MONOCULAR TRACKING 3D PEOPLE BY GAUSSIAN PROCESS SPATIO-TEMPORAL VARIABLE MODEL

Junbiao Pang[1, 2], Laiyun Qing[2], Qingming Huang[1, 2], Shuqiang Jiang[2], Wen Gao[3]

[1]Graduate School of Chinese Academy of Sciences, Beijing, P.R. China
[2]Key Lab. of Intelligent Information Processing, Institute of Computing Technology Chinese Academy of Sciences, Beijing, P.R. China
[3]Institute of Digital Media, Peking University, Beijing, P.R. China
{jbpang, lyqing, qmhuang, sqjiang, wgao}@jdl.ac.cn

ABSTRACT

Tracking 3D people from monocular video is often poorly constrained. To mitigate this problem, prior knowledge should be exploited. In this paper, the Gaussian process spatio-temporal variable model (GPSTVM), a novel dynamical system modeling method is proposed for learning human pose and motion priors. The GPSTVM provides a low dimensional embedding of human motion data, with a smooth density function that provides higher probability to the poses and motions close to the training data. The low dimensional latent space is optimized directly to retain the spatio-temporal structure of the high dimensional pose space. After the prior on human pose is learned, the particle filtering can be used tracking articulated human pose; particle filtering propagates over time in the embedding space, avoiding the curse of dimensionality. Experiments demonstrate that our approach tracks 3D people accurately.

Index Terms— Gaussian process, dimension reduction, machine learning, motion estimation, particle filtering

1. INTRODUCTION

Tracking 3D people from monocular video is a fundamental problem for human motion analysis in computer vision community. It has many important applications: video surveillance, gesture analysis, advanced human computer interface, etc. The task of 3D people tracking can be defined as follows: given the initial 3D people state in the first video frame, tracking algorithms will update the 3D people's state continuously, given the successive frames. Due to the poor constraints caused by self-occlusions, ambiguities and image measurement noise as well as high dimensional state space of human motion, prior models of the poses and motions should be exploited to enhance the tracker performance.

While some powerful models of 3D human pose are emerging, what characteristics of a prior model are more suitable for tracking 3D people? Intuitively, a model that better encodes the sophisticated dynamics and the spatial information of human poses is demanded. Learning such a model is challenging because of the nonlinearity of human dynamics and the high dimensional human poses.

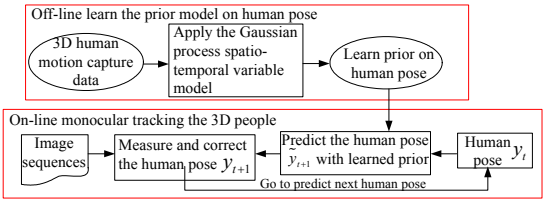

Fig. 1 Overview of the tracking algorithm

Some approaches modeling the sophisticated human motion have involved parameterization of the human pose through nonlinear dimensionality reduction, using local geometrical attributes of the high dimensional poses [1, 2]. While these methods yielding mapping from the pose space to the embedding space does not provide a probabilistic model over poses, nor a dynamical model. Thus the additional step is required to construct a dynamical model. For example, Agarwal and Triggs [3] learn a mapping from silhouettes to poses using relevance vector machine and a second-order auto regression (AR) dynamical model. Though spatio-temporal isomap (ST-Isomap) handles the spatio-temporal structure of the high dimensional data in the low dimensional space, it does not provide a probabilistic density model over poses, nor a mapping back from the latent space to the pose space [4].

The other approaches model the human poses prior in an embedding pose space based on probabilistic model. In [5, 6, 7, 8], locally linear coordination (LLC) [5] and Gaussian process latent variable model (GPLVM) [6, 7] are used for tracking applications, but the embedding learned by LLC does not encode the dynamics of human poses, nor does GPLVM. In [8], Gaussian process dynamic model (GPDM) and a second-order Markov model are utilized for tracking. GPDM learn a non-linear embedding with an AR dynamics process in latent space [9]. However, the GPDM does not model the spatial relationship of high dimensional data.

This work was supported by National High-Tech Research and Development Program (863 Program): 2006AA01Z117

This paper introduces a new model, Gaussian process spatio-temporal variable model (GPSTVM), for effective learning the motion prior for people tracking. The GPSTVM comprises a low dimensional latent space with associated spatio-temporal process. It provides more genuine embedding of the human poses from both spatial and temporal perspectives. After the prior on human pose is learned, the particle filtering can be used tracking articulated human pose; particle filtering propagates over time in the embedding space, avoiding the curse of dimensionality. The predicted human pose is projected onto image plane for measurement. Fig.1 shows the tracking algorithm.

The remainder of the paper is structured as follows. The GPSTVM and the model result are presented in Section 2. In Section 3 we describe the particle filtering integrated with learned prior for monocular 3D people tracking. Experiments are presented in Section 4. Section 5 concludes the paper and discusses future research work.

2. GAUSSIAN PROCESS SPATIO-TEMPORAL VARIABLE MODEL

The GPSTVM is obtained by incorporating the neighborhood information as a hard constraint on object function during the training stage, and by modeling a first-order Markov process in the latent space as a smooth term.

Firstly, we model a mapping from latent space $x_t \in \mathbb{R}^d$ to pose space $y_t \in \mathbb{R}^D$ ($d \ll D$) with AR model and model dynamics with first-order AR model in latent space

$$y_t = \sum_j a_j \varphi_j(x_t) + n_{y,t} \tag{1}$$

$$x_t = \sum_i b_i \phi_i(x_{t-1}) + n_{x,t} \tag{2}$$

where $A = [a_1, a_2, ...]$ and $B = [b_1, b_2, ...]$ are weight, φ_j and ϕ_i are basis function, $n_{x,t}$ and $n_{y,t}$ are additive zero-mean white Gaussian noise.

From Bayesian perspective, the specific forms of φ_j, and the weights A should be marginalized out. Marginalizing over φ_j and A can be done in close form [10] to yield a multivariate Gaussian data likelihood of the form:

$$p(Y \mid X, \tilde{\alpha}) = \frac{1}{\sqrt{(2\pi)^{ND} \mid K_Y \mid^D}} \exp(-\frac{1}{2} tr(K_Y^{-1} Y Y^T)) \tag{3}$$

where $Y = [y_1, ..., y_t, ..., y_N]^T$, $X = [x_1, ..., x_t, ..., x_N]^T$, K_Y is a kernel matrix. The elements of kernel matrix are defined by a kernel function, $(K_Y)_{i,j} = k_Y(x_i, x_j)$ which are taken as common RBF for two coordinates x and x' in latent space X.

$$k_Y(x, x') = \alpha_1 \exp(-\frac{\alpha_2}{2} \| x - x' \|^2) + \alpha_3^{-1} \delta_{x,x'} \tag{4}$$

where $\tilde{\alpha} = \{\alpha_1, \alpha_2, \alpha_3\}$ comprises the kernel hyper parameters that control the output variance, the RBF support width, and the variance of the additive noise $n_{y,t}$.

Secondly, we propose to use the neighborhood information of each point y_t in high dimensional space to model the spatial structure. For computational convenience, we assume that all these neighborhoods are linear, i.e. each data point can be optimally reconstructed using a linear combination of its neighbors [11]. Hence our objective is to minimize

$$\varepsilon = \sum_t \| y_t - \sum_{j \in N_t} w_{tj} y_j \|^2 \tag{5}$$

where N_t represents the neighborhood of y_t, and w_{tj} is the contribution of y_j to y_t. We further constrain $\sum_{j \in N_t} w_{tj} = 1$, $w_{tj} \geq 0$. Obviously, the more similar y_j to y_t, the lager w_{tj} will be. Thus w_{tj} can be used to measure how similarity y_j to y_t. One issue should be addressed here is that usually $w_{tj} \neq w_{jt}$. It can be easily inferred that

$$\varepsilon_i = \| y_t - \sum_{j \in N_t} w_{tj} y_j \|^2 \tag{6}$$

thus the reconstruction weights w_{tj} in the low dimensional space can be solved by estimating the constrained least squares.

Intuitively, the latent coordinates X should retain the spatial relationship of the high dimensional data. X is obtained by minimizing the embedding cost function

$$\eta = \sum_{t=1}^{N} \| x_t - \sum_j w_{tj} x_j \|^2 = XMX^T \tag{7}$$

where M is given by

$$M_{tj} = \delta_{tj} - w_{tj} - w_{jt} + \sum_k w_{kt} w_{kj} \tag{8}$$

here， δ_{ij} is 1 if $i = j$ and 0 otherwise. The cost function (7) can be viewed as hard constraint during optimizing the latent coordinates.

Thirdly, following [9], we model the dynamics in latent space. Incorporating the first-order Markov property and marginalizing out the parameters B and basis function ϕ_i. The density over latent space reduces to

$$p(X \mid \tilde{\beta}) = p(x_1) \frac{1}{\sqrt{(2\pi)^{(N-1)d} \mid K_X \mid^d}} \exp(-\frac{1}{2} tr(K_X^{-1} X_{out} X_{out}^T)) \tag{9}$$

where $X_{out} = [x_2, ..., x_N]^T$, K_X is the $(N-1) \times (N-1)$ kernel matrix constructed from $X_{in} = \{x_1, ..., x_{N-1}\}$ [10]. Kernel K_X uses common RBF for two data x and x' in set X_{in}:

$$k_X(x, x') = \beta_1 \exp(\frac{-\beta_2}{2} \| x - x' \|^2) + \frac{\delta_{x,x'}}{\beta_3} \tag{10}$$

2.1. Learning

Learning the GPSTVM from poses Y entails minimizing the negative log-posterior with hard constraint (7). Following [9], we adopt simple prior on the hyper parameters $p(\tilde{\alpha}) \propto \prod_i \alpha_i^{-1}$, $p(\tilde{\beta}) \propto \prod_i \beta_i^{-1}$ to discourage overfitting. Together, the priors, the latent mapping, and the dynamics define a generative model for spatio-temporal series poses.

$$p(X,Y,\tilde{\alpha},\tilde{\beta}) = p(Y \mid X,\tilde{\alpha})p(X \mid \tilde{\beta})p(\tilde{\beta})p(\tilde{\alpha}) \quad (11)$$

The latent coordinates and hyper parameters are found by minimize the negative posterior under the spatial hard constraint (7)

$$\min_{X,\tilde{\alpha},\tilde{\beta}} L = -\ln p(X,\tilde{\alpha},\tilde{\beta} \mid Y) = \frac{d}{2}\ln \mid K_X \mid + \frac{1}{2} tr(K_X^{-1} X_{out} X_{out}^T) + \sum_j \ln \alpha_j + \frac{D}{2}\ln \mid K_Y \mid + \frac{1}{2} tr(K_Y^{-1} Y Y^T) + \sum_j \ln \beta_j \quad (12)$$

$$s.t: X^T M X - \eta = 0$$

where reconstruction error η is selected manually. We have experimented with Sequential Quadratic Programming (SQP) to solve the latent coordinates. SQP allows the use of hard constraints on the object function. However, hard constraints can only be used for underconstrained function, otherwise the system quickly becomes infeasible and the solver fails. A more general solution is to convert the constraints into soft constraints by adding a term $\| X^T M X - \eta \|^2$ to the objective with a large weight [12]. Thus, object function is changed into unconstrained optimization

$$\min_{X,\tilde{\alpha},\tilde{\beta}} \tilde{L} = (1-\lambda)L + \lambda \| X^T M X - \eta \|^2 \quad (13)$$

We optimize the $X,\tilde{\alpha},\tilde{\beta}$ numerically with conjugate gradients.

2.2. Model results

The test walking sequence data comes from CMU motion capture data. It consists of 3 normal periodical walking. The dimension of data is 62. Our experimental setting is as follows. λ is equal to 0.85, and the reconstruction error η is equal to 2. The latent coordinates are initialized with PCA.

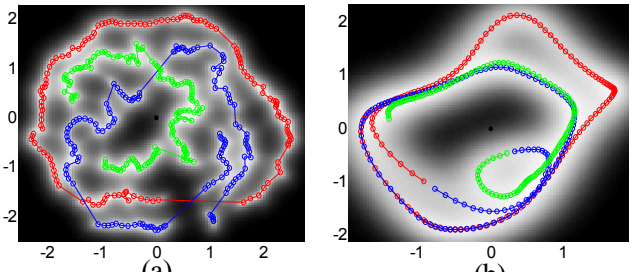

Fig. 2 Latent space of 3 walking sequences with different styles and speed. (a) GPDM, (b) GPSTVM.

Fig. 2 shows the latent space of 3 walking sequences with different styles and speeds; different color represents the different walking sequence. Note that the latent trajectories in Fig. 2(a) illustrates that GPDM produces fragmental latent space. Fig.2 (b) shows that GPSTVM produces a smooth configuration of latent positions. Spatial constraint term in object function explains it. Fig. 2 also shows a visualization of the inverse reconstruction variance, i.e. $2\ln\sigma_{y|x,X,Y,\tilde{\beta}}$. This shows the confidence with which the model reconstructs a data from a latent position x. Brighter colors correspond to lower variances; this indicates more reliable reconstruction from latent space. Thus, GPSTVM is more reliable than GPDM during mapping from latent space to original pose space.

3. TRACKING

In the application to 3D people tracking, we use particle filtering [15] integrated with the GPSTVM for tracking task. At each time instance t, the complete body pose is controlled by a state vector $s_t = (P_t, x_t)$. P_t represents the global position of the body, and x_t is the point in latent space.

We manually initialize the 3D position P_t and 3D pose y_0. Given y_0, the initial latent points x_0 can be obtained by minimizing the following likelihood function [12],

$$\bar{L}(x_t, y_t) = \frac{\| y_t - f(x_t) \|^2}{2\sigma^2(x_t)} + \frac{D}{2}\ln\sigma^2(x_t) + \frac{1}{2}\| x_t \|^2 \quad (14)$$

where $f(x) = Y^T K_Y^{-1} \overline{k_Y}(x)$, $\sigma^2(x) = k_Y(x,x) - \overline{k_Y}(x) K_Y^{-1} \overline{k_Y}(x)$. $\overline{k_Y}(x)$ is a vector with elements $k_Y(x, x_j)$ for all other latent points x_j in the GPSTVM.

We model the dynamics as second-order AR model. W-

$$x_{t+1} = Ax_t + Bx_{t-1} + Cv_t, v_t \sim N(0,\Sigma) \quad (15)$$

here the matrices A, B, C and Σ defining the dynamics are learned form motion capture data; more detail on how to learning the parameters can be found in [14]. To compute the measurement for the current prediction, first the silhouette of the current video frame is extracted through background subtraction, and the chamfer matching cost between the projected model and image silhouettes is considered to be proportional to the negative log-likelihood [13]. We use the same human model proposed by [13], which consists of a group of cylinders.

The Particle filtering based on the GPSTVM prior will be described in Algorithm. 1.

Algorithm. 1 Particle filtering based on GPSTVM

1. Initialization: manually initialize 3D pose and calculate the latent coordinates x_0 through Eq. 14
2. Prediction: sample x_t through dynamics defined by Eq. 15
3. Measurement: first map the samples x_t into corresponding pose y_t; second project the 3D pose y_t onto image plane for measurement and evaluate the samples weights.
4. Output: predict current pose.
5. Resample particles and go to step 2

4. EXPERIMENTAL RESULTS

The proposed algorithm has been tested in tracking walking humans. The test set, calibration information and ground truth are obtained from [13]. The test set has four image sequences, captured by four synchronized cameras from

different viewpoints. The challenges of the test set in tracking include large motion, motion blur, loose clothing and self-occlusions.

In the experiments, we use only one of the sequences, and maintain the 2 dimensional latent pace and 200 particles for our algorithm. One training data from motion capture data has 400 frames with 28 freedoms. We compare our algorithm against: (1) annealed particle filtering [16] and (2) particle filtering. We set annealed particle filtering with 10 layers, 100 particles per layer and four synchronized image sequences simultaneously for measurement. All algorithms run on a 3.0 GHz PC with 512 MB RAM under Matlab. Our tracking algorithm takes approximately 3 seconds per frame, while the annealed particle filtering and particle filtering take approximately 120 seconds per frame.

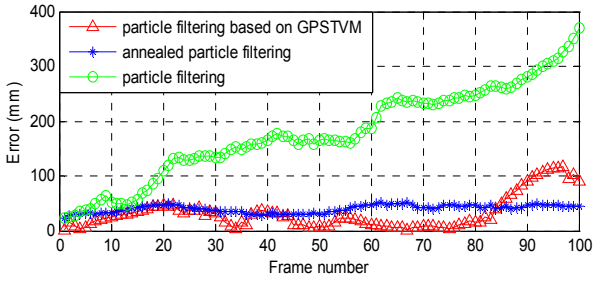

Fig. 3 Estimation error of 3 tracking algorithms

Fig. 3 shows the accuracy of the different tracking algorithms. As proposed in [13], the error is measured as the absolute distance in millimeters between the ground truth and estimated marker positions on the body limbs. As can be seen in the graph of Fig. 3, our method is consistently more robust. Based on the performance reported in [13] (up to 50 frames), our algorithm can track the walk motion longer and more accurately. The error increases rapidly after the 80-th frames, since the people changes his motion style into 'turn left' and we do not model this motion prior for tracker.

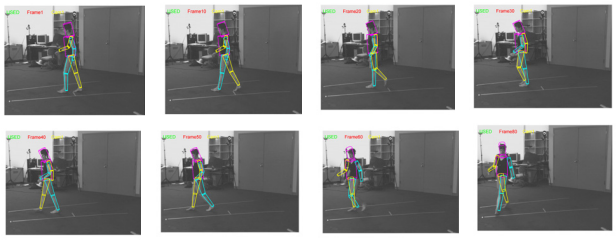

Fig. 4 Experimental results for monocular tracking 3D people. Visit website: http://www.jdl.ac.cn/user/jbpang/GPSTVM.htm for video.

Due to the limitation of space, Fig. 4 just illustrates the example tracking results. We can see that our approach tracks the walking straight successfully. In some frame, global position estimation error causes the pose estimation failure, for instance, the 60-th frame. Smarter sampling method or more particles will make up it.

5. CONCLUSIONS

In this paper, we have proposed an algorithm to track 3D people accurately, despite the self-occlusion and the noisy image measurements. Our main contribution is focused on GPSTVM, a novel model for learning human motion prior. Currently we only learn the walk prior. Essentially different motions can be learned using GPSTVM; hence more complicated motion can be tracked using same algorithm. In future, how to deal with the transition of different motion should be carried out.

6. ACKNOWLEDGEMENT

This work was supported in part by Science100 Plan of Chinese Academy of Sciences: 99T3002T03, Beijing Natural Science Foundation: 4063041 and National 242 Project: 2006A09.

7. REFERENCES

[1] A. Elgammal, C. Lee, "Inferring 3D body pose from silhouettes using activity manifold learning," Proc. CVPR, Vol. 2, pp.681-688, 2004

[2] C. Sminchisescu, A. Jepson, "Generative modeling for continuous non-linearly embedded visual inference." Proc. ICML, 2004

[3] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," IEEE Trans. PAMI, vol. 28, no.1, pp.44-58, 2006

[4] O. C. Jenkins and M.J. Mataric, "A spatio-temporal extension to isomap nonlinear dimension reduction," Proc. ICML, 2004

[5] R. Urtasun, D.J. Fleet, A. Hertzmann and P. Fua, "Priors for people tracking from small training sets," Proc. ICCV, vol. 1, pp. 403-410, 2005

[6] R. Li, M.H. Yang, S. Sclaroff and T. P. Tian," Monocular Tracking of 3D Human Motion with a Coordinated Mixture of Factor Analyzers," Proc. ECCV, vol. 2, pp. 137-150, 2006

[7] T. Tian, R. Li, S. Sclaroff, "Articulated pose estimation in a learned smooth space of feasible solutions," Proc. CVPR Learning workshop, 2005

[8] R. Urtasun, D.J. Fleet and P. Fua, "3D people tracking with gaussian process dynamical models," Proc. CVPR, vol. 1, pp.238-245, 2006

[9] J. Wang, D.J. Fleet and A. Hertzmann, "Gaussian process dynamical models," Proc. NIPS, 2005

[10] N.D. Lawrence,"Gaussian process latent variable models for visualization of high dimensional data," Proc. NIPS, 2004

[11] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding" Science, vol. 290, pp. 2323-2326, 2000

[12] K. Grochow, S. Martin and A. Hertzmann,"Style-based inverse kinematics," SIGGRAPH, pp.522-531, 2004

[13] L. Sigal, S. Bhatia, S. Roth and M.J. Black, "Tracking loose-limbed people," Proc. CVPR, vol.1, pp.421-428, 2004

[14] J. Pang, Q. Huang and S. Jiang," Monocular tracking 3D people with back constrained scaled Gaussian process latent variable models", Proc. Asia-pacific workshop on visual information processing, pp. 1119-125, 2006

[15] P. Perez, C. Hue, J. Vermaak, M. Gangnet, "Color-based probabilistic tracking," Proc, ECCV, pp. 661-675, 2002

[16] J. Deutscher and I. Reid,"Articulated body motion capture by stochastic search". IJCV, vol. 61, pp.185-205, 2004