

# Tracking People on a Torus

Chan-Su Lee and Ahmed Elgammal  
Department of Computer Science  
Rutgers University

{chansu, elgammal}@cs.rutgers.edu

May 10, 2007

## Abstract

We present a framework to track and 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  $\times$  different views) and lie on a low dimensional manifold in the visual input space. The approach we introduce here is based on learning both the visual observation manifold and the kinematic manifold of the motion 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). We use a torus manifold to represent such data for both periodic and non-periodic motions. Experimental results show accurate estimation of 3D body pose and view from a single camera.

## 1 Introduction

Tracking human body and recovery of 3D body pose is a challenging problem for human motion analysis with many applications such as visual surveillance, human-machine interface, and gesture recognition. Traditionally, this problem has been addressed through generative approaches that map from 3D body configuration space to the visual observation space e.g. [11]. Therefore, the recovery of the 3D configuration is formulated as a search problem for the best configuration that minimizes an error metric given the visual observation, e.g. [12, 15]. Such approaches typically requires a body model and a calibrated camera in order to obtain hypothesis observations from configurations. Similarly, 2D view-based body models can be used, however, this is limited in dealing with continuous view variability. Alternatively, discriminative approaches have been suggested which learn mapping functions from the visual observation to the 3D configuration [13, 5, 1, 18]. However such mapping is hard to constrain and therefore hard to generalize to unseen observations. Recently, researchers [2, 3, 17, 10, 20, 9] have increasing interest into constraining the problem by exploiting the fact that despite the high dimensionality of the body configuration space, many human motion activities lie intrinsically on low dimensional manifolds. Similarly, the observed motion, in terms of body shape contours, or other feature configurations, lies on a low dimensional manifold as well (visual manifold). Exploiting such property as well as the relation between the configuration manifolds and the visual manifolds are essential to constrain the solution space for many problems such as tracking, posture estimation, and activity recognition. However, it is not clear what is the best way to exploit such manifold constraints. Is it through learning the visual observation manifold or the body configuration manifold in the 3D configuration space, or both?. 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

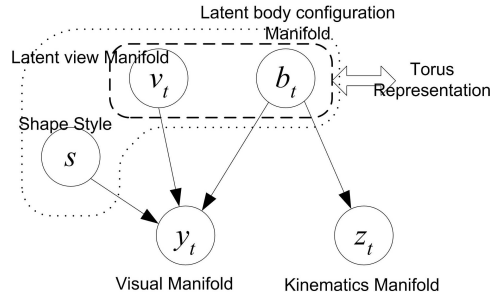


Figure 1: Graphical Model

supervised setting is suitable to model human motions that lie intrinsically on a one dimensional manifold whether closed and periodic such as walking, jogging, running, etc., or open such as golf swing, kicking, tennis serve, 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.

## 2 Framework

Consider a motion observed from a camera (stationary or moving). Such motion can be represented as a kinematic sequence  $Z^T = z_1, \dots, z_T$  and observed as a sequence of observation  $Y^T = y_1, \dots, y_T$ . In this paper, by observation, we mainly mean shape contours. With an accurate 3D body model, camera calibration, and geometric transformation information, we can explain  $Y^T$  as a projection of an articulated model. The dynamic sequence  $Z^T$  lies on a manifold, let's call it kinematic manifold. Also, the observations lie on a manifold, visual manifold. In fact, observations are lying on a product manifolds, the body configuration and the view manifolds.

What is the relation between the kinematic manifold and the visual input manifold. We can think of a graphical model connecting the two manifolds through two latent variables: body configuration variable,  $b_t$  and a view point variable,  $v_t$ . The body configuration variable is shared between both the kinematic manifold and the visual manifold. The view point variable represents the relative camera location to a human centered coordinate system. Another variable affecting the observation is the shape variability among different subjects, i.e., human shape space, or shape style as denoted by [4]. We denote this variable by  $s$ , which is time invariant variable.

We can summarize our goals as follows:

- 1) We need to relate the kinematic manifold with the visual input manifold in order to be able to infer configuration from input
- 2) We need model the visual manifold with all its variabilities due to the motion, the view point, and shape style. In particular, we need to be able to deal with both body configuration and view points as a continuous variables. This facilitates tracking subjects with varying view points due to camera motion or changing subject view w.r.t. the camera.
- 3) We need the tracking state space to be low dimensional and continuous. Moreover, and despite the nonlinearity in dynamics in both the kinematics and the observations, we need the model to exhibits simple dynamics, i.e., linear dynamics or even constant speed dynamics

So, let us start with a simple periodic motion such as a simple aerobic exercise or gait, observed from a view circle around the person. Later we show how to deal with more complex motions and also extend to the whole view sphere. Given a set of observed shapes representing a product space of two one-dimensional manifolds representing body configuration and view, how can we learn a useful representation. Nonlinear manifold learning techniques, such as LLE [14], Isomap [19], etc., have been popular recently in learning low dimensional representations of both visual and kinematic data. Unfortunately such techniques are limited

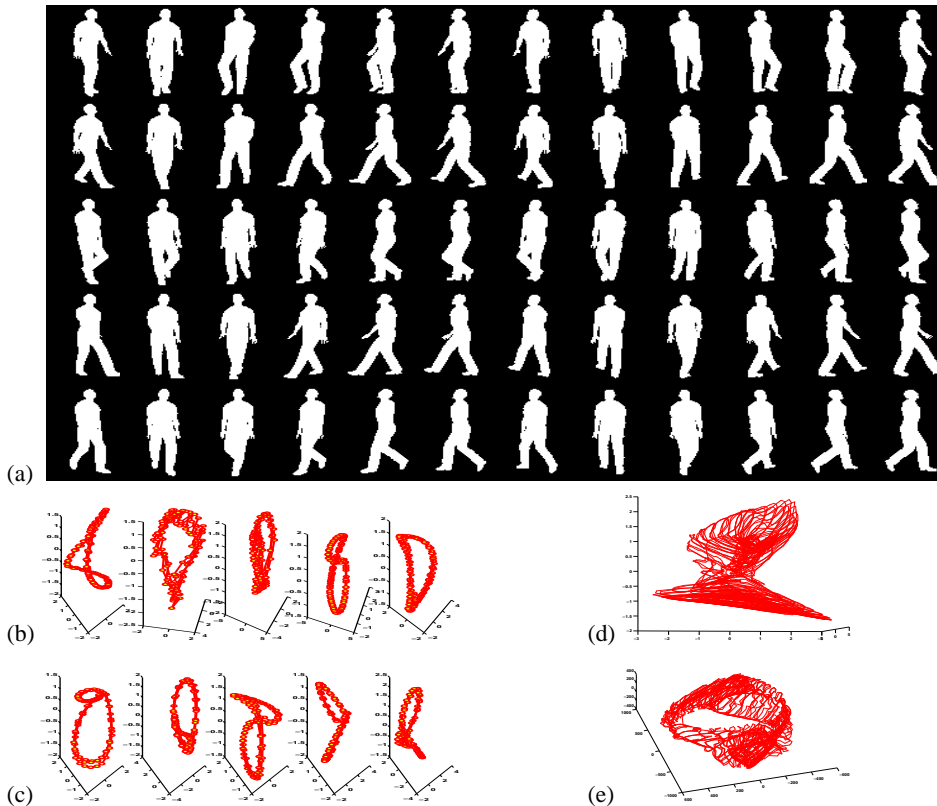


Figure 2: Data-driven view and body configuration manifolds:(a) Examples of sample data with view and configuration. Rows: body pose at  $0, \frac{1}{5}T, \frac{2}{5}T, \frac{3}{5}T, \frac{4}{5}T$ . Cols: view  $0, 30, 60, \dots, 330$ . (b) Intrinsic configuration manifold when view angle is  $0, 60, 120, 180$ , and  $240$ . (c) View manifold for five different fixed body pose. (d) (e) Combined view and body configuration manifold by LLE and Isomap.

when dealing with complex manifolds such as joint motion and view manifolds and will not necessarily lead to useful representations. This can be observed in Fig. 2-d,e where LLE and Isomap are used to embed data with continuous view and configuration variability as shown in Fig. 2-a. The resulting embedding, although reflects the actual manifold local structure, is not useful as a representation for tracking. Moreover, if we consider different people, the joint manifold is expected to twist differently depending on the shape of the person performing the motion. Therefore, the resulting representation will not be useful to generalize to other people. The conclusion is the data-driven embedding of the joint view-configuration manifold is not practical to be used in tracking, synthesis, or analysis tasks.

### Supervised Generative Manifold Learning:

Traditional manifold learning approaches are unsupervised where the goal is to find a low dimensional embedding of the data which preserve the manifold topological structure. However, if the manifold topology is known, manifold learning can be formulated in a different way. Manifold learning is then the task of learning the deformation of the manifold from an ideal case. For example, for the gait case, observed from the same view point, as shown in the examples in Fig. 2-b, the gait manifold is a one dimensional closed manifold which is topologically equivalent to a unit circle. So, we can think of the gait manifold as a twisted or deformed circle in the visual input space. Since we already know the topology, the task of manifold learning can be viewed as: how to deform a unit circle to reach the actual data manifold. Or, in other words, how to generate the data knowing an equivalent “idealistic” topological structure. In fact, this view can be even extended if the data manifold does not share the exact topology from the ideal manifold. For

example, the gait manifold can intersect itself in the visual space but still, we can learn the deformation from a unit circle to the data. Similarly, if we consider the view manifold for a certain body posture, the resulting manifolds are topologically equivalent to unit circle as can be seen in Fig. 2-c.

For the case of joint configuration and view manifold where the view varies along a view circle, this is a product space and ideally is equivalent to the produce of two circles, i.e., torus manifold. i.e., the data in Fig. 2-a lies on a deformed torus in the input space. So we need to learn deformation from the torus to the data. If we consider the full view sphere, the resulting manifold is a deformed order-3 torus or  $S^1 \times S^1 \times S^1$  structure.

On the other hand, the kinematic manifold, which is invariant to view point, is also a deformed circle in the kinematic space. Starting from a torus, the kinematic manifold can be achieved through collapsing the torus along one of its axis to form a circle and then deform that circle. Therefore, a torus manifold acts as an “ideal” manifold to represent both the latent body configuration and view variables,  $b_t, v_t$ . In one side, the torus can deform to form the visual manifold,  $y_t$ , and on the other side, it can deform to form the kinematic manifold  $z_t$ .

### 3 View and Configuration Joint Representation

#### 3.1 Torus Manifold

A torus manifold, a two dimensional manifold embedded in three dimensional space with a single hole, is useful to represent both periodic and non-periodic dynamic human motion observed from a viewing circle.

The torus manifold can be constructed from a rectangle, which can be represented by two orthogonal coordinates with range  $[0, 1] \times [0, 1]$ , by gluing both pairs of opposite edges together with no twists [6]. Therefore, the torus surface can be parameterized by two variables  $u, v \in [0, 1]$ .

As justified in Sec. 2 the torus can be used as a conceptual embedding for the joint view (along one viewing circle) and configuration manifold. The view and body configuration manifold can be parameterized in the rectangle coordinate with the two orthogonal axis of the torus manifold. Any manifold point in the torus can have two circles: one is in the plane of the torus, which we use to model the view variable and parameterized with  $\mu$ , and the other is perpendicular to it which we use to represent the body configuration and parameterized by  $\nu$ .

Generalization to the full view sphere around the person is straight forward. In this case the joint configuration and view manifold can be mapped to a family of tori, which is a subset of the product space of three circles  $S^1 \times S^1 \times S^1$ , one for the body configuration, one for the horizontal view circle and one for the vertical view circle. In practice, only small range of the vertical view circle is considered, therefore, this can be modeled as a set of rectangles each representing a torus manifold for a given view circle, i.e., can be parameterized by three parameters  $\mu, \nu, \xi$  for body configuration, view angle and elevation view angle.

#### 3.2 How to embed points on the torus

Given a sequence of kinematic data  $Z$  representing a motion, we can use graphics software to render body silhouettes from different view points along a given viewing circle. We denote this data by  $Y$ . It is desired to embed this data on the torus in a conceptual way that does not necessarily reflect their Euclidean distance in the kinematic space nor in the visual input space, instead the objective is to embed them on the torus in a way to simplify the tracking. There are two ways we can achieve such embedding.

**Constant Speed Dynamics:** For tracking, we not only know the topology of the manifold but we may also know the desired dynamics in the state space. For example, for periodic motion such as walking and running, although the nonlinearity in dynamics in both the kinematic and the visual input manifolds, we need the latent state variable to reflect a constant speed on the latent manifold. The nonlinear mapping in Eq. 1 should transform this linear dynamics to nonlinear dynamics. This can be achieved by embedding the points on equidistance points along the configuration axis of the torus.

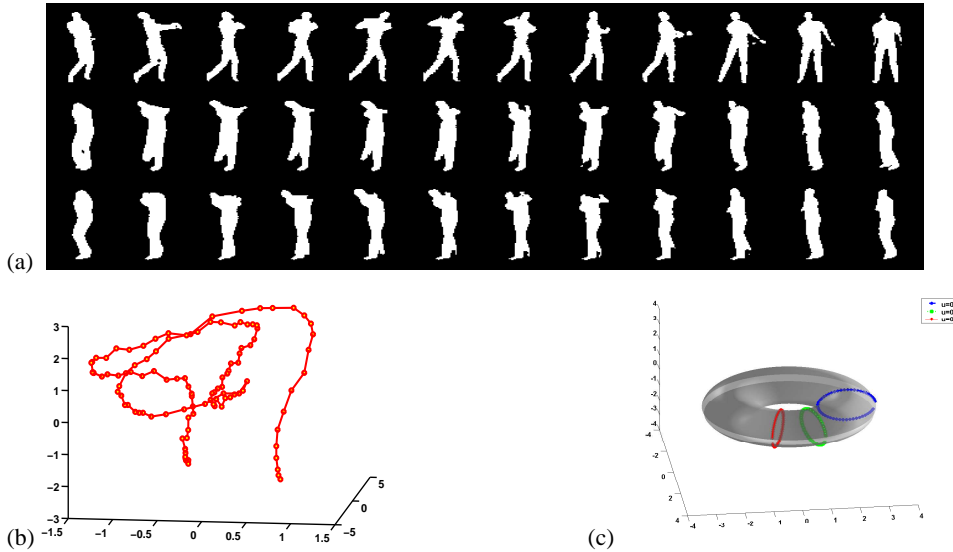


Figure 3: Torus manifold with gap. (a) Example sequence of a golf swing from three different views  $\mu = 0, 0.2, 0.3$ . (b) Embedding of golf swing motion capture data. (c) Visualization of a torus manifold with gap with trajectories of the three different views used for synthesis in (a)

**Geodesics-based Embedding:** For non-periodic motion, such as golf swing, where data might exhibit different acceleration along the course of the motion, it is desired to embed the data on the torus in a way that preserves their kinematic manifold structure. This can be achieved through embedding the points such that the geodesic distance on the torus is proportional to the geodesics on the kinematic manifold. Another constraint stems from the fact that in non-periodic motion, the manifold is an open trajectory and therefore, configuration manifold should be mapped to a part of the torus configuration axis.

To achieve this, we first embed the kinematic manifold using LLE or any other nonlinear embedding techniques. This leads to an open trajectory embedding. Such embedding is used for 1) measuring the gap between the beginning and end pose of the motion in order to map the manifold to a proportional part of the torus. 2) to measure the geodesics along the kinematic manifold. The points are embedded on the torus in such a way that only a part of the torus  $\nu$  axis is used proportional to the embedded manifold length. Let  $x_i, i = 0, \dots, N$  be the embedding coordinate of the kinematic sequence  $z_i, i = 0, \dots, N$ . The coordinate of point  $z_i$  on the torus  $\nu$ -axis is denoted by  $\nu_{z_i}$  and is set to be  $\nu_{z_i} = S_i/S$  where  $S_i$  is the geodesic distance of point  $x_i$  to the starting point,  $x_o$ , i.e.,  $S_i = \sum_{j=1}^N \|x_j - x_{j-1}\|$  and  $S$  is defined to be  $S = S_N + \|x_N - x_o\|$ . The gap between the beginning body pose embedding point and final body pose embedding points on the torus will be  $Gap = \frac{\|x_N - x_o\|}{S}$ . Fig. 3 (a) shows an example a golf swing motion from three different view points and its low dimensional embedding of the kinematics using LLE is shown in (b). Fig. 3 (c) shows a torus manifold with a gap between the start and the end body pose embedding for the case of a golf swing.

## 4 Learning Manifold Deformation

### 4.1 Learning Manifold Deformation

Learning a mapping from a topological structure to the data, where that topological structure is homeomorphic to the data, can be achieved through learning a regularized nonlinear warping function. Let  $\mathcal{T}$  denotes the torus manifold and  $\mathcal{M}$  denotes a data manifold where  $\mathcal{T}$  and  $\mathcal{M}$  share the same topology. Given a set of point  $x_i \in R^d, i = 1 \dots, K$  on  $\mathcal{T}$  and their corresponding points  $y_i \in R^D, i = 1 \dots, K$  on a manifold  $\mathcal{M}$ , we can learn a nonlinear mapping function  $g : R^d \rightarrow R^D$  from  $\mathcal{T}$  to  $\mathcal{M}$ . According to the representer

theorem [7], such function admits a representation in the form of  $y = \sum_j b_j k(x, z_j)$  where  $z_j$  are a finite set of points in the input space, not necessarily data points, and  $k(\cdot, \cdot)$  is a kernel function. If radial basis kernels are used then this is a form of radial basis function interpolation. Given the data embedded on the torus as described above, we can learn the deformation between the torus and both the visual manifold and the kinematic manifold. This can be achieved through learning two regularized nonlinear mapping functions in the form of Eq. 1 as follows:

**Torus to Visual Manifold:** Deforming the torus to the visual manifold can be achieved through learning a nonlinear mapping in the form of Eq. 1. Given the embedding coordinates on the torus,  $(\mu_v, \nu_b)$  and their corresponding visual data (silhouettes),  $y^{vb} \in R^D$ , for discrete pose  $b$  and view  $v$ , we can fit the mapping function  $g : R^2 \rightarrow R^D$  which map from the torus to the shape space in the form

$$y = g(\mu, \nu) = \mathbf{D} \cdot \psi(\mu, \nu) \quad (1)$$

satisfying  $y^{vb} = g(\mu_v, \nu_b)$ . In this case, we need a set  $N$  of basis functions covering the torus surface which are set uniformly across the surface. Using this model, for any view  $v$  and body configuration  $b$  sequence, we can generate a new observations where  $\mu_v$  is view representation in  $\mu$  axis, and  $\nu_b$  is body configuration representation in  $\nu$  axis of the torus manifold.

**Torus to Kinematic Manifold:**

Deforming the torus to the kinematic manifold can be achieved through learning a nonlinear mapping from the torus configuration axis to the kinematic manifold. Given the embedding on the torus,  $(\mu_v, \nu_b)$  and their corresponding kinematic points  $z_b \in R^d$  we fit the mapping function  $f : R \rightarrow R^d$  in the form

$$z = f(\nu) = \mathbf{B} \cdot \psi(\nu) \quad (2)$$

stratifying that  $z^b = f(\nu_b)$ . Given this mapping, any point on the torus  $(\mu, \nu)$  can be directly mapped to a 3D joint position configuration.

## 4.2 Learning Different People Manifolds from Partial Views

Our goal is to be able to achieve adaptive tracking where the tracker can adapt to the person contour shape. Elgammal and Lee [4] presented an approach for decomposing “style” variations in the space of nonlinear mapping coefficients from an embedded manifold to the observation space. Similar approach can be used here to learn style dependent mappings in the form of Eq. 1 from the torus to each person’s data. The torus represents a unified manifold representation invariant to the person.

Given different people sequences from different sparse view points, each sequence can be embedded on the torus as described in Sec. 3. Let  $Y_{v_k}^s$  be sequences of visual data for person  $s$  from view points  $v_k$ , we can embed such sequences on the torus which leads to a set of torus coordinates  $(\mu_{v_k}, \nu_b)$ . The view points do not need to be the same across subjects and the sequences do not need to be the same length; only the beginning and end of the motion is needed to be aligned on the torus  $\nu$ -axis. Given the embedding points and their corresponding contours, person-specific mapping functions in the form of Eq. 1 can be fitted which leads to an  $D \times N$  coefficient matrix  $D^s$ . Notice that the kernel space defined by  $\psi(\cdot)$  in Eq. 1 is the same across all subjects since the same RBF basis are used on the torus. Given the coefficient matrices, we can fit a model in the form of

$$y_{vb}^s = \mathcal{A} \times_1 a^s \times_2 \psi(\mu_v, \nu_b) \quad (3)$$

where  $a^s$  is a vector characterizing the person shape style and  $\mathcal{A}$  is third order tensor with dimensions  $D \times S \times N$  where  $S$  is the dimensionality of the shape space and  $\times_n$  is the tensor multiplication as defined in [8]. The model in Eq. 3 generalized over the model proposed by Elgammal and Lee [4] which is limited to single view and it’s extension to multiple views required learning several view-based representations. The model proposed here, provides a continuous representation of the view point and the body configuration in one latent representation space.

## 5 Bayesian Tracking on the Torus

Given observations which represent body contours or detected image edges, we need to estimate the state on the torus, i.e., the configuration and the view, as well as the shape style parameter. Recovering the state on the torus directly yield the body kinematics in 3D through the mapping function 2. The Bayesian tracking framework enables a recursive update of the posterior  $P(\mathbf{X}_t|\mathbf{Y}^t)$  over the object state  $\mathbf{X}_t$  given all observations  $\mathbf{Y}^t = \mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_t$  up to time  $t$ . The state can be updated based on observation likelihood estimation  $P(\mathbf{Y}_t|\mathbf{X}_t)$  with transition probability  $P(\mathbf{X}_t|\mathbf{X}_{t-1})$  and previous state estimation.

On the torus manifold, view and body configuration are represented together. The manifold provides natural continuous representation of joint distribution for particle filtering. The state has three components: view configuration, body configuration on the manifold and shape parameter. We denote the state at time  $t$  by  $\xi_t = [\lambda_t, \mu_t, \nu_t]$  where  $\mu_t$  and  $\nu_t$  are the torus coordinate corresponding to view and body configurations.  $\lambda_t$  is the shape state where the shape at time  $t$  is assumed to be a convex combination of shape classes in the training set. That is, the shape style vector  $a$  in Eq. 3 is written as a linear combination of  $K$  shape style vectors  $a^k$  in the style space  $a_t = \sum_{k=1}^K w_t^k a^k$ ,  $\sum_{k=1}^K w_t^k = 1$ . The shape state is represented by the coefficients  $w_t^k$ , i.e.,  $\lambda_t = [w_t^1, \dots, w_t^K]$ . Using traditional particle filter, We represent view, configuration, and shape style joint state with  $N_\xi$  particles  $\{\xi_t^{(k)}, \pi_t^{(k)}\}_{k=1}^{N_\xi}$  with weights  $\pi$ .

**Dynamic Model:** Since the pose, the view, and the shape style parts of the state are independent given the observation, the dynamic model is a product of three dynamic models, i.e.

$$P(\mathbf{X}_t|\mathbf{X}_{t-1}) = P(\lambda_t|\lambda_{t-1})P(\mu_t|\mu_{t-1})P(\nu_t|\nu_{t-1})$$

. The shape state is suppose to be time-invariant, but in tracking, since the subject shape style is not known, the shape style needs to change with frames till it adapts to the correct shape style. Therefore, the propagation of particles in the shape space is controlled by a variance variable that decay with time. As introduced in Sec. 3, the configuration can be embedded on the torus in a way to achieve constant speed dynamics in the state space which is very suitable in the case of periodic motion such as gait. Therefore, the propagation of particle in the body configuration domain can use such constant speed model which can adapt to each person dynamics through tracking. Alternatively, the particles can just be propagated through a random walk process on the torus.

**Observation Model:** The generative model in Eq. 3 fits directly to the Bayesian framework to generate observation hypotheses from states. We can generate samples for given particle  $\xi_{t+1}^{(k)} = [\lambda_{t+1}^{(k)}, \mu_{t+1}^{(k)}, \nu_{t+1}^{(k)}]$  by

$$\mathbf{y}_{t+1}^{(k)} = \mathcal{A} \times \left[ \sum_{l=1}^K w^l a^l \right] \cdot \psi(\mu_{t+1}^{(k)}, \nu_{t+1}^{(k)})$$

, i.e., each particle will directly generate a hypothesis contour in a given configuration, in give view, in the form of a level set function. The observation itself, can be in form of extracted silhouettes (using background subtraction) or just edges extracted directly from the images.

## 6 Experimental Results

We evaluate our algorithm quantitatively using Brown HUMANEVA-I dataset [16]. We also show adaptation of shape model with online adaptation and with shape style estimation given a collection of training shape styles in advance. We also test view and body configuration estimation from real and synthetic silhouettes with view variations.

**Brown HUMANEVA-I dataset Evaluation:** We measured 3D reconstruction error using Brown HUMANEVA-I dataset [16]. We generated synthetic training data of walking silhouette from motion capture data using animation software Poser®. 12 different views ( $10^\circ, 30^\circ, \dots, 360^\circ$ ) are collected for walking on a circle motion. We extracted silhouettes using background subtraction. Joint locations of the validation set and of

one cycle of training sequence are extracted and normalized to represent *normalized pose* which is invariant to subject’s rotation and translation.

We estimate body pose from the maximum a posterior (MAP) estimation of body configuration from the particle filtering. We used 900 particles ( $N_{\beta} = 900$ ) in the experiment to represent view and body configuration on the torus manifold. We measured errors in estimated body pose by average absolute distance between individual markers as in [16] Fig. 4-a shows average error in each frame.

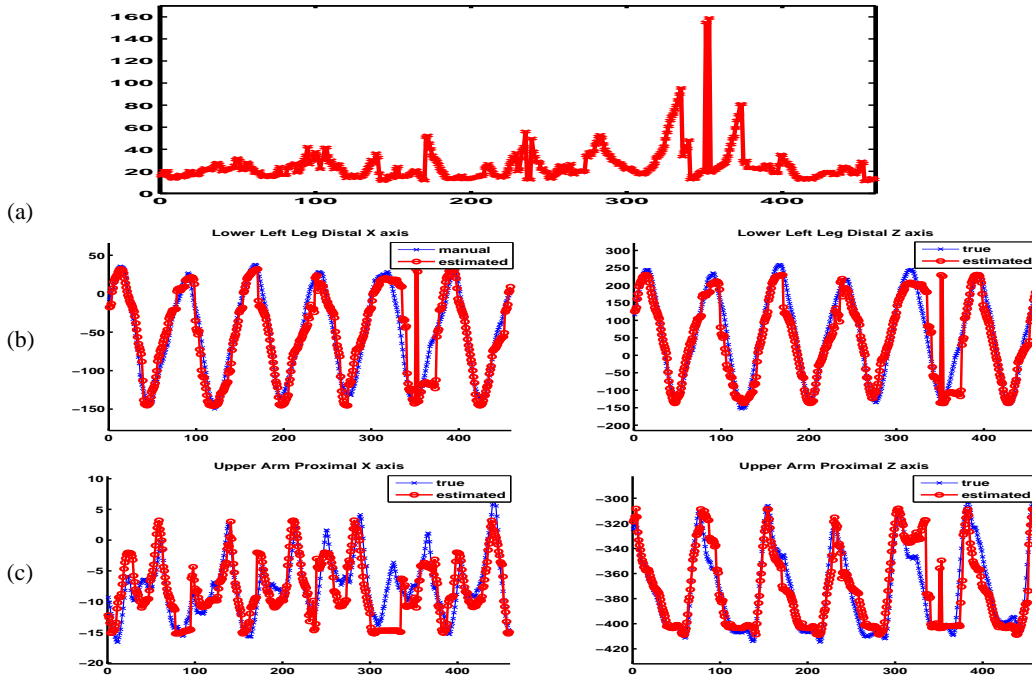


Figure 4: Error measurement of 3D body pose reconstruction for HUMANEVA-I: X-axis: frame number, Y-axis: joint location value (unit:*mm*). (a) Average errors in joint locations in each frame. (b)(c) True and estimated joint location *x* and *z* values for *Lower left leg distal* and *Upper right arm proximal*.

**Comparison to other Representations:** We compared the torus representation with other embedding approaches for the task of body pose estimation. Since we used a torus as a two-dimensional manifold embedding for view and body configuration representation, we also used two-dimensional embedded representation obtained from LLE [14] and Isomap [19]. We used the same number of particles in the two-dimensional embedding space for all approaches. We also compared nearest neighborhood search to see the best result we can get from the collected data itself. Table 1 shows average error for the different approaches. For the case of nearest neighbor ( NN ), we searched for the nearest silhouette from training sequence and used its corresponding 3D joint location as reconstruction. Torus embedding shows much better performance than other manifold representation.

Table 1: Average error in normalized 3D body pose estimation in different embedding

Embedding Type	LLE	Isomap	NN	Torus
Average Error in <i>mm</i>	62.19	61.08	49.52	24.08

**Shape Style Adaptation:**

As we can model shape variations in different people as style change, we can adapt to observed shape

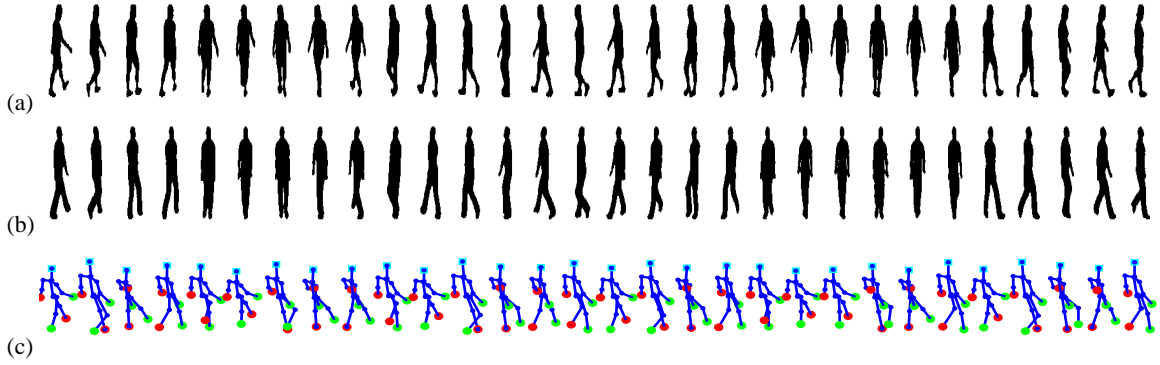


Figure 5: Test sequence and reconstruction in selected HUMANEVA-I data (S1-walking) : (a) Input silhouettes. (b) Reconstructed silhouette based on estimated view and body configuration. (c) Reconstructed 3D body pose.

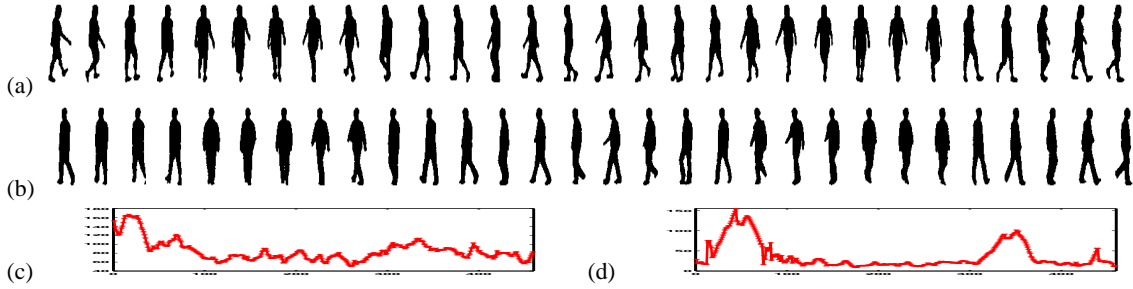


Figure 6: Style adaptive circular gait sequence tracking: (a) Original test silhouettes. (b) Estimated silhouettes with style adaptation. (c) Measured shape contour error in each frame in style adaptation. (d) Measured 3D reconstruction error (Average errors in joint locations in each frame in different embedding).

by estimating the style factor  $a^s$  in Eq. 3 to explain observed shape. New person’s style can be represented by combination of training person style. In our experiment, we captured people walking sequence on the treadmill with multiple camera. For our experiment, we collected sequences from 4 different people with 7 different views using synchronized camera. We started from mean style. As style adaptation goes on, the 3D reconstruction errors and 2D image reconstruction errors are decreased. Fig. 6 shows experiment results for the HUMANEVA-I dataset we used in the previous experiment. At the beginning, shape contour error (c) are large but it decrease as the style estimation get more accurate parameters. Similarly, the estimated 3D body pose shows decrease in error as time passes after large errors at the beginning when we used just mean style.

**Jump sequences:** We evaluated the approach with a jump motion (example of open manifold) where the subject can rotate in the air while jumping. We used motion captured data to learn the model using geodesics-based embedding on the torus. Fig. 7 shows estimation of view and body configuration in outdoor environment. Despite inaccurate silhouette extraction (Fig. 7-a), our model estimate body configuration accurately (Fig. 7-e).

Fig. 8-a shows jump motion with body rotation in the air. Estimated view parameter shows constant view parameter change due to body rotation in Fig. 8-d. Simultaneously the estimated configuration parameter enables reconstruction of 3D body pose (Fig. 8-f).

**Edge-based Contour Tracking:** We tested the approach with real data without background subtracted contours. Instead Chamfer matching is used as an observation model given edges extracted from the images.

We tested for a walking sequence along a circle with fixed camera view. Fig. 9-a,b shows tracking results for walking sequence with view variation. You can see spiral motion on the torus manifold due to simultaneous change of view and body configuration. The tracking on the torus manifold can achieve reliable tracking result with prior dynamic constraints on the manifold even weak edge cues.

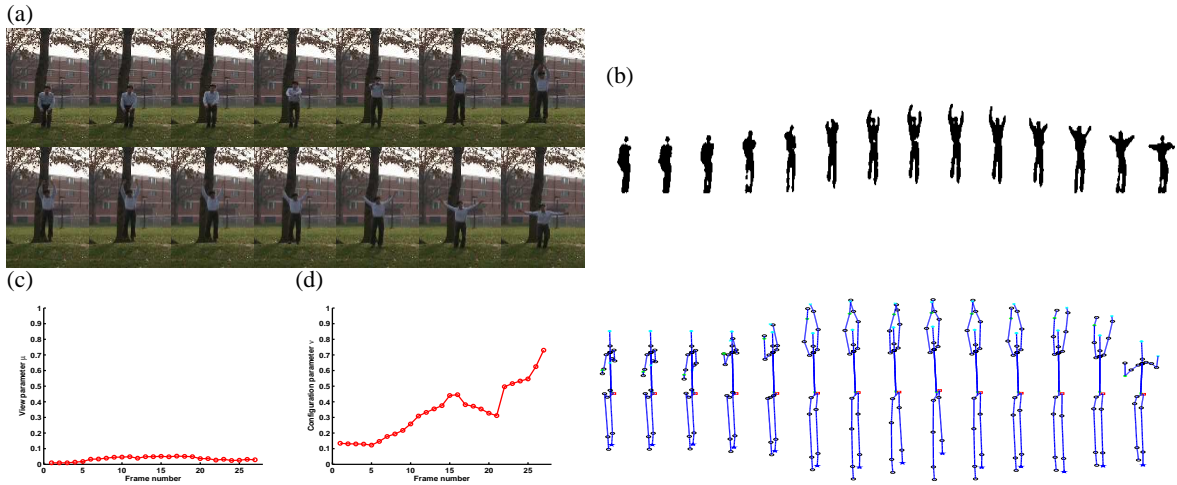


Figure 7: Outdoor fixed view jump motion. (a) Input image. (b) Input silhouette. (c) Estimated view. (d) Estimated body configuration. (e) 3D body pose reconstruction based on estimated body configuration.

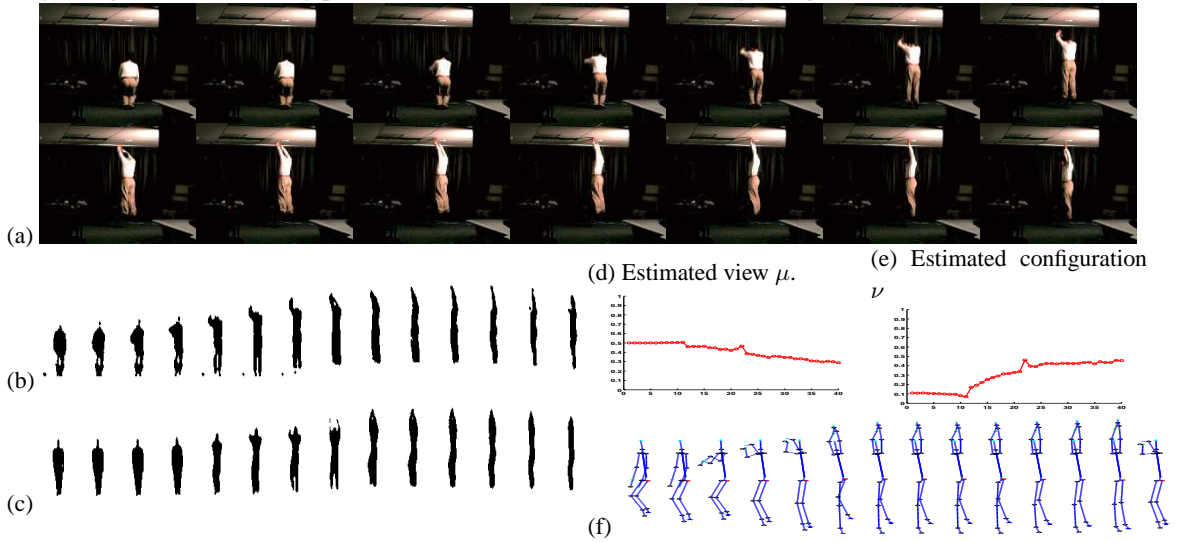


Figure 8: Indoor jump motion with rotation. (a) Input image. (b) Input silhouettes. (c) Reconstructed silhouettes. (d) Estimated view. (e) Estimated body configuration. (f) 3D body pose reconstruction based on estimated body configuration.

**Golf Swing Tracking:** In this experiment we tested tracking performance of golf swing from unknown camera and view. In this experiment, we can recover correct view and body configuration. Fig. 9-c,d shows tracking results. Since the view is unknown, we start from a uniform distribution, i.e., the particles are spread along the big circle on the torus (the same  $\mu$ ) at the beginning and it converged to one area.

## 7 Conclusions

We formulated view variant human motion tracking as tracking on a torus surface. We use the torus as a state space for both body configuration and view. We learn how the torus deform to the actual visual manifold and to the kinematic manifold through two nonlinear mapping functions. The torus model is suitable for one dimensional manifold motions, whether periodic, as walking, running, etc., or non periodic, as golf swings, jumping, etc. The experimental results showed that such model is superior than other representations for the

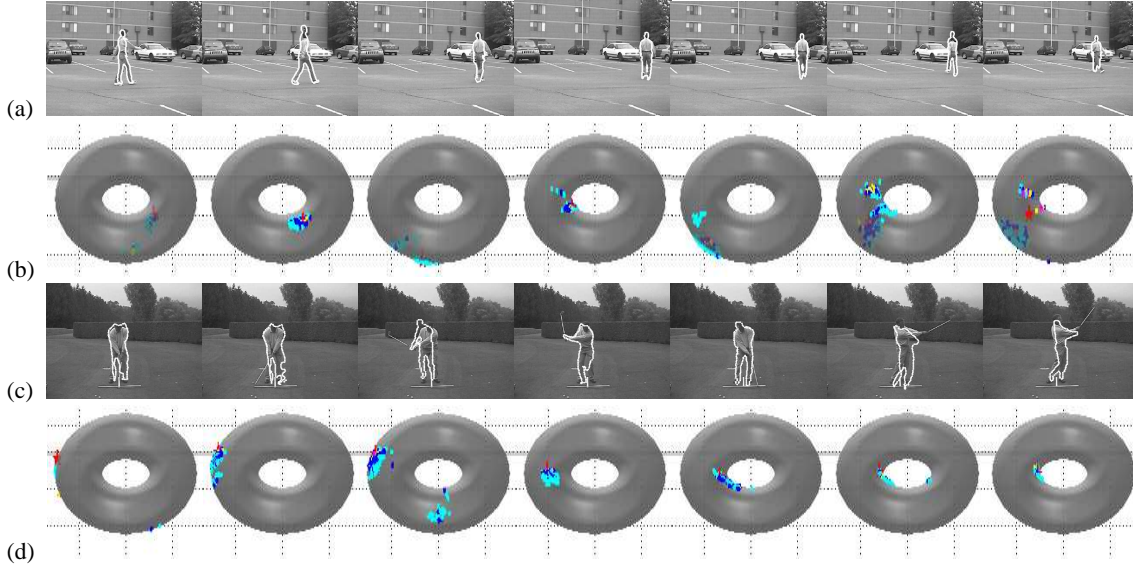


Figure 9: Circular gait sequence tracking: (a) Estimated shape contours. (b) View and configuration particle distributions on torus manifold. Golf swing tracking: (c) Estimated shape contours (d) View and configuration particle distributions on torus manifold.

task of tracking and pose/view recovery since it provides a low dimensional, continuous, uniformly spaced state representation. We also show, how the model can be generalized to the full view sphere and how to adapt to different people shapes.

## References

- [1] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevance vector regression. In *CVPR*, volume 2, pages 882–888, 2004.
- [2] M. Brand. Shadow puppetry. In *Proc. of ICCV*, volume 2, pages 1237–1244, 1999.
- [3] A. Elgammal and C.-S. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *Proc. CVPR*, volume 2, pages 681–688, 2004.
- [4] A. Elgammal and C.-S. Lee. Separating style and content on a nonlinear manifold. In *Proc. CVPR*, volume 1, pages 478–485, 2004.
- [5] K. Grauman, G. Shakhnarovich, and T. Darrell. Inferring 3d structure with a statistical image-based shape model. In *ICCV*, pages 641–648, 2003.
- [6] A. Gray. *Modern Differential Geometry of Curves and Surfaces with Mathematica*. CRC Press, 2nd edition, 1997.
- [7] G. S. Kimeldorf and G. Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41:495–502, 1970.
- [8] L. D. Lathauwer, B. de Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM Journal On Matrix Analysis and Applications*, 21(4):1253–1278, 2000.
- [9] V. I. Morariu and O. I. Camps. Modeling correspondences for multi-camera tracking using nonlinear manifold learning and target dynamics. In *CVPR (1)*, pages 545–552, 2006.
- [10] A. Rahimi, B. Recht, and T. Darrell. Learning appearance manifolds from video. In *CVPR*, 2005.
- [11] J. M. Rehg and T. Kanade. Visual tracking of high DOF articulated structures: an application to human hand tracking. In *ECCV (2)*, pages 35–46, 1994.
- [12] J. M. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *ICCV*, pages 612–617, 1995.
- [13] R. Rosales, V. Athitsos, and S. Sclaroff. 3d hand pose reconstruction using specialized mappings. In *Proc. ICCV*, pages 378–387, 2001.

- [14] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.
- [15] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *ECCV(2)*, pages 702–718, 2000.
- [16] L. Sigal and M. J. Black. Humaneva: Cynchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, 2006.
- [17] C. Sminchisescu and A. Jepson. Generative modeling of continuous non-linearly embedded visual inference. In *ICML*, pages 140–147, 2004.
- [18] C. Sminchisescu, A. Kanaujia, Z. Li, and D. N. Metaxas. Discriminative density propagation for 3d human motion estimation. In *CVPR (1)*, pages 390–397, 2005.
- [19] J. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319 – 2323, 2000.
- [20] R. Urtasun, D. J. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In *ICCV*, pages 403–410, 2005.