Body Pose Tracking From Uncalibrated Camera Using Supervised Manifold Learning

Chan-Su Lee[*]
Department of Computer Science
Rutgers University
Piscataway, NJ 08854
chansu@cs.rutgers.edu

Ahmed Elgammal
Department of Computer Science
Rutgers University
Piscataway, NJ 08854
elgammal@cs.rutgers.edu

Abstract

We present a framework to estimate 3D body configuration and view point from a single uncalibrated camera. We model shape deformations corresponding to both view point and body configuration changes through the motion. Such observed shapes present a product space (different configurations × different views) and therefore lie on a two dimensional manifold in the visual input space. The approach we introduce here is based on learning the visual observation manifold in a supervised manner. Instead of learning an embedding of the manifold, we learn the geometric deformation between an ideal manifold (conceptual equivalent topological structure) and a twisted version of the manifold (the data). For the case of a walking motion, we use a torus manifold to represent the data. In our experiment, we learned the torus manifold from synthetic data and estimated view and body configuration for circular walking sequence in HUMANEVA-I data set. 3D body pose was inferred from estimated body configuration parameter using only one training 3D body pose cycle in each subject. Experimental results show accurate estimation of 3D body pose and view from a single camera.

1 Introduction

The approach we introduce here is based on learning the visual observation manifold in a supervised manner. Traditional manifold learning approaches are unsupervised where the goal is to find a low dimensional embedding of the data. However, if the manifold topology is known the manifold learning can be formulated in a different way. Manifold learning is then the task of learning a mapping from/to a topological structure to/from the data where that topological structure is homeomorphic to the data.

In this paper we argue that this supervised setting is suitable to model human motions that lie intrinsically on a one dimensional manifolds whether closed and periodic such as walking, jogging, running, etc., or open such as golf swing, kicking, etc. We show that we can model the visual manifold of such motions (in terms of shape) as observed from different view points by mapping such manifold to a torus manifold.

The evaluation shows that the proposed model can achieve an average of about 30mm error in the recovered body joints' positions *without any training on observations of the subjects* from the HUMANEVA-I dataset. Only one cycle of motion captured data is needed for each subject to learn the mapping from our intrinsic representation (the torus) to the 3D joint positions.

[*] http://www.cs.rutgers.edu/c̃hansu/

2 Modeling View and Configuration: Torus Manifold Embedding

Modeling both the view and body configuration manifolds for human motion jointly is a very challenging task as it requires modeling two continuous manifolds in a joint space. For example if we consider learning the shape manifold of a person walking, a one dimensional manifold motion, observed from different view points along a view circle at fixed camera height. Such setting, although limited, is very useful for many applications such as surveillance, as well as in sport analysis, where walking and running are the most frequent motions. Such observed shapes present a product space (different configurations × different views) and therefore lie on a two dimensional manifold in the visual input space. Assume we have a dense sampling of such data, we can model the view and body configuration manifolds in two orthogonal axis on a two dimensional manifold. For a periodic motion such as gait, since the data we consider are two dimensional where both the view and the configuration are closed manifolds, this is topologically equivalent to a torus which is also a two dimensional manifold embedded in a three dimensional Euclidean space.

Given the torus structure, learning the manifold is then the problem of learning a generative mapping from the torus to the data. In other words, it is learning the geometric deformation between an ideal manifold (the torus) and a twisted version of the manifold (the data). This can be achieved through learning a nonlinear mapping through generalized Radial basis function mapping as was shown in [1].

We can represent any point on the torus manifold by a function g of the two variables μ and ν as $[x\ y\ z] = g(\mu, \nu)$. Given input shapes with all views and all poses $\boldsymbol{y}_{vb}$ and their corresponding torus embedding $\boldsymbol{x}_{vb}$, we can learn a nonlinear mapping in the form

$$\boldsymbol{y}_{vb} = \boldsymbol{D} \cdot \psi(\boldsymbol{x}_{vb}) = \boldsymbol{D} \cdot \psi(g(\mu_v, \nu_b)) = \boldsymbol{D} \cdot \varphi(\mu_v, \nu_b). \tag{1}$$

Such model can be learned by solving a linear system as in [1]. Using this model, for any view given v and body configuration b sequence, we can generate a new observations where μ_v is view representation in the μ axis, and ν_b is body configuration representation in ν axis of the torus manifold.

3 Tracking on Torus Manifold Using Particle Filtering

The Bayesian tracking framework enables recursive update of the posterior $P(\boldsymbol{X}_t|\boldsymbol{Y}^t)$ of the object state $\boldsymbol{X}_t$ given all observations $\boldsymbol{Y}^t = \boldsymbol{Y}_1, \boldsymbol{Y}_2, .., \boldsymbol{Y}_t$ up to time t:

$$P(\boldsymbol{X}_t|\boldsymbol{Y}^t) \propto P(\boldsymbol{Y}_t|\boldsymbol{X}_t) \int_{\boldsymbol{X}_{t-1}} P(\boldsymbol{X}_t|\boldsymbol{X}_{t-1}) P(\boldsymbol{X}_{t-1}|\boldsymbol{Y}^{t-1})$$

We can update the state posterior based on observation likelihood estimation with transition probability and previous time step state posterior, where $P(\boldsymbol{Y}_t|\boldsymbol{X}_t)$ is the observation (measurement) model and $P(\boldsymbol{X}_t|\boldsymbol{X}_{t-1})$ is the dynamic model. The generative model in Eq. 1 fits directly to the Bayesian framework to generate observations from states to estimate the observation model $P(\boldsymbol{Y}_t|\boldsymbol{X}_t)$. The state is represented by the view parameter μ and the configuration parameter ν, i.e., $\boldsymbol{x}_t = (\mu_t, \nu_t)$.

It is important how well the state represents the motion in a low dimensional space. Given the generative model, we have the two dimensional manifold representing the view and configuration. Utilizing such manifold embedding, we can constrain tracking to a two dimensional tracking space with a pure linear dynamic system which leads to efficient tracker.

We use particle filter to realize the sequential estimation of view and body configuration. Each particle represent a point on the torus manifold, which has corresponding view and body configuration parameter. We represent a view and body configuration combination by N_β particles $\{\boldsymbol{x}_t^{(i)}, \pi_t^{(i)}\}_{i=1}^{N_\beta}$. In order to estimate likelihood for each particle, we first generate shape for each given manifold point using Eq. 1. Updating of weights of each particle can be achieved by measuring similarity between the observed shape and the synthesized shape from the generative model.

4 Experimental Results

Data Collections: We generated synthetic training data of walking silhouette from motion capture data using animation software Poser®. 36 different views $(10^o, 20^o, \cdots, 360^o)$ are collected

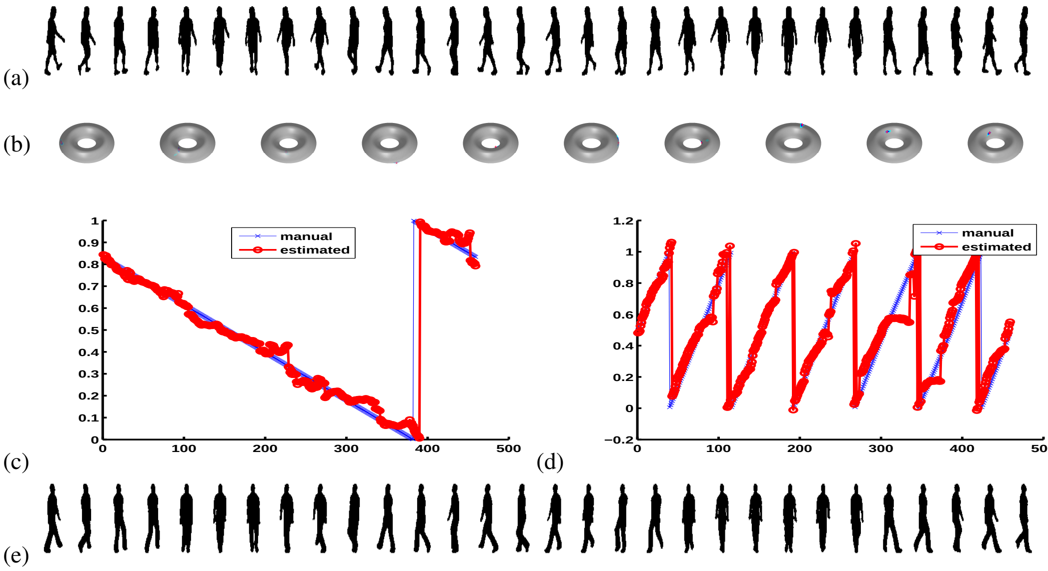

Figure 1: Reconstruction of 3D body pose: (a) Input silhouettes. (b) Trajectory on torus manifold according to view and configuration change. (c) Estimated view parameter μ. (d) Estimated body configuration parameter ν. (e) Estimated MAP synthesized silhouette.

for a cycle motion capture data. The motion captured data used for synthesis is not part of the HUMANEVA-I dataset. Given this synthetic data we learn the manifold deformation from the torus as was described in section 2

Evaluation Data: We extracted silhouette sequence for validation frames of each subject in HUMANEVA-I dataset [2]. For extracted silhouette frames, we performed background subtraction after learning background model using nonparametric kernel density estimation [3]. The extracted foreground image is normalized by height and center of foreground blobs. Joint locations of validation set and of one cycle training sequence are extracted and normalized to represent *normalized pose* which is invariant to subject's rotation and translation. We achieved this pose normalization by computing joint location after rotating each joint transformation into body centered coordinate and re-centering translation based on mean node location in each frame. We collected three subjects' validation sequences to estimate the performance of inferring 3D body pose. Table 1 shows frames used for performance test in each subject. Fig. 1 (a) shows example silhouette images for subject 'S1'.

Body Pose Estimation: We estimate body pose from the maximum a posterior (MAP) estimation of body configuration in the particle filtering. Fig. 1 (d) shows estimated body configuration parameters. We compared the estimated parameters with approximated compute body configuration parameter using manual marked cycles. We also estimated view parameters from MAP embedding points simultaneously as in Fig. 1 (c). We used 600 particles ($N_\beta = 600$) in the experiment to represent view and body configuration on the torus manifold. In the estimated body configuration parameter, it shows increase in error around 350th frames where the camera is front view of the subject's walking directions. We may reduce the errors in the front or back view if we estimate body configuration from multiple camera by fusing estimated body configuration from different camera.

We reconstructed 3D body pose from estimated body configuration parameter and one 3D body pose cycle from training sequence for each subject. We learned mapping between body configuration parameter and 3D body pose from the selected training sequence. After that, we can infer 3D body pose for any estimated body configuration parameter. Fig. 2 shows measured reconstruction error in each joint locations (a) and average error in each frame (b) for subject 'S1'. True and estimated limb endpoints are compared in Fig. 2(d)(e). Table 1 shows average errors in each subject's test sequence (We discard *HeadProximal* location of subject $S3$ during the error estimation as the validation data has inconsistency.). At first, we tested performance without considering dynamics of the body configuration and view. In this case, particles just drift randomly in each frame. Then, we applied dynamics of body configuration state and view state assuming constant velocity for view and body configuration on the torus manifold. We estimate the velocity of body configuration ν by average cycle number in the training sequence. For the view velocity, we compute view state μ's velocity by frame number to circle the whole cycle. Experiments show better body configuration estimation when we count dynamics of the particle propagation. Overall average error in each joint location from three subjects is 31.36 millimeters (mm) with dynamics in the particle propagation.

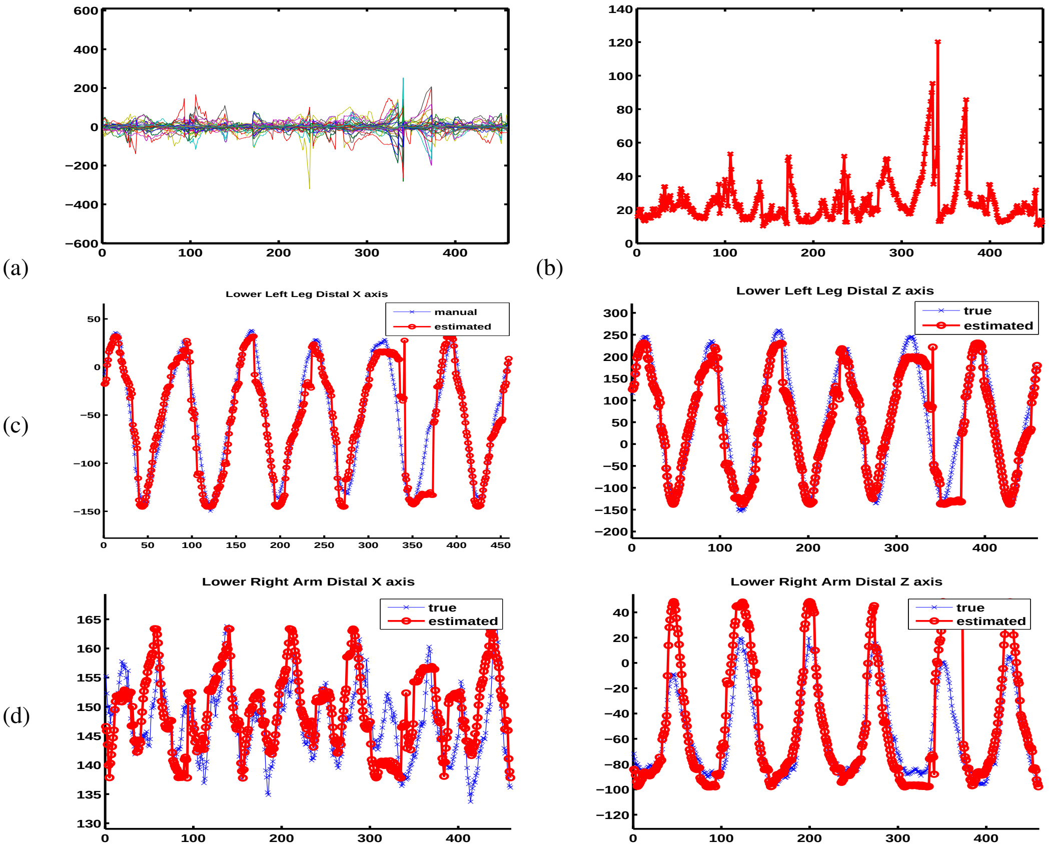

Figure 2: Error measurement of 3D body pose reconstruction for S1 :X-axis: frame number, Y-axis: joint location value (unit:mm). (a) Reconstruction errors in each joint location. (b) average joint location error in each frame. (c)(d) True and estimated joint location x and z values for *Lower left leg distal* and *Lower right arm distal*. (e) Estimated MAP synthesized silhouette.

Table 1: Average error in normalized 3D body pose estimation from single camera

Subject	**Start**	**End**	**Duration**	**Cycle**	**Mean Error(No Dyn.)**	**Mean Error(With Dyn.)**
S1	76	534	459	6	26.16 mm	24.71 mm
S2	21	436	416	5	37.11 mm	31.16 mm
S3	91	438	348	5	40.47 mm	38.21 mm
S1,S2,S3					34.58 mm	31.36 mm

References

[1] Ahmed Elgammal. Nonlinear manifold learning for dynamic shape and dynamic appearance. In *Workshop Proc. of GMBV*, 2004.

[2] Leonid Sigal and Michael J. Black. Humaneva: Cynchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, 2006.

[3] Ahmed M. Elgammal, David Harwood, and Larry S. Davis. Non-parametric model for background subtraction. In *ECCV '00: Proceedings of the 6th European Conference on Computer Vision-Part II*, pages 751–767, London, UK, 2000. Springer-Verlag.