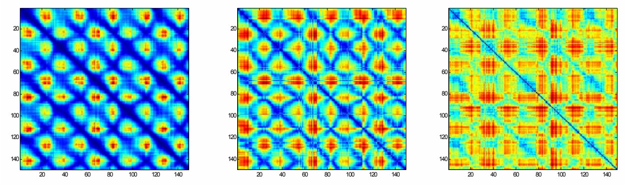


Fig. 2. Distance matrices for 149 frames side view walking silhouettes. Dark blue-colored pixels indicate small distances. The periodic pattern is caused by both repeated movement in different gait cycles and the half cycle ambiguity. Left: Computed from KLD for GMM. Middle: Computed from GMM vectorization using 46 key frames. Right: Computed from shape context vectorization using 90 clusters with 8000 training samples.

$$x_t = A\phi(x_{t-1}) + n_{x,t} \quad (1)$$

$$y_t = B\varphi(x_t) + n_{y,t} \quad (2)$$

where A and B are regression weights, $n_{x,t}$ and $n_{y,t}$ are normal Gaussian noise. $\phi(x_{t-1})$ and $\varphi(x_t)$ can be linear or nonlinear kernel functions.

Marginalizing over A and B gives the latent dynamics and the latent variables:

$$p(X|\Phi) \propto exp(-\frac{1}{2}tr(K_x^{-1}(\hat{X} - \tilde{X})(\hat{X} - \tilde{X})^T) \quad (3)$$

$$P(Y|X,\Psi) \propto exp(-\frac{1}{2}tr(K^{-1}YY^T)) \quad (4)$$

where $\hat{X} = [x_2...x_t]^T$ and $\tilde{X} = [x_1...x_{t-1}]^T$. K_x is the kernel associated with the dynamics Gaussian process and is constructed on the matrix $\tilde{X}$. We use an RBF kernel and white noise term for GPDM dynamics:

$$k_x(x_t, x_{t-1}) = \alpha_d exp(-\frac{\gamma_d}{2}||x_t - x_{t-1}||^2) + \beta_d^{-1}\delta_{t,t-1} \quad (5)$$

where $\Phi = (\alpha_d, \gamma_d, \beta_d)$ are dynamics parameters.

$\varphi(x_t) = [k(x_t, x_1), ...k(x_t, x_i)..., k(x_t, x_N)]$ is column vector of the kernel function $k(x_t, x_i)$. Here we use the RBF kernel

$$k(x_t, x_i) = \alpha exp(-\frac{\gamma}{2}||x_t - x_i||^2) + \beta^{-1}\delta_{x_t,x_i} \quad (6)$$

where α is the overall scale of the output, γ is the inverse width of the RBFs. The variance of the noise is given by β^{-1}. $k(x_t, x_i)$ is the elements of the kernel matrix K. $\Psi = (\alpha, \beta, \gamma)$ are the unknown model parameters.

GPDM learning is to learn the model parameters Ψ, Φ and latent variable X. That is equivalent to minimize the negative log of the object function with respect to the Ψ, Φ:

$$L_d = \frac{D}{2}ln|K| + \frac{1}{2}tr(K^{-1}YY^T + \frac{d}{2}ln|K_x| + \frac{1}{2}tr(K_x^{-1}(\hat{X} - \tilde{}X)(\hat{X} - \tilde{X})^T) \quad (7)$$

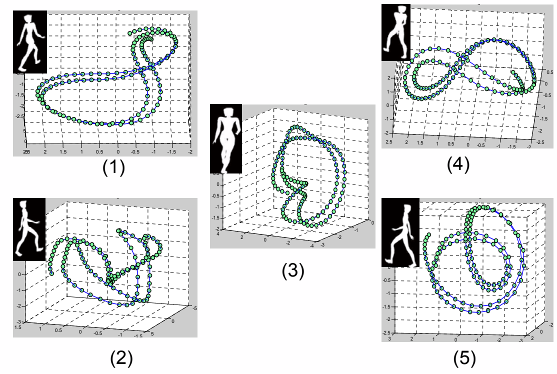


Fig. 3. GPDM latent space for 5 views walking silhouettes.

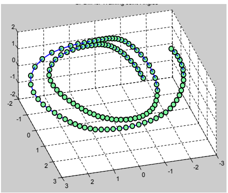


Fig. 4. GPDM latent space for local joint angles learned from 2 walking cycles.

The optimization will simultaneously generate the latent point X and the mapping parameters. In the application of GPDM for silhouettes and joint angles, the points will be embedded in the latent space Ls and Lj. In Figure 3, the 3D latent points of silhouettes for each view are shown. The latent space for the learned joint angles are shown in Figure 4.

Once the model is learned, the corresponding x_j for a given new input data y_j can be obtained by solving the likelihood object function:

$$L_m(x_j, y_j) = \frac{||y_j - \mu(x_j)||^2}{2\sigma^2(x_j)} + \frac{D}{2} ln\sigma^2(x_j) + \frac{1}{2}||x_j||^2 \quad (8)$$

where

$$\begin{aligned} \mu(x_j) &= \mu + Y^T K^{-1} \varphi(x_j) && (9) \\ \sigma^2(x_j) &= k(x_j, x_j) - \varphi(x_j)^T K^{-1} \varphi(x_j) && (10) \end{aligned}$$

$\mu(x_j)$ is the mean pose reconstructed from the latent point x_j, $\sigma^2(x_j)$ is the reconstruction variance . μ is the mean of the training data Y. $\varphi(x_j)$ is the kernel function of x_j with training data. Given input y_j, the initial latent position is obtained $x_j = argmin_{x_j} L_m(x_j, y_j)$. Given x_j, using equation (9), the high dimensional data $\hat{y_j}$ can be obtained too.

3.2. Regression Between Two Latent Spaces

Given the low dimensional representations of human poses Lj and silhouettes Ls, the regression between Lj and Ls are learned using sparse relevance vector machine(RVM)[7]. From Ls to Lj is backward mapping and from Lj to Ls is forward mapping. RVM is also one Gaussian process for classification and regression. The hyperparameters are introduced in RVM to control the weights W. This will generate sparse non-zero weights compared with SVM. The detail of the learning step can be found in [7]. In our training, the relevance vectors account about $10\% - 20\%$ of the total data which will save the computation resources.

4. TRACKING USING PARTICLE FILTER

We use the particle filter to track the 3D human movement. Let Lj_t be the latent point of J_t, the joint angles of 3D body pose. The state is defined as Lj_t. Given the sequence of images $I_{1:t}$, the evolution of the distribution of the state is approximated by a particle filter with importance sampling:

$$p(Lj_t | I_{1:t}) = w_t^i Lj_t^i \quad (11)$$

where the importance weights of the particles are given by:

$$w_t^i = \frac{p(I_t | Lj_t^i) p(Lj_t^i | Lj_{t-1}^i)}{q(Lj_t^i | Lj_{t-1}^i, I_t)} \quad (12)$$

Particles are drawn from the proposal distribution q. We use the proposal density as the mixture of two distributions.

$$q(Lj_t^i | Lj_{t-1}, I_t) = \alpha q_{bm}(Lj_t | I_t) + (1 - \alpha) p(Lj_t | Lj_{t-1}) \quad (13)$$

where α determines the convergence of the proposal distribution to the observation and in our experiment we use $\alpha = 0.5$. $q_{bm}(Lj_t | I_t) = p(Lj_t | Ls_t)$ is the distribution from backward mapping,where Ls_t is the latent point of I_t.

Dynamics learned from GPDM are used for $p(Lj_t^i | Lj_{t-1}^i)$. Because of the low dimensional space, we only draw 100 particles from the proposal distribution. This saves the computation resources greatly.

For the likelihood,

$$p(I_t | Lj_t^i) \propto p(I_t | f(Lj_t^i)) = P(I_t | \hat{I}_t^i) \quad (14)$$

where $f(Lj_t^i) = \hat{I}_t^i$ is the mapped image features from Lj_t^i through forward mapping. Assume the likelihood in visual input is Guassian distribution $P(I_t | \hat{I}_t^i) = \frac{||I_t - \hat{I}_t^i||^2}{2\sigma_I^2}$. Through this mapping, the complex 3D human model is avoided.

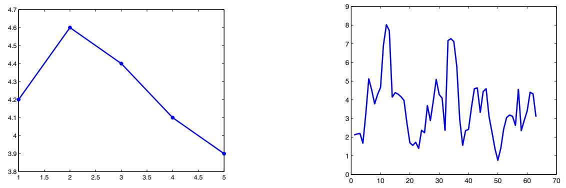
Given the estimated result of Lj_t, J_t will be obtained and the 3D body pose will be inferred. For camera view estimation, given the learned view-based manifolds for each view, determining the view point reduces to finding the manifold that minimizes the mapping error of a sequence of inputs $I_{1:t}$. Given input sequence $I_{1:t}$ and its projections $Ls_{1:t}$, we chose the manifold that minimizes $||I_{1:t} - h(Ls_{1:t})||$.

5. EXPERIMENTAL RESULTS

The proposed framework has been tested using both synthetic and real image sets. The joint angle data for synthetic data generation were from the CMU motion capture database [5]. The joint angle includes 42 joints. A total number of 320 training silhouettes were rendered using Maya with 5 viewpoints (uniformly placed on the sphere equator centered at the hip of the body).

For the model learning, we use three dimensional embedding space for both silhouettes Ls and local joint angle Lj because this is the smallest dimension which we can use to discriminate different poses and learn complex motions. The learned GPDM trajectories are shown in Figure 3 and 4.

Firstly we test the system using the synthetic data to evaluate the accuracy of our tracking system. The mean RMS error between the true and the estimated joint angles is used. We test one motion sequence from a different subject not included in training and the 3D model is different too. We test 5 views of the walking data which has 63 frames(2 walking cycles) for each view. The mean RMS errors for different views are shown in Figure 5(a). The RMS error for each frame of view 2 test sequence (shown in Figure 6) are shown in Figure 5(b).



(a) (b)

Fig. 5. (a).The mean of RMS errors for test sequence with different views. (b). Reconstruction pose RMS errors for each frame.

Some of the estimated poses for view 2 of the test sequence are shown in Figure 6. The framework can track the walking well.

A real video (40 frames, two steps side view walking) are also used to evaluate the proposed system. Due to the noise presented in the video, the extracted silhouettes are not as clean as the synthesized ones. However, the proposed approach can still produce perceptually sound results. Some recovered poses are shown in Figure 7. From frame 29 to frame 32 there is a jump for the leg movement because of the capture reason. But the system still can estimate the correct poses.

6. CONCLUSION

In this paper, we propose a framework to robust track the view-based 3D human movement using GPDM and particle filter. GPDM captures the structures of the silhouettes and motion joints in manifold spaces. Those low dimensional spaces make the tracking more efficient. The mapping between two low dimensional spaces are performed using relevance vector machine. So better sampling scheme and weight evaluation can be realized. Also no 3D model is needed. The experimental results show that the framework can work for different subjects and noisy image observation.

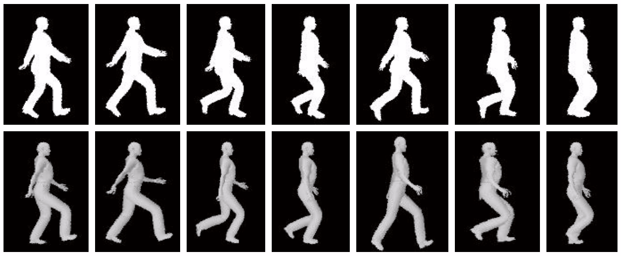


Fig. 6. Reconstructed poses of test sequence of view 2. First row: Input silhouettes. Seconde row: The reconstructed poses.

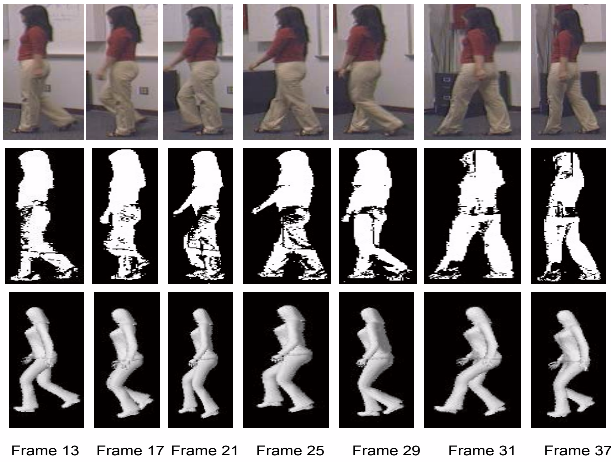


Fig. 7. Tracking of the real video images. First row:Input video images. Second row:Extacted noisy silhouettes. Third row: Reconstructed tracking results.

7. REFERENCES

[1] C. Sminchisescu, A. Kanaujia, Z. Li and D. Metaxas, ”Conditional Visual Tracking in Kernel Space”, *NIPS*,2005

[2] K.Grauman, G.Shakhnarovich and T.Darrell,”Inferring 3D Structure with a Statistical Image-Based Shape Model”,*ICCV2003*

[3] A. Elgammal and C.-S. Lee, ”Inferring 3D Body Pose from Silhouettes using Activity Manifold Learning”, *CVPR*, 2004

[4] C.Curio and M.A.Giese,”Combining View-based and Model-based Tracking of Articulated Human Movements”, *Workshop on Application of Computer Vision(WACV)*,2005

[5] CMU Human Motion Capture Database. Available online at *http://mocap.cs.cmu.edu/search.html*

[6] J. Wang, D.J. Fleet and A. Hertzmann, ”Gaussian Process dynamical models”, *NIPS*, 2005

[7] M. Tipping, ”Sparse Bayesian learning and the Relevance Vector Machine”, *Journal of Machine Learning Reserach*, vol.1, pp. 211-244, 2001.

[8] R. Urtasun, D.J. Fleet and P. Fua, ”3D people tracking with Gaussian process dynamical models”, *CVPR*, 2006

[9] H.Sidenbladh, M.Black and D.Fleet,”Stochastic Tracking of 3D Human Figures Using 2D Image Motion”, *ECCV*, 2000

[10] G. Mori and J. Malik, ”Estimating human body configurations using shape context matching”, *ECCV*, 2002.

[11] A. Agarwal and B. Triggs, ”3D Human Pose from Silhouettes by Relevance Vector Regression”, *CVPR*, 2004.

[12] R. Poppe and M. Poel, ”Comparison of Silhouette Shape Descriptors for Example-based Human Pose Recovery”, *FGR*, 2006

[13] F.Guo,G.Qian, ”Learning and Inference of 3D Human Poses from Gaussian Mixture Modeled Silhouettes”, *ICPR*, 2006