MULTI-VIEW TRACKING OF ARTICULATED HUMAN MOTION IN SILHOUETTE AND POSE MANIFOLDS

Feng Guo and Gang Qian

Arts, Media and Engineering Program and
Department of Electrical Engineering,
Arizona State University,
Tempe, AZ 85281

ABSTRACT

This paper presents a multi-view articulated human motion tracking framework using particle filter with manifold learning through Gaussian process latent variable model. The dimensionality of the input image observation and joint angles are reduced using Gaussian process models to improve the tracking efficiency. The forward and backward mappings between the two low dimensional spaces are then obtained using relevance vector machine and Batesian mixture of experts (BME). Improved sampling schemes and auto-initialization are obtained using BME. Without using a 3D body model, effective likelihood evaluation is obtained through RVM using images from multiple views. Tracking results obtained using real videos with complex dance movement show the efficacy of the proposed approach.

Index Terms— articulated movement tracking, Gaussian process latent variable model, particle filtering

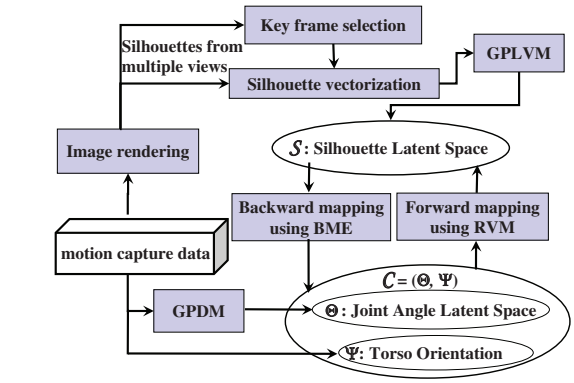
1. INTRODUCTION

Vision-based articulated motion tracking is a challenging problem for computer vision. Existing 3D movement tracking algorithms can be roughly classified into two categories, namely, generative-based and discriminative-based approaches. Generative-based approaches use an articulated 3D model for tracking. The model is projected onto an image plane and an error function is computed to indicate the quality of the match. Recently, nonlinear probabilistic generative models, such as Gaussian process latent variable model (GPLVM) [1] have been used to obtain a probabilistic low-dimensional representation of body joint data, which can reduce the tracking complexity. In addition, as variants of GPLVM, Gaussian process dynamical models (GPDM) [2, 3] have proved powerful to capture the underlying dynamics of movement and meanwhile with reducing the dimensionality of the pose space. Such models have been successfully used as priors for kinematic tracking of walking [3]. Multi-view based approaches are common for generative-based 3D body trackers to reduce ambiguity. For example, visual hulls is often extracted to reconstruct the body shape, based on which the articulated body pose can be recovered [4, 5, 6].

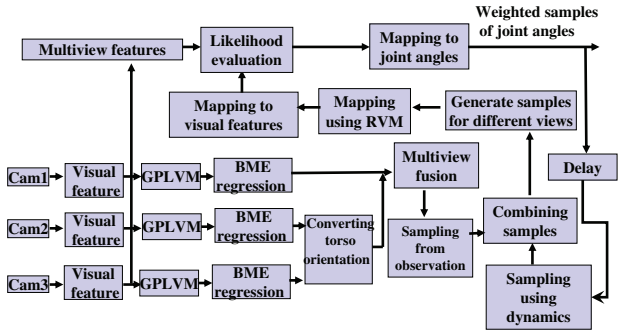
Discriminative-based approaches infer body poses directly from training data, using machine learning techniques, including relevance vector machine [7], nearest-neighbor [8], and local linear embedding [9]. In these methods, the kernel principal component analysis (PCA) [10] and probabilistic PCA [8] are commonly used to reduce the dimensionality of the image and pose spaces. To tackle the one-to-many mapping problem from image to pose, expert-based learning has been used [11, 12]. The basic idea is to split the input image space into a nested set of regions and the data mapping that falls in these regions is approximated. Multiple views have also been exploit in discriminative-based methods [13, 14, 15].

Recently, we have developed a monocular 3D human motion tracking system by integrating the generative and discriminative-based approaches using manifold learning and Gaussian process in a particle filtering framework [16]. This system explores the underlying dynamics of the movement modeled by GPDM as the prior information. It maps the image observation to low dimension by GPLVM. This also reduces noises caused by image segmentation or different appearance of the input image. This generates the small number of the parameters and particles for tracking and provide quicker performance and higher accuracy. The mapping from silhouette to kinematics is utilized to better draw particles according to the most recent observation and provide initialization. The system is view-independent so that it can determine both body kinematics such as joint angles and torso orientation given input from any views. Ambiguity is overcome by mapping multi-model using Bayesian mixture of expert. For the likelihood calculation, the system is able to conduct fast computation without need to generate synthetic images in tracking for particle weight evaluation.

It is clear from existing research that using multiple views can effectively reduce tracking ambiguity. In this paper, we present a multi-view tracking framework by extending our previous monocular tracking approach to improve tracking



(a)



(b)

Fig. 1. An overview of the proposed framework, (a): training phase; (b): tracking phase.

accuracy and reduce ambiguity. Effective fusion methods of images from multiple views are developed. Our experimental results obtained using both synthetic and real video data demonstrate that in addition to the fact that using multiple views can reduce tracking ambiguity and improve tracking accuracy, complex movement such as dance movement can also be tracked reliably using the proposed multi-view tracking approach.

2. PROPOSED MULTI-VIEW TRACKING APPROACH

An overview of the architecture of the proposed system is presented in Figure 1.

2.1. System Training

The training phase in the multi-view framework is identical to that in the monocular framework [16], since camera geometry is not required in system training. Similar to [16], the training consists of training data preparation and model learning. In data preparation, synthetic images are rendered using animation software, e.g. Maya, and motion capture data. The learning part has five major components. First, key frames are selected from the synthetic images using multidimensional scaling [17] and k-means. Based on these key frames, an input silhouette can be vectorized using its distances to all the key frames [16]. Then GPLVM is used to construct the low dimensional manifold $\mathcal{S}$ of the image silhouettes from multiple views using their vectorized descriptors. The third component is the dimension reduction of pose data with modeling of motion dynamic priors. GPDM is used to obtain the manifold Θ of full body pose angles. This latent space is then augmented by the torso orientation space Ψ to form the *complete* pose latent space $\mathcal{C} \equiv (\Theta, \Psi)$. The forward and backward nonlinear mappings between $\mathcal{C}$ to $\mathcal{S}$ are constructed in the learning phase. The forward mapping is established from $\mathcal{C}$ to $\mathcal{S}$ using RVM, which will be used to efficiently evaluate sample weights in the tracking phase. The multimodal (one-to-many) backward mapping from $\mathcal{S}$ to $\mathcal{C}$ is obtained using BME. This backward mapping from the silhouette manifold S to the joint space of the pose manifold and the torso orientation C is needed to conduct both autonomous tracking initialization and sampling from the most recent observation. More details on system training can be found in [16].

2.2. Tracking From Multiple Views

In tracking, weighted movement particles in $\mathcal{C}$ are propagated based on the image observation from multiple views up to the current time instant and learned movement dynamic models. In our discussion on multi-view tracking, we assume that we have n cameras installed around the circle with known relative looking directions. The tracking procedure using multiple view goes as follows. Body silhouettes are first extracted from input images and then vectorized. Given the input silhouettes from each view at t, latent points of the silhouettes $\mathbf{s}_t = \{s_i\}_{i=1}^n$ can be found in the silhouette manifold $\mathcal{S}$ using the learned GPLVM. Then BME is invoked to find a few plausible pose estimates in $\mathcal{C}$ for each s_i. BME results using multiple views are then integrated to reduce the set of candidate solutions. The integrated solution can also be used to initialize the tracking in the first frame. In the tracking mode, movement samples are drawn according to both the BME outputs and movement dynamics represented by the learned GPDM. The sample weights are evaluated according to the distance between the observed and predicted silhouettes. The empirical posterior distribution of poses is then obtained as the weighted samples.

To be specific, a particle filter defined over $\mathcal{C}$ is used for 3D movement tracking. The state parameter at time t is $c_t = (\theta_t, \psi_t)$, where θ_t is the latent point of the body joint angles, and ψ_t is the torso orientation. Given a sequence of latent silhouette points $\mathbf{s}_{1:t}$ obtained from input images using GPLVM, the posterior distribution of the state is approximated by a set of weighted samples $\{w_t^{(i)}, c_t^{(i)}\}_{i=1}^M$. The importance weights of the particles are propagated over time

$$w_t^{(i)} \propto w_{t-1}^{(i)} \frac{p(\mathbf{s}_t|c_t^{(i)})p(c_t^{(i)}|c_{t-1}^{(i)})}{q(c_t^{(i)}|c_{t-1}^{(i)}, \mathbf{s}_t)} \quad (1)$$

Particles are propagated over time from a proposal distribution q. To take into account both the movement dynamics and

the most recent observation $\mathbf{s}_t$, the proposed approach selects q from the following mixture of two distributions

$$q(c_t|c_{t-1}, \mathbf{s}_t) = \pi q_b(c_t|\mathbf{s}_t) + (1-\pi)p(c_t|c_{t-1}) \quad (2)$$

where $q_b(c_t|\mathbf{s}_t)$ is chosen as the integrated BME outputs from different views. The second term in (2) is from movement dynamics learned using GPDM and a first-order AR model for the torso orientation

$$p(c_t|c_{t-1}) = p(\theta_t|\theta_{t-1})p(\psi_t|\psi_{t-1}) \quad (3)$$

In (2), π is the mixture coefficient of the BME-based prediction and the dynamics-based prediction components. The experiment for the proposed approach uses $\pi = 0.5$. Because $\mathcal{C}$ is only a 5D space, only 100 particles are drawn from the proposal distribution, which greatly saves the computation cost. In the following, we present our method to ingrate the BME outputs from different views and the computation of joint likelihood using multi-view image features.

2.2.1. Integration of BME Outputs

The BME output from one view is a mixture of Gaussian $p(c|s_i) = \sum_{k=1}^{K} \alpha_{i,k}\mathcal{N}(\mu_{i,k}, \Sigma_{i,k})$, $i = 1, \cdots, n$. $\mu_{i,k} = (\theta_{i,k}, \psi_{i,k})$ is the mean of the kth mixture component of the BME output from the ith view, where $\theta_{i,k}$ is latent point of the pose in pose manifold and $\psi_{i,k}$ the torso orientation based on the i^{th} view. To properly integrate BME outputs from multiple views, the torso orientation estimates need to be aligned. In our approach, we align $\psi_{i,k}, i = 2, \cdots, n, k = 1, \cdots, K$ to the first view, i.e., we compute $\psi_{i,k}^{(a)} = \psi_{i,k} - \Delta\psi_i$, where $\Delta\psi_i$ is the camera looking direction angle difference between view 1 and view i. $\Delta\psi_i, i = 2, \cdots, n$ are known from the system calibration. In the following discussion, $p(c|s_i), i = 1, \cdots, n$ represents the orientation-aligned BME output. The integrated BME output is given by $p(c|\mathbf{s}_t)$ as follows

$$\begin{aligned} p(c|\mathbf{s}_t) &= \frac{p(\mathbf{s}_t|c)p(c)}{p(\mathbf{s}_t)} = \frac{\prod_i^n p(s_i|c)p(c)}{p(\mathbf{s}_t)} \\ &= \frac{\prod_i^n p(c|s_i)p(s_i)}{p(c)^{n-1}p(\mathbf{s}_t)} \propto \prod_i^n p(c|s_i) \end{aligned} \quad (4)$$

where a uniform prior of c is assumed. Based on 4, $p(c|\mathbf{s}_t)$ is given by the product of a series of mixtures of Gaussian, which can still be represented by a mixture of Gaussian. The reduce the number of mixtures in $p(c|\mathbf{s}_t)$, in $p(c|s_i)$, only mixture components with $\alpha_{i,k} > 0.1$ are taken into account. Once $p(c|\mathbf{s}_t)$ is obtained by integrating BME output from multiple views, it can be used to initialize the tracking in the first frame, and inform the sampling step in the following frames according to (2) to generate complete sample set.

2.2.2. Likelihood Evaluation Using Multiple Views

Another major challenge in multi-view tracking is to evaluate the sample weights using observation from all the views. The likelihood of $\mathbf{s}_t$, w.r.t. a sample c is

$$p(\mathbf{s}_t|c) = \prod_i^n p(s_i|c) = \prod_i^n p(s_i|c_i) \quad (5)$$

where $c_i = (\theta, \psi_i)$ is transformed from c with correct orientation angle with respect to the ith view so that $\psi_i = \psi_1 + \Delta\psi_i$. $p(s_i|c_i)$ can be evaluated using learned RVM forward mapping. See [16] for details.

3. EXPERIMENTAL RESULTS

Dance movement is used for the evaluation of the proposed system as a complex movement. In total 344 frames of motion capture data from two sequences of dance movement were used in training, down-sampled from 1376 frames. There are 66 local joint angles in the original motion capture data, but only 48 major joint angles are considered in these experiments. Two different 3D models were used to produce training silhouettes with diverse appearance. In total 4128 training silhouettes were rendered using poses with changing torso orientations from 12 non-overlapping ranges. Then in each camera view, an angle is uniformly drawn over an interval of 30°. Hence, for each given motion capture frame, there are 12 complete pose frames with different torso orientations. Using these image silhouettes and the corresponding ground truth data of joint angles and torso orientation, a 5D manifold of silhouettes $\mathcal{S}$ is obtained using GPLVM and a 3D local joint angle manifold Θ is obtained using GPDM. Trajectories of two dance sequences in Θ obtained using GPDM are shown in Figure 2. The silhouette vectorization is based on key frames obtained from 2100 silhouettes of one training sequence. Several key frames are shown in Figure 3. In total 42 key frames are obtained which means the vector of dance silhouette is 42 dimensions. The proposed multi-view tracking system has been tested using both synthetic and real image sets using three cameras. To evaluate the performance of the proposed system, synthetic data were first used to evaluate the accuracy of the tracking system. The average RMS errors obtained using synthetic testing sequences from different views are 10.9°, 7.2° and 16° respectively. The multi-view approach resulted in an RMS error of 4.9° with improved tracking accuracy.

Real videos collected synchronously by three cameras were used to test the proposed algorithm. The three cameras were located at roughly mid-body height. Silhouettes were obtained using background subtraction. Figure 4 presents some of the silhouettes from three views. 3D rendering of some of the multi-view tracking results are shown in the last two rows of Figure 4. Body pose estimates are visually accurate in most part of the video.

4. CONCLUSION

A multi-view framework for articulated motion tracking in silhouette and pose manifolds is proposed to integrate the discriminative- and generative-based approaches. The proposed framework makes use of multiple views to improve

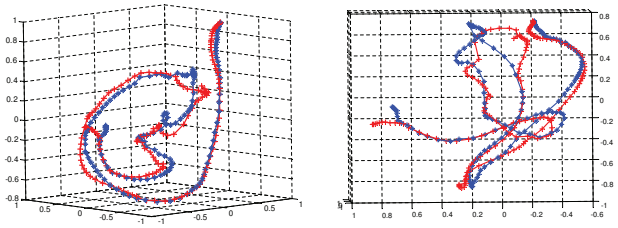


Fig. 2. Two views of a 3D GPDM learned using dance data set Θ_T from two sequences.

Fig. 3. Some key frames for dance silhouettes vectorization.

tracking accuracy, reduce ambiguity and handle complex movement such as dance.

5. ACKNOWLEDGEMENT

This paper is based upon work partly supported by U.S. National Science Foundation on CISE-RI no. 0403428 and IGERT no. 0504647. [1]

6. REFERENCES

[1] Neil D. Lawrence, "Gaussian process latent variable models for visualization of high dimensional data," in *Proceedings of Conference on Neural Information Processing Systems*, 2003.

[2] Jack M. Wang, David J. Fleet, and Aaron Hertzmann, "Gaussian process dynamical models.," in *Proceedings of Conference on Neural Information Processing Systems*, 2006, pp. 1441–1448.

[3] D.J. Fleet R. Urtasun and P. Fua, "3D people tracking with gaussian process dynamical models," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[4] K. Grauman, G. Shakhnarovich, and T.J. Darrell, "A bayesian approach to image-based visual hull reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003, pp. I: 187–194.

[5] I.Mikic, M.M.Trivedi, E.Hunter, and P.C.Cosman, "Human body model acquisition and tracking using voxel data," *International Journal Of Computer Vision*, vol. 53, no. 3, pp. 199–223, July 2003.

[6] S. Cheng and M. Trivedi, "Articulated human body pose inference from voxel data using a kinematically constrained gaussian mixture model," in *EHuM2: Workshop on Evaluation of Articulated Human Motion and Pose Estimation*, 2007.

[7] Ankur Agarwal and Bill Triggs, "3D human pose from silhouettes by relevance vector regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[8] Kristen Grauman, Gregory Shakhnarovich, and Trevor Darrell, "Inferring 3D structure with a statistical image-based shape model," in *Proceedings of the IEEE International Conference on Computer Vision*, Nice, France, October 2003.

[9] Ahmed Elgammal and Chan-Su Lee, "Inferring 3D body pose from silhouettes using activity manifold learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[10] Z. Li D. Metaxas C. Sminchisescu, A. Kanaujia, "Conditional visual tracking in kernel space," in *Proceedings of Conference on Neural Information Processing Systems*, 2005.

[11] Ankur Agarwal and Bill Triggs, "Monocular human motion capture with a mixture of regressors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[12] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Discriminative density propagation for 3D human motion estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[13] R. Rosales, M.Siddiqui, J.Alon, and S.Sclaroff, "Estimating 3d body pose using uncalibrated cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001, pp. I:821–827.

[14] Teofilo E. de Campos and D.W. Murray, "Regression-based hand pose estimation from multiple cameras," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, pp. I: 782–789.

[15] A. Bissacco, M.H. Yang, and S. Soatto, "Fast human pose estimation using appearance and motion via multi-dimensional boosting regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[16] Feng Guo and Gang Qian, "Monocular 3d tracking of articulated human motion in silhouette and pose manifolds," *EURASIP Journal on Image and Video Processing*, 2008.

[17] I. Borg and P. Groenen, *Modern multidimensional scaling. Theory and applications*, Kluwer Academic Publishers, 2005.

Fig. 4. Three-view tracking results of dance movement. Rows 1-2, 3-4, and 5-6 show sample input images and silhouettes from views 1, 2, and 3. The last two rows show the tracked poses rendered from view 1 and a novel view.

[1] Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. National Science Foundation (NSF).