LEARNING LOCAL MODELS FOR 2D HUMAN MOTION TRACKING

Wenzhong Wang[1,2], *Xiaoming Deng*[1], *Xianjie Qiu*[1], *Shihong Xia*[1], *and Zhaoqi Wang*[1]

[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]Graduate School of Chinese Academy of Sciences
{wangwenzhong, dengxiaoming, qxj, xsh, zqwang}@ict.ac.cn

ABSTRACT

We present a novel approach to tracking 2D human motion in uncalibrated monocular videos. Human motion usually exhibits time-varying patterns, and we propose to use locally learnt prior models to capture this characteristics. For each input image, our method automatically learns a local probability density model and a local dynamical model from a set of training examples that are close matches to the input. We evaluate the image likelihood by matching a deformable 2D human body model to the input images. The local models and the image likelihood are integrated to optimize the pose for the current input. Experiments on both synthetic and real videos demonstrate the effectiveness of our method.

Index Terms— Motion Tracking, Local Learning

1. INTRODUCTION

Estimating body poses from monocular images is important for human motion analysis, motion recognition, and visual surveillance. Due to self-occlusions, image noises, and motion variations, it is hard to find optimal solutions directly from monocular videos. Consequently recent research often incorporate some prior knowledge (e.g. joint angle limits) to guide the searching in the solution space.

These works typically model prior densities for static poses or temporal dynamics for motion sequences. These models include Gaussian Mixture Models [1, 2], density estimation [3], Hidden Markov Models [4], autoregressive process [5] and Gaussian Process Dynamical Models [6]. Some of these models implicitly assume that the dynamics is time invariant which is not hold in reality. In [7], Ankur Agarwal et al cluster the body poses into connected regions exhibiting similar dynamical patterns and modeling the dynamics in each region as a Gaussian autoregressive process. However, the number of clusters must be carefully chosen, and modeling the inter-cluster transitions is non-trivial.

In this paper, we propose a novel scheme to incorporate prior knowledge into the estimation process (Fig.1). The key idea is that the human motion exhibits time-varying characteristics which should be better modeled using time-varying models(Fig.2). We build a motion database consists of both human motion and the 2D silhouettes for each pose in the motion. Each data point in the database is a pair of pose x and silhouette feature s: $D = \{(x_i, s_i)\}$. For each input, we search the motion database for examples that are close to the input query. These examples constitute the local region for this query. Our method automatically learns a local density model and a local dynamical model in this local region for each input. These two models together with the image likelihood are integrated into an optimization process to produce a pose which can explain the input silhouette. In this paper we focus on side-view human motion tracking. Given an input silhouette z, we try to estimate the underlying 2D body configuration x. In the probabilistic perspective, our method maximizes the a posteriori (MAP) $p(x|z_t)$ at each time instant t.

This work is supported by the National Key Technology R&D Program of China (2008BAI50B07), the State Key Program of National Natural Science Foundation of China (60533070), China Postdoctoral Science Foundation, the K. C. Wong Education Foundation, Hong Kong, and the Co-building Program of Beijing Municipal Education Commission.

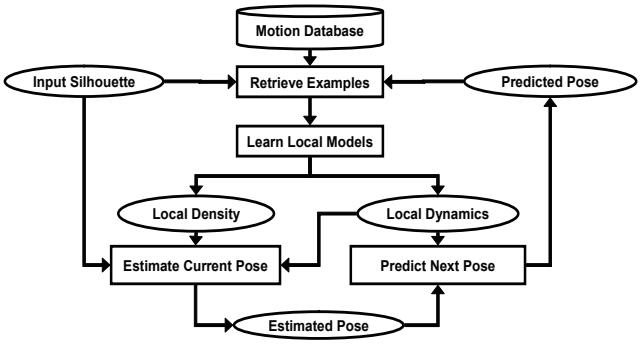

Fig. 1. Overview of the estimation process.

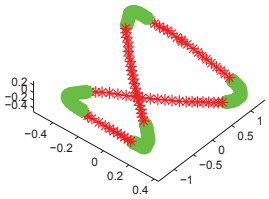

Fig. 2. Time-varying motion dynamics. Plotted is a human walking cycle projected onto the first 3 principle components. The walking motion exhibits different temporal patterns in the green and red phases which should be better captured with different models.

2. ESTIMATION METHOD

2.1. Representing Human Body and Images

The selection of human body models is important to estimate the likelihood of the image. We build *deformable* 2D human models directly from the real silhouette images. This model consists of five deformable parts: two legs, two arms, and one torso (including head and neck). Each body part is described by three joints and a closed

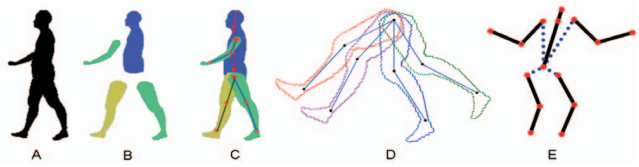

Fig. 3. 2D body model. In this example, the missed right arm is replicated from the left arm.

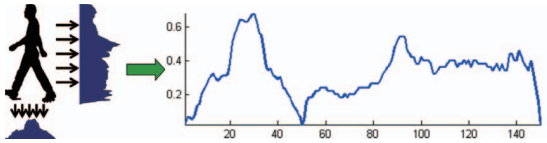

Fig. 4. Image Features.

contour which is (manually) extracted from a real image. Linking the joints yields two bones for each part. The body part is deformed using the *Skeleton Subspace Deformation* technique[8]. Fig.3 shows the construction process: The body parts(B) are extracted from a silhouette(A). All the five parts are assembled to yield a single body model(C). The contour deforms according to the articulation of the bones(D). The body model is parameterized as a kinematic tree(E).

The limb parts are connected to the root using virtual bones (dashed lines in Fig.2 E). The body model has 30 parameters: the bone lengths $(l_1, \cdots, l_{14})$, bone directions $(x = (x_1, \cdots, x_{10})$, $v = (v_1, \cdots, v_4))$ (x_i denotes the directions of real bones depicted as hard lines in Fig.2 E, these directions are called *joint angles* in the following text. v_i denotes the directions of the virtual bones and are called *virtual angles*), and the root position $p = (p_x, p_y)$. In our current implementation, the bone lengths are fixed and calculated in the modeling process. So our model has 16 degrees-of-freedom (DOF) $(x = (x_1, \cdots, x_{10}), v = (v_1, \cdots, v_4), p = (p_x, p_y))$. Note that the local models are learned with only the joint angles x.

The advantage of this model is that it can be deformed in accordance with the underlying motion in a relatively natural way. This model retains some degrees of the realistic details of the human body silhouette under consideration, and thus provides more accurate estimation of the image likelihood.

The image feature used in this work is the width and height profile[8]. We project a silhouette along its columns and rows, and obtain two histograms (height profile and width profile). These histograms are then normalized by dividing them with the width and height of the bounding box of the silhouette, respectively. In this work, the width and height profiles are sampled to 100 and 50 bins, respectively, and then concatenated to yield a single 150-bin histogram(Fig.4).

2.2. Learning Local Models

Our local models include local densities and local dynamical models.

The local dynamics is modeled as a second order Gaussian autoregressive process (ARP) [7, 5]:

$$x_t = A_t x_{t-1} + B_t x_{t-2} + v_t, v_t \sim N(0, Q_t) \quad (1)$$

where A_i and B_i are $m \times m$ $(m = 10)$ matrices giving the influences of x_{t-1} and x_{t-2} on x_t, and v_t is a Gaussian noise.

This gives the following dynamical prior on pose x_t:

$$p_d(x_t|x_{t-1}, x_{t-2}) = N(A_t x_{t-1} + B_t x_{t-2}, Q_t) \quad (2)$$

We also model the statical prior which restricts the solutions to plausible body poses. This prior is expressed as a probability density. We assume that in an arbitrary time instant, the poses in the local region are normal distributed [9]:

$$p_s(x_t) = N(r_t, R_t) \quad (3)$$

where r_t and R_t are the mean vector and covariance matrix of all poses in the local region. The dynamical prior and statical prior are then combined to yield the local prior [2]:

$$p(x_t) = p_s(x_t) p_d(x_t|x_{t-1}, x_{t-2}) \quad (4)$$

Estimating model parameters: To construct the local models, we search the motion database for examples similar to the input query q. The distance between q and the training point p is calculated as the weighted sum of the pose distance and silhouette distance:

$$d(q, p) = \alpha d_x(x_q, x_p) + (1 - \alpha) d_s(s_q, s_p) \quad (5)$$

where d_s is the Euclidean distance between the feature vectors of two silhouettes, and the pose distance d_x writes:

$$d_x(x, y) = \frac{1}{m} \sum_{i=1}^{m} \sqrt{(\sin x_i - \sin y_i)^2 + (\cos x_i - \cos y_i)^2} \quad (6)$$

Due to the inherent ambiguities of silhouettes(Fig.5, four different poses in the first row can generate very similar silhouettes in the second row.). We expect to retrieve examples distributed in different areas of the pose space which are far apart from each other. In consequence, the local models learnt from these samples prove to be meaningless. Alternatively, we may use poses as queries to retrieve similar examples. However, a single pose is itself ambiguous since it encodes no motion velocity information. In the case of side-view motion, it can not tell whether a person is stepping with the left leg or the right leg(the 3rd row of Fig 5). To get around this, we use a small chunk of consecutive poses to evaluate the similarity between the query pose and the training poses. Likewise, we use a small chunk of consecutive silhouettes to measure the similarity between the query silhouettes and training ones (this chunk-based shape similarity is still ambiguous, in the case of side-view walking motion shown in Fig.5, it has the same ambiguity as a single pose has). Note that using only pose-chunk similarity can retrieve examples without ambiguities; the silhouette-chunk similarity is used to bias the overall similarity so that the retrieved silhouettes are close to the input ones. We find that this treatment works well in the experiments, and the selected training examples always form consistent local regions.

The consistency of the local region is of great importance for the success of our method, since data from different parts of the pose space may have distinct dynamical patterns and probability distributions. Models learnt from such an inconsistent data set would average over different pose classes, and cannot describe the local characteristics of any pose classes.

The weight α in Eqn.(5) is selected empirically, we typically set $\alpha = 0.8$. The d_s's are normalized as follow:

$$d_s(s_q, s_p) = \frac{\|s_q - s_p\| - \min_{s \in D} \|s_q - s\|}{\max_{s \in D} \|s_q - s\| - \min_{s \in D} \|s_q - s\|} \quad (7)$$

The d_x's are normalized likewise.

For the query q, we find N closest training examples $\{(x_k, s_k)|k = 1 \cdots N\}$. These examples form the *local region* (LR) for the current query. To ensure that the local models are properly estimated,

Fig. 5. Ambiguities of single silhouette and pose.

the local region must contain sufficient training examples. Otherwise, the models will be under-fitted.

The static prior model is estimated using all poses in this local region (Eqn.(3)).

To estimate the parameters of dynamical model, for each pose x_k in the local region, we extract a small motion clip T_k from the original motion sequence containing x_k. Let x_k be the ith pose $x_k(i)$ in the motion sequence, then T_k is a small motion patch consists of $2l+1$ consecutive poses $\{x_k(i-l), \cdots, x_k(i), \cdots, x_k(i+l)\}$. ($l = 5$ in our implementation). We use these small motion clips to estimate the parameters of the dynamical model:

We rewrite the linear part of equation (1) as[1]:

$$x_t = C_t \tilde{x}_t \tag{8}$$

where $C_t = (\ A_t \quad B_t\ ), \tilde{x}_t = (x_{t-1}^T, x_{t-2}^T)^T$

We then construct two matrices U_k and V_k from T_k:

$$U_k = (x_{k,3}, \cdots, x_{k,2l+1}) \tag{9}$$

$$V_k = \begin{pmatrix} x_{k,1} & x_{k,2} & \cdots & x_{k,2l-1} \\ x_{k,2} & x_{k,3} & \cdots & x_{k,2l} \end{pmatrix} \tag{10}$$

where $x_{k,i}$ is the ith pose in T_k.

We stack the U_k and V_k to obtain two matrices U and V:

$$U = (U_1, \cdots, U_N), V = (V_1, \cdots, V_N) \tag{11}$$

Then C_t is estimated by solving the following problem:

$$U = C_t V \tag{12}$$

This is solved as a regularized least squares problem and the regularizing parameter is determined by cross-validation. Once A_t and B_t are found, the covariance Q_t is estimated from the residuals between $\{x_t\}$ and $\{A_t x_{t-1} + B_t x_{t-2}\}$.

2.3. Image Matching Likelihood

We choose to use the sum-of-squares error to measure how well a hypothesized body model explains the observed silhouette:

$$E_{likelihood}(x, v, p) = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (C_M(i,j) - C_I(i,j))^2 \tag{13}$$

where M and N are the dimensions of the images, C_I and C_M are Distance Transform images of the observed silhouette and the model silhouette, respectively.

The image likelihood gives the probability of observing image z given the 2D body configuration:

$$p(z|x, v, p) \propto exp(-E_{likelihood}) \tag{14}$$

[1] x's are column vectors.

2.4. Tracking Framework

Our goal is to maximize the posterior probability (MAP) of pose x given the input image z, i.e.

$$x_t = argmax_x p(x|z_t) \tag{15}$$

Applying the *Bayesian* theory, we have

$$p(x_t|z_t) = p(z_t|x_t)p(x_t) = p(z_t|x_t)p_s(x_t)p_d(x_t|x_{t-1}, x_{t-2}) \tag{16}$$

So our aim is to minimize the following energy function:

$$O(x, v, p) = aE_{likelihood} + bE_{static} + cE_{dynamic} \tag{17}$$

where a, b and c are weights to balance the contributions of different energy terms to the objective, and

$$E_{static} \propto -logp_s(x_t) \tag{18}$$

$$E_{dynamic} \propto -logp_d(x_t|x_{t-1}, x_{t-2}) \tag{19}$$

The first term in Eqn(17), E_{static}, measures the a-prior likelihood of the current pose. This term restricts the synthesized pose to satisfy the probability distribution determined by the training examples in the local region [9]. This term can also conform the unobservable DOF's in a plausible range.

The second term, $E_{dynamic}$, measures the temporal smoothness of the estimated motion up to current time instant. It is used to ensure the temporal smoothness of the synthesized motion.

The third term, $E_{likelihood}$, indicates how well the hypothesized pose explains the input silhouette. This item ensures that the estimated pose can generate a silhouette that is close to the input one.

Algorithm for pose estimation:

Input: shape features $\{s_{t-k}, \cdots, s_t\}$[2], predicted pose y_t, previous estimations p_{t-1}, v_{t-1}, $\{x_{t-k}, \cdots, x_{t-1}\}$, local dynamical model $\{A_t, B_t, Q_t\}$, previous and current silhouettes I_{t-1}, I_t.

Output: pose estimation $\{x_t, v_t, p_t\}$, predicted pose y_{t+1}, local dynamical model $A_{t+1}, B_{t+1}, Q_{t+1}$.

Step 1: Retrieve Local Region from the database using $\{s_{t-k}, \cdots, s_t\}$, $\{x_{t-k}, \cdots, x_{t-1}\}$ and y_t.

Step 2: Estimate local density parameters r_t and R_t.

Step 3: Estimate local dynamical parameters $\{A_{t+1}, B_{t+1}, Q_{t+1}\}$.

Step 4: Estimate pose (x_t, v_t, p_t):

Initialization: Let $p^0 = p_{t-1} + COM(I_t) - COM(I_{t-1})$. (COM=Center Of Mass); $v^0 = v_{t-1}$; $x^0 = y_t$.

Optimization: $(x_t, v_t, p_t) = argmin_{x,v,p} O(x, v, p)$.

Step 5: Predict next pose: $y_{t+1} = A_{t+1}x_t + B_{t+1}x_{t-1}$.

---

3. EXPERIMENTS AND RESULTS

We select 12 walking motion sequences from the CMU Mocap dataset (*http://mocap.cs.cmu.edu*). These motions are performed by 15 different scanned 3D human body models(7 males and 8 females, we try to introduce shape variations into the training set using different body models), and each performance is projected from side views to produce a series of 2D poses as well as the corresponding silhouettes, resulting in a dataset consists of 180 2D motion sequences. We select 18 series from this dataset to form the validation set, and the remainder are used to train the models (for

[2] we set the length of a chunk as $k = framerate/5$.

each validation sequence, we further drop out all training sequences generated using the same motion with different human models from the training set, so the training set contains no identical motions with the validation sequences). Our training database consists of about 62000 frames of 2D pose.

We conduct experiments on the validation set to verify the effectiveness of our method. The results are illustrated in Fig.6(A,B,C). The mean errors of joint angle are 2.5021 degrees. Fig.6C shows the ground truth (blue) and the estimated motion (red) projected onto their first three principle components. The two trajectories exhibit very similar shapes meaning that the estimated motion successfully captures the underlying dynamics of the original motion.

We compares the performances of our local dynamical models (LD) with a global, second order Gaussian ARP model (GD) [5], and our deformable human body model (DM) with the *Cardboad* body model (CM) [10]. The results are illustrated in Fig.6(D). We learn a single ARP model for the 2D human walking using the whole training set. Experiments show that this global model is unreliable for tracking long sequences. In our implementation of global models, the error accumulates rapidly and the tracking fails after 60 frames. Our local dynamical model in conjunction with the Cardboad body model can successfully track through the whole sequence (about 290 frames) with acceptable errors. The combination of our local model and deformable body model gives the best results.

We also conduct experiments on several real videos. Fig.7 shows the tracking results on three of them(The first video is obtained from [11], others from the *CASIA Gait Database* [12]). The video silhouettes are manually segmented, and the joint locations in the first two frames are manually labeled. For each video, the motion database is down sampled to meet the different frame rate and motion speed of the video. We also resize the input silhouettes so that they are of the same height. Our method produces visually plausible motions. In the case of self-occlusions, the method gives reasonable joint angle estimations. However, it cannot correctly estimate the virtual angles, since no prior knowledge or observable information is available for the occluded limbs.

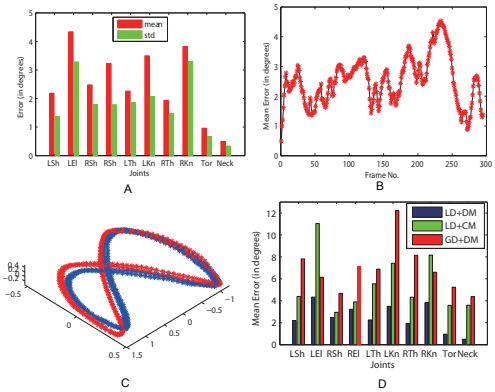

Fig. 6. Tracking results of validation sets

4. CONCLUSION

In this work, we have proposed a learning-based method to recover 2D human motion from monocular side-view image sequences. Our approach learns a series of local prior models from a set of examples. These local models express the prior knowledge about the pose

Fig. 7. Tracking results for real videos.

to be recovered as time-varying probability densities and temporal conditionals. We demonstrate the effectiveness of the method with experiments on both synthetic and real videos.

The success of our method strongly depends on the accuracy of the local models. Properly learning these models requires a large amount of training data. Our method is demanding of computational time since every time a query arrives, a local model must be constructed from the ground.

As for future work, the human body model needs to be augmented with appearance model. We intend to improve the efficiency of our method. We are also interested in applying our method to tracking more complex motions and recovering 3D human motion from silhouettes.

5. REFERENCES

[1] N. R. Howe, M. E. Leventon, and W. T. Freeman, "Bayesian reconstruction of 3d human motion from single-camera video," in *NIPS*, 1999.

[2] Tobias Jaeggli, Esther Koller-Meier, and Luc Van Gool, "Learning generative models for monocular body pose estimation," in *ACCV*, 2007.

[3] T. Brox, B. Rosenhahn, D. Cremers, and H. Seidel, "Nonparametric density estimation with adaptive, anisotropic kernels for human motion tracking," in *Workshop on Human Motion*, 2007.

[4] M. Brand, "Shadow puppetry," in *ICCV*, 1999.

[5] A. Agarwal and B. Triggs, "Learning to track 3d human motion from silhouettes," in *ICML*, 2004.

[6] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *PAMI*, 2008.

[7] A. Agarwal and B. Triggs, "Tracking articulated motion using a mixture of autoregressive models," in *ECCV*, 2004.

[8] Y. Liu, R. Collins, and Y. Tsin, "Gait sequence analysis using frieze patterns," in *ECCV*, 2002.

[9] J. Chai and J. K. Hodgins, "Performance animation from low-dimensional control signals," in *SIGGRAPH*, 2005.

[10] S. X. Ju, M. J. Black, and Y. Yacoob, "Cardboard people: A parameterized model of articulated image motion," in *FGR*, 1996.

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

[12] L. Wang, H. Ning, W. Hu, and T. Tan, "Gait recognition based on procrustes shape analysis," in *ICIP*, 2002.