Modeling View and Posture Manifolds for Tracking

Chan-Su Lee and Ahmed Elgammal
Department of Computer Science
Rutgers University
Piscataway, NJ, 08854, USA
{chansu,elgammal}@cs.rutgers.edu

Abstract

In this paper we consider modeling data lying on multiple continuous manifolds. In particular, we model the shape manifold of a person performing a motion observed from different view points along a view circle at fixed camera height. We introduce a model that ties together the body configuration (kinematics) manifold and the visual manifold (observations) in a way that facilitates tracking the 3D configuration with continuous relative view variability. The model exploits the low dimensionality nature of both the body configuration manifold and the view manifold where each of them are represented separately.

1. Introduction

Despite the high dimensionality of the human body configuration space, many human activities lie intrinsically on low dimensional manifolds. Exploiting such property is essential to constrain the solution space for many problems such as tracking, posture estimation, and activity recognition. Recently, there have been increasing interests in learning low dimensional representations of the body configuration manifolds, as in [3, 15, 18, 2, 8], for tracking and posture estimation.

The goal of this paper is to model the visual manifold of an articulated object observed from different view points. Modeling visual manifolds is a challenging task. In particular, we focus on modeling human motion as observed from different view points. Traditionally, generative model-based approaches have been used for tracking and posture estimation, where a 3D body model and a camera model are used, and the problem is formulated as a search problem in high dimensional spaces (articulated body configuration and geometric transformation), e.g. [13]. Alternatively, discriminative mappings have also been introduced, e.g. [10, 1]. The model introduced here is generative. However, it generates observations for a certain motion as observed from different view points without any explicit 3D body model, rather, this is achieved through modeling the visual manifold corresponding to different postures and views.

Modeling the visual manifolds for rigid objects under different views and illuminations have been studied in [9] for object recognition. However, dealing with articulated objects are more challenging. Consider a simple example of a human performing a periodic motion, like walking, and observed from different view points along a view circle. It was shown in [3] that, from a given view point, the observed motion lies on a low dimensional manifold (one dimensional for gait). This corresponds to the configuration manifold observed from a single view point. On the other hand, given a single body posture observed from different view points along a viewing circle, the observations will lie on a one-dimensional manifold as well. That is the view manifold for that particular posture. i.e., each posture has its own view manifold and each view has its own configuration manifold. If both the motion and the view are one-dimensional manifolds (e.g., gait observed from a view circle), then this product space was shown to be equivalent to a torus manifold [7]. In that work, a torus was used to model such two dimensional manifold (configuration $\times$ view) jointly. However, the approach in [7] is limited to the particular sitting of a one-dimensional motion. The fundamental question we address here is: *How to learn a representation of the view manifold that is invariant to the body posture and, therefore, exhibits the one-dimensional behavior expected due to the camera setting.*

Contributions: Our work here aims to:

I- To model the posture, the view, and the shape manifolds of an observed motion with three separate low dimensional representations: 1) a view-invariant, shape-invariant configuration manifold; 2) a configuration-invariant, shape-invariant view manifold; 3) a configuration-invariant, view-invariant shape representation.

II- To model view and posture manifolds in a general setting where the motion is not assumed to be one dimensional. We show results with complex motions.

III- To link the configuration manifold, learned from 3D

motion-captured date, with the visual manifold. A distinguishing feature about our work here is that we utilize both the input (visual) and output (kinematics) manifolds to constrain the problem. That is, we model the kinematic manifold and the observation manifold, tied together with a parameterized generative mapping function.

We consider tracking and inferring view and body configuration of human motion from a single monocular camera where the person can change his/her pose with respect to the camera while being tracked (or equivalently the camera can be moving). In this paper we limit the view variability to a one view circle. However, this is not a theoretical limitations of the approach, but rather, a practical choice. Our goal mainly is to model the person's pose w.r.t. the camera and not the camera motion. In many applications, typically, the camera is fixed and mounted at a fixed hight and the person change his/her orientation w.r.t. the camera. A one view circle is a good approximation of the expected view point variabilities in such scenarios as will be shown from the experimental results where no camera calibration is assumed.

The paper organization is as follows: Section 2 summarizes the framework. Sections 3 and 4 describe the learning procedure. Section 6 shows some experimental results on different motions with varying complexity.

2. Framework

We consider two manifolds: 1) the body configuration manifold during the motion in the kinematic space 2) the visual input manifold (the observations) of the same motion observed from different view points along a view circle at a fixed camera height. It is clear that the kinematic manifold can be embedded using linear of nonlinear dimensionality reduction techniques to achieve a low dimension representation of the manifold, which can be used for tracking. For example, Gaussian Process Dynamic Models (GPDM) [20] can achieve such embedding, as well as learn a dynamic model for such manifold. The challenge is the visual manifold, since it involves both body configuration and view variabilities. Embedding such complex manifold will not result in any useful representation that can be used for inferring the configuration and view separately. That can be noticed in Fig. 1-c where LLE [11] is used to embed the visual manifold of a ballet motion from different views.

Here we summarize our approach:

1) Using joint angles data, we obtain an embedding of the kinematics, representing the motion manifold invariant to the view. We learn a parameterization of the motion manifold in the embedding space and learn the dynamics through learning a flow field.

2) Given view-based observations, we learn view-based nonlinear mapping functions from the kinematic manifold embedding space to the observations in each of the views.

3) Given the view-based mapping functions coefficients, arranged as a tensor, we factorize the view factor using high order singular value decomposition (HOSVD) [5].

4) Given the view factors, we explicitly model the view manifold in the coefficient space, which leads to a representation of the view manifold invariant to body configuration.

5) We also factorize individuals' shape variabilities within the same model.

The result is two low-dimensional embeddings: one for configuration and one for the view, as well as a generative model that can generate observation given the two manifolds' parameterizations. This fits perfectly into the Bayesian tracking framework as it provides, in a direct way: 1) a low dimensional state representation for each of the view and the configuration, 2) a constrained dynamic model since the manifolds are modeled explicitly, 3) an observation model, which comes directly from the generative model used.

3. Learning Configuration and View Manifolds

3.1. Learning View-invariant Configuration Manifold

As a common representation of the body configuration, invariant to view point, we use an embedding of the kinematic manifold, which represents the body configuration in a low dimensional space. Such kinematic manifold embedding is also invariant to different people shapes and appearances. We can obtain a low dimensional representation of the kinematic manifold by applying nonlinear dimensionality reduction for motion capture data using approaches such as LLE [11], Isomap [16], GPLVM [6], etc[1]. Since we need to achieve an embedding of the kinematics, invariant to the person's transformation with respect to the world coordinate system, we represent the kinematics using joints' location in a human-centered coordinate system. We aligned for global transformation in advance in order to only count for motion due to body configuration changes.

Fig. 2-a shows an embedded kinematic manifold for a gait motion. As expected, for a periodic motion as in the gait case, the embedding shows the kinematic manifold as a one-dimensional twisted closed manifold, which can be embedded free of intersections in a three dimensional embedding space. For more complex motions, the manifold is not necessarily one-dimensional. However, we can always achieve an embedding of the kinematic manifold in a low dimensional Euclidean space. Fig. 3-a shows an example embedding for the ballet dance routine data, which is shown in Fig. 1.

3.2. Learning Posture-invariant View Manifold

Given an embedding of the kinematic manifold, we can achieve a representation of different views by analyzing the

[1] In particular, we used LLE in this paper. The choice of which embedding technique to be used is not relevant to the approach.

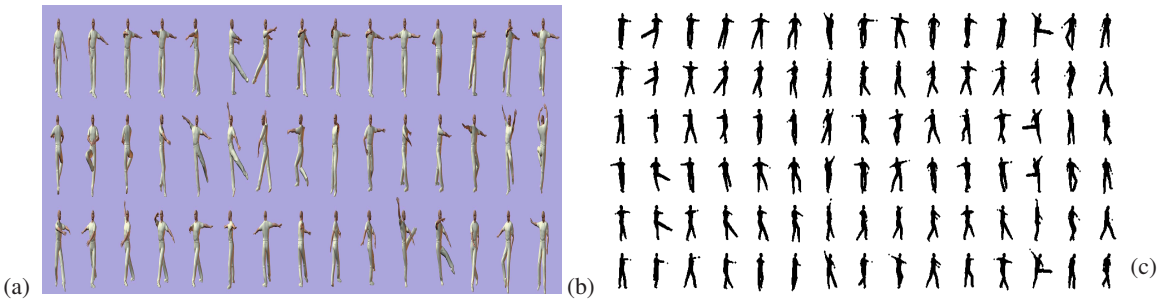

Figure 1. Example of a complex motion from different views: (a) Example postures from a ballet motion. We selected $8^{th}, 16^{th}, \cdots, 360^{th}$ frames from a sequence. (b) Sampled shapes in different views. Rows: different views ($30^o, 90^o, \cdots, 330^o$). Columns: body postures at frames $25^{th}, 50^{th}, \cdots, 375^{th}$ (c) Combined view and body configuration manifold embedding by LLE.

coefficient space of nonlinear mappings between the kinematic manifold embedding and view-dependent observation sequences. Elgammal and Lee [4] introduced a framework to separate "style" factors in the space of the coefficients of nonlinear functions that map from a unified "content" manifold and style-dependent observations. In our case, we consider the kinematic manifold embedding as the"content" manifold and the view is considered as a "style" factor, where, such "style" variations are factorized in the space of nonlinear mapping coefficients from an embedded manifold to the view-dependent observations. However, unlike [4], the view (style factor) in our case lies on a continuous manifold. Also, unlike [4] where the content manifolds were view-dependent, in our case, the use of the kinematic manifold provides a view invariant content representation and, therefore, differences between view-dependent observed data will be preserved in the nonlinear mapping of each view-dependent input sequences.

Given a set of N body configuration embedding coordinates on the kinematic manifold, $X = \{\boldsymbol{x}_1 \cdots \boldsymbol{x}_N\}$ and their corresponding view-dependent shape observations (silhouettes) $\boldsymbol{Y}^k = \{\boldsymbol{y}_1^k \cdots \boldsymbol{y}_N^k\}$ for each view k where $k = 1, \cdots, V$, we can fit view-dependent regularized nonlinear mapping functions in the form of generalized radial basis function

$$\boldsymbol{y}^k = \boldsymbol{B}^k \psi(\boldsymbol{x}), \tag{1}$$

for each view k. Here, each observation $\boldsymbol{y}$ is represented as D dimensional vector and we denote the embedding space dimensionality by e. $\psi(\cdot)$ is an empirical kernel map [12] $\psi_{N_c}(\boldsymbol{x}) : \mathbb{R}^e \rightarrow \mathbb{R}^{N_c}$ defined using N_c kernel functions centered around arbitrary points $\{\boldsymbol{z}_i \in \mathbb{R}^e, i = 1 \cdots N_c\}$ along the kinematic manifold embedding where

$$\psi_{N_c}(\boldsymbol{x}) = [\phi(\boldsymbol{x}, \boldsymbol{z}_1), \cdots, \phi(\boldsymbol{x}, \boldsymbol{z}_{N_c})]^{\mathsf{T}}, \tag{2}$$

where $\phi(\cdot, \cdot)$ is a radial basis function (we use Gaussian functions). Each $D \times N_c$ matrix $\boldsymbol{B}^k$ is a view-dependent coefficient matrix that encodes the view variability. Given such view-dependent mapping coefficients, we can fit a model in the form

$$\boldsymbol{y}_i^k = \boldsymbol{\mathcal{A}} \times_1 \boldsymbol{v}^k \times_2 \psi(\boldsymbol{x}_i), \tag{3}$$

where $\boldsymbol{\mathcal{A}}$ is a third-order tensor with dimensionality $D \times V \times N_c$ and $\times_j$ is the mode-j tensor multiplication [5]. This equation represents a generative model to synthesize observation

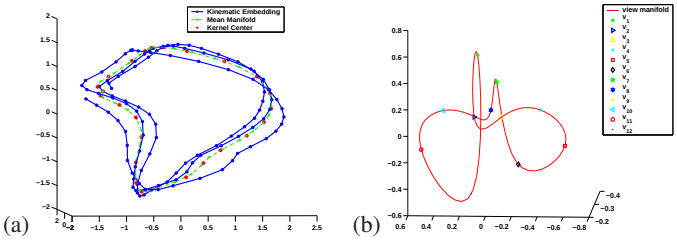

Figure 2. Configuration and View Manifolds for Gait:(a) Embedded kinematic manifold (b) Configuration-invariant view manifold (The first three dimensions are shown).

vector $\boldsymbol{y}_i^k \in \mathbb{R}^D$ of view k and configuration i given a view vector $\boldsymbol{v}^k$, and body configuration represented by embedding coordinate $\boldsymbol{x}_i \in \mathbb{R}^e$ on the kinematic manifold embedding.

To fit such model, the view-dependent coefficient matrices $\boldsymbol{B}^k, k = 1, \cdots, V$ are stacked as columns in a $(DN_c) \times V$ matrix $\boldsymbol{C}$ and then the view factors are decomposed by fitting an asymmetric bilinear model [17]. i.e., $\boldsymbol{C} = \boldsymbol{A} \cdot [\boldsymbol{v}^1 \cdots \boldsymbol{v}^V]$. The third-order $(D \times V \times N_c)$ tensor $\boldsymbol{\mathcal{A}}$ in Eq. 3 is the tensor representation of the matrix $\boldsymbol{A}$, which can be obtained by unstacking its columns.

The resulting representation of the view variations is discrete and high dimensional. The dimensionality of the view vector in Eq. 3 depends on the number of views, i.e., V dimensional. This high dimensional representation is not desirable as a state representation for tracking. The dimensionality can be reduced when fitting the asymmetric model by finding fewer number of view bases. Fig. 2-b and Fig. 3-b show the embedded posture-invariant view manifold in the mapping coefficient space for gait and ballet dance motion respectively, which clearly shows a one dimension manifold that preserves the proximity between nearby views. Here, the first three dimensions are shown. The actual view manifold can then be explicitly represented as will be shown in Sec. 4.

3.3. Learning Observation Shape Variability

The model in Eq. 3 can be further generalized to include a variable for shape style variability between different people, i.e., to model different people shapes. The use of the kinematic manifold provides an invariant representation to observation variabilities, which allows us to generalize the model. Given view-dependent shape observations for different people, we

can fit view-dependent, person-dependent mapping functions in the form of Eq. 1, which yields a set of coefficient matrices $\boldsymbol{B}^{kl}$ for each person l and view k. Given such coefficient matrices, we can fit a generalized model in the form

$$\boldsymbol{y}_i^{kl} = \boldsymbol{\mathcal{D}} \times_1 \boldsymbol{s}^l \times_2 \boldsymbol{v}^k \times_3 \psi(\boldsymbol{x}_i), \tag{4}$$

where $\boldsymbol{\mathcal{D}}$ is a forth-order tensor with dimensionality $D \times S \times V \times N_c$. This equation represents a generative model to synthesize an observation vector $\boldsymbol{y}_i^{kl} \in \mathbb{R}^D$ of a view k, a shape style l and a configuration i, given a view vector $\boldsymbol{v}^k \in \mathbb{R}^V$, a shape style vector $\boldsymbol{s}^l \in \mathbb{R}^S$, and a body configuration represented by an embedding coordinate $\boldsymbol{x}_i \in \mathbb{R}^e$ on the kinematic manifold embedding. Fitting such model can be achieved using HOSVD [5, 19]

4. Parameterizations of View and Configuration Manifolds

4.1. Parameterizing the View Manifold

Given the view space defined by the decomposition in Eq. 3, different view vectors are expected to lie on a low dimensional nonlinear manifold. Obviously, a linear combination of view vectors in Eq. 3 will not result in valid view vectors. We need to explicitly model the view manifold in the coefficient space to be able to predict and synthesize new views. Therefore, we model view variations as a one-dimensional nonlinear manifold by a one-dimensional continuous variable using spline fitting with C^2 connectivity constraints between the last and the first sample views, since the view manifold is presumed to be closed. As a result, we represent the view manifold by a one-dimensional view parameter θ where a certain view $\boldsymbol{v}_t$ can be represented as $\boldsymbol{v}_t = g_v(\theta_t)$. Fig. 2-b and 3-b show a spline-parameterized one-dimensional view manifold embedded in a three dimensional space.

4.2. Parameterizing the Configuration Manifold

In general, we make no assumption about the dimensionality of the body configuration manifold. However, we discriminate between two cases: 1) the case of a one-dimensional motion, whether periodic, such as walking, running, etc., or non periodic open trajectory, such as golf swings, tennis serves, etc. 2) the case of a general motion where the actual configuration manifold dimensionality is unknown, e.g., dance or aerobics, etc.

For one-dimensional motions, the kinematic manifold can be represented using a one-dimensional spline parameter $\beta_t \in \mathbb{R}$ and a spline function $g_b : \mathbb{R} \rightarrow \mathbb{R}^e$ that maps from the parameter space into the embedding space and satisfies $\boldsymbol{x}_t = g_b(\beta_t)$. Using the spline parameter is advantages over the embedding $\boldsymbol{x_t}$ since it leads to a constant-speed dynamic model since the parameter β_t will change in a constant-speed between frames while the embedding x_t will change in variable steps on the manifold. This can be seen in the results as in Fig. 4-e and Fig. 7-c.

For complex motions, such as aerobics, dance, etc., where the manifold dimensionality is unknown, a two-dimensional embedding space is used to represent the manifold. In such case, the kernel functions centers in Eq. 2 are fit to the embedded manifold through fitting a Gaussian mixture model. To learn the dynamics in such case, we learn a flow field in the embedding space.

Given a sequence of N body configuration embedding coordinates on the kinematic manifold, $X = \{x_1 \cdots x_N\}, x_t \in \mathbb{R}^2$ we can directly obtain flow vectors, representing the velocity in the embedding space, as $v(x_t) = x_t - x_{t-1}$. Given this set of flow vectors, we estimate a smooth flow field over the whole embedding domain where the flow $v(x)$ at any point x in the space can be estimated as $v(x) = \sum_{i=1}^{N} b_i k(x, x_i)$ using Gaussian kernels $k(\cdot,\cdot)$ and linear coefficients $b_i \in \mathbb{R}^2$, which can be obtained by solving a linear system. The smooth flow field is used to estimate how the body configuration will change in the embedding space, which is used in tracking to propagate the particles. Fig. 3-d shows an example of the motion flow field for a ballet dance motion.

4.3. Parameterizing the Shape Space

The shape variable $\boldsymbol{s}$ in Eq. 4 can be high dimensional. To constrain the shapes generated by the model in Eq. 4, we represent a shape as a linear convex combination of shape clusters in the training data. That is, the shape style vector $\boldsymbol{s}$ is written as a linear combination of Q shape style vectors $\boldsymbol{s}^q$ in the shape space such that $\boldsymbol{s}_t = \sum_{q=1}^{Q} w_t^q \boldsymbol{s}^q, \sum_{q=1}^{Q} w_t^q = 1$. The shape state at time t is denoted by λ_t and represented by the coefficients w_t^q, i.e., $\lambda_t = [w_t^1, \cdots, w_t^Q]$.

5. Tracking on the Manifold Using Particle Filtering

The generative models in Eq. 3 and 4 fit directly to the Bayesian tracking framework to generate observations from the state $\boldsymbol{X}_t$ to estimate the observation likelihood $P(\boldsymbol{Y}_t|\boldsymbol{X}_t)$ at time t. The state is represented by the view parameter θ_t and the configuration parameter β_t, and shape parameter λ_t, i.e., $X_t = (\theta_t, \beta_t, \lambda_t)$. We use a particle filter to realize the tracker. Separate particle representations for the view manifold, configuration manifold, and shape space are used.

For a body configuration particle i, a view particle j, and a style particle k, the observation probability can be computed $P(\boldsymbol{y}_t|\theta_t, \beta_t, \lambda_t) = N(\boldsymbol{\mathcal{A}} \times_1 \sum_{q=1}^{Q} w_t^q \boldsymbol{s}^q \times_2 g_v(\theta_t) \times_3 \psi(\beta_t), \boldsymbol{\Sigma})$ with observation covariance $\boldsymbol{\Sigma}$, to update the particles' weights. To propagate the particles, we use the flow field to propagate the body configuration particles and a random walk to propagate both the view and shape particles. For one-dimensional motions, we use a constant speed dynamic

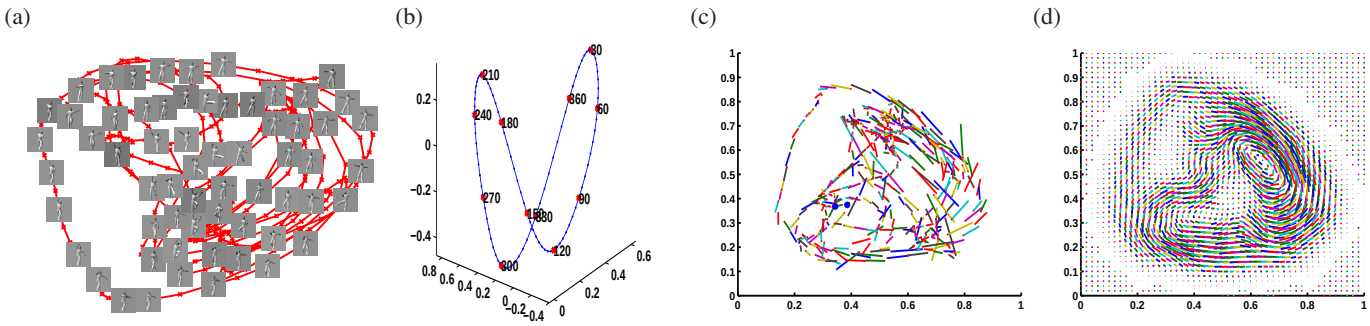

Figure 3. Configuration and View Manifold for a ballet motion: (a) Embedded kinematic manifold in a 2D (b) One-dimensional configuration-invariant view manifold embedding (The first three dimensions are shown) (c) Velocity field on the configuration manifold (d) Interpolation of the velocity field.

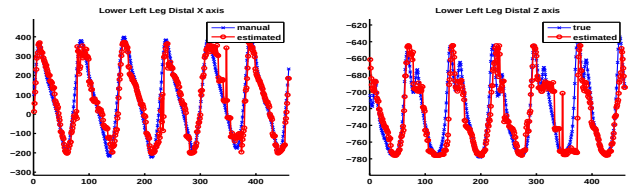

Figure 5. Evaluation of joints' location estimation (HUMANEVA-I): estimated joint's locations and ground truth for each frame: x and z values for *Lower left leg distal*

model, which directly follows by construction from the spline fitting and leads to superior tracking results.

6. Experimental Results

We tested the performance of our approach with different types of motions using synthetic data and real data. In order to learn the model, *for all the experiments, unless otherwise stated, we use the following setting:* We used synthesized shapes rendered from discrete views from real motion-captured data. To evaluate the approach we used both synthesized and real data. The synthesized data facilitates quantitative analysis of the configuration and view estimation. In the experiments shown here, we mainly used silhouettes to represent the observations. However, the approach provides a generative model for contours and can be easily integrated with edge based observations with a proper observation model. We used an implicit function representation for the silhouettes. To evaluate the 3D configuration estimation, the embedded body configuration is mapped to a 3D joint angles' location space through learning an RBF mapping from the embedding space to the joint angles' space.

6.1. Estimation Using View Manifold and One-Dimensional Motion Manifold

Brown HUMANEVA-I dataset: We tested the 3D body posture estimation accuracy using the Brown HUMANEVA-I dataset [14], which provides ground truth data for 3D joint locations for different types of motions. We used 3 circular walking sequences, which have continuous view variations w.r.t. the camera. We normalized the original joints' locations in the HUMANEVA-I dataset into a body-centered coordinate system. We trained the model using synthetic data with 12 discrete views rendered based on our own motion-captured walking motion, i.e., we did not learn the model from any of the subjects in the HUMANEVA-I dataset. Fig. 4 shows the estimated view, body configuration, and corresponding 3D body posture reconstruction. The estimated parameters fit very well a constant speed linear dynamic system for both the configuration and view parameters. Average errors of all joint angles are 26.29 mm for subject $S1$ during 512 validation frames, 25.38 mm for $S2$ during 439 frames, and 30.61 mm for $S3$ during 348 frames. Fig. 5 shows examples of two joint angles reconstruction. The tracking is achieved using only 30 particles for configuration and 30 particles for view.

Golf swing - one dimensional open manifold: A golf swing is a one-dimensional non-periodic motion. Fig. 6-g shows an embedding of a golf swing kinematic manifold. We learned view manifold after synthesizing 107 frames in 12 discrete views from motion-captured data. We tested on synthetic date with a simulated continuous constant speed camera motion in a 360° circular trajectory during the golf swing motion. Fig. 6-d shows the estimated joint pdf using 30 particles for the body configuration β and 30 particles for the view θ. The estimated view in Fig. 6-f correctly reflects the constant change of view (notice that only 12 views are learned, and all intermediate views were correctly estimated).

Shape-adaptive tracking: To show the generalization to track different subjects and adapting to their shapes, we captured 8 different views of 4 subjects walking on a treadmill. Given such data, we fit the model in Eq. 4. We tested the model on outdoor sequences where people are walking in S-shaped trajectories. Fig. 7-d shows the estimated view, which exhibits directional change due to the S-shape walking trajectory. The estimated view parameter decreased from 1 to 0.5 (180 degree counter-clockwise variations), and then it changed back from 0.5 to 1 (180 degree clockwise variations), which simulates the actual S-shape walking pattern.

6.2. Estimation From General Motion Manifolds

Simple sport motions like ball passing, catch/throw can not be parameterized by a one-dimensional manifold due to the

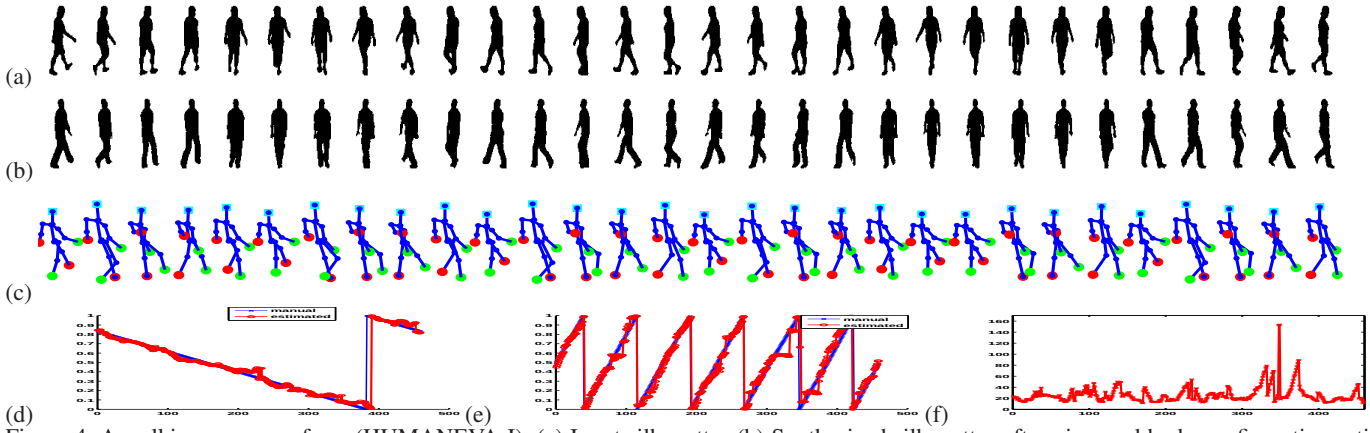

Figure 4. A walking sequence from (HUMANEVA-I): (a) Input silhouettes (b) Synthesized silhouettes after view and body configuration estimation (c) Reconstructed 3D postures (d) Estimated view parameter (e) Estimated body configuration parameter (f) Joint location error in each frame in *mm*.

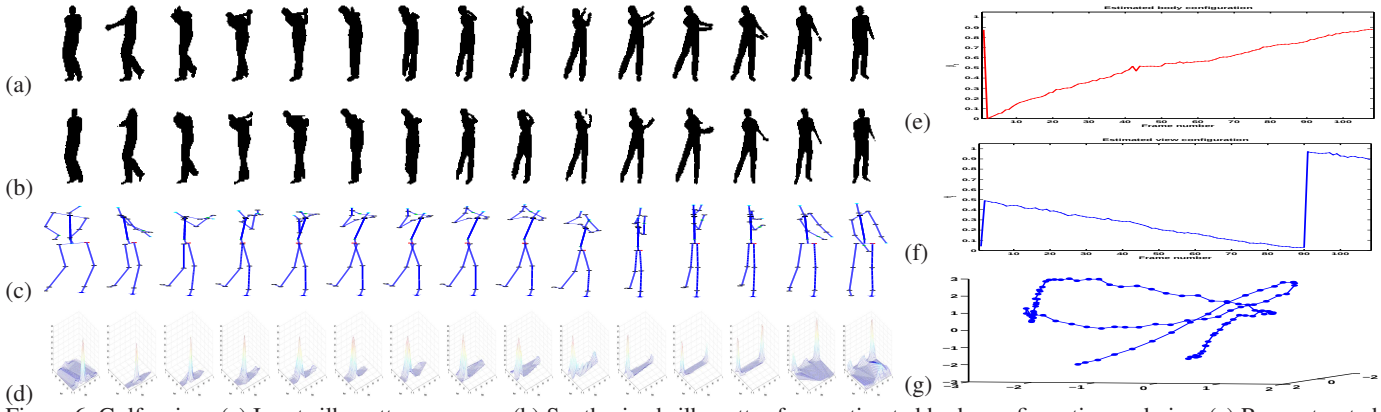

Figure 6. Golf swing: (a) Input silhouette sequences (b) Synthesized silhouettes from estimated body configuration and view (c) Reconstructed 3D body posture (d) Estimated probability densities for view and body configuration parameters (e) Estimated body configuration parameter (f) Estimated view parameter (g) an embedding of the kinematic manifold.

variability in body configuration when the motion is repeated. For example, when we catch and throw a ball repeatedly in the air, the catch action changes according to the falling ball locations. Moreover, many activities like dancing, aerobics are high dimensional in their kinematic manifold.

Catch/throw motion: We used catch and throw sequences with variations of motion in each catch and throw cycles, which is represented as different trajectories in the body configuration embedding. We used 90 and 60 particles for configuration and view. Fig. 8 shows the results with details in the caption. Fig. 8-f shows the estimated view for the test sequence shown in Fig. 8-b, which exhibits camera motion with a constant speed.

Aerobic Dancing Sequence A two-dimensional manifold embedding is used to represent a repetitive dancing sequence as in Fig 9-b. Fig. 9-c shows the learned configuration-invariant view manifold. We tested the performance of the view and body configuration estimation using synthetic rendered data. Fig. 10-a-d show the results for a test sequence with view variations from 0° to 90°. Average joints' location error in each frame is shown in Fig. 10-d. We used 60 particles for each of the configuration and the view.

Ballet Motion: Ballet motion has frequent body rotation and the motion is very complicated since arms and legs are moving independently. However, the motion is still constrained by the physical dynamics in the motion and the rules in the ballet dancing. Fig. 10-e,f,g show the estimated configuration and view for the ballet motion that was shown in Fig. 1 and Fig 3. We used 100 and 30 particles for configuration and view.

6.3. Comparative Evaluation:

We evaluated the performance of the proposed approach compared to other representations. The goal is to compare other representations for the visual manifold with both view and configuration variability. We rendered 12 discrete views along a view circle. We collected one cycle with 40 frames for each view. The silhouette images are cropped and normalized to 50×40. So, each silhouette image is represented by 2000 dimensional vector.

We compared the performance of the proposed representation with a nearest-neighbor (NN) search, inferring body pose using a torus manifold embedding [7], and an embedded representation using Gaussian process latent variable model (GPLVM) [6]. For the case of NN, we directly find out 3D pose from the nearest training instance. For GPLVM, given an

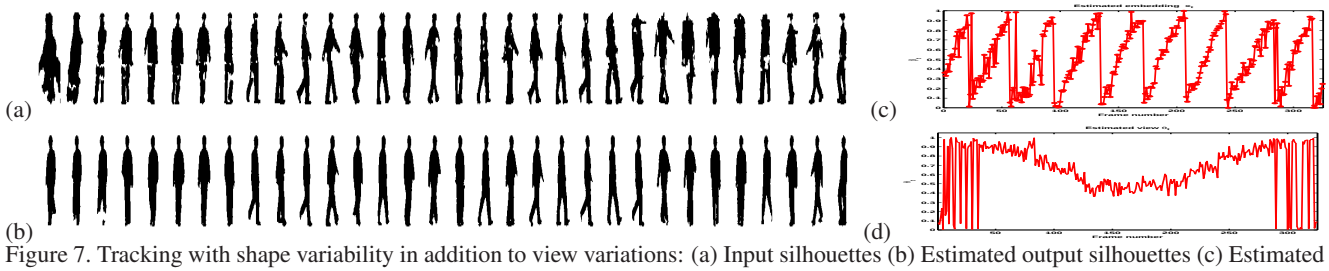

Figure 7. Tracking with shape variability in addition to view variations: (a) Input silhouettes (b) Estimated output silhouettes (c) Estimated body configuration parameter (d) Estimated view parameter.

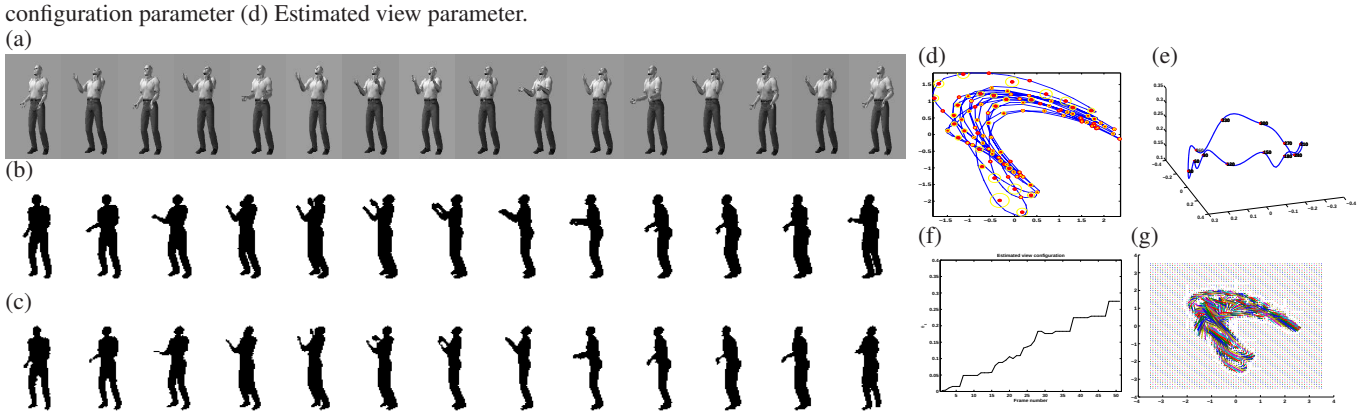

Figure 8. Catch/throw motion (Evaluation): (a) Rendered image sequence (frames 3, 25, 47, 69, $\cdots$, 333) (b) A test sequence with a moving camera (c) Estimated shape sequence after view and configuration estimation (d) Two-dimensional configuration manifold embedding and selected basis points. (e) configuration-invariant view manifold in a 3D space (f) Estimated view. (g) moton flow field on the embedding space.

embedding of the data, we can directly find the embedding for each input image silhouette. We used the provided optimization routine to find out embedding points for a given input. The test data is a walking 3 cycles sequence with continuous view variations. To compare the performance, any recovered embedding points is mapped to a 3D body joints' location using similar RBF mapping for all the approaches (except NN). The average error is shown in Table 1. The proposed approach shows better performance. In this experiments, NN shows relatively good results since the test data does not have any noise and the training posture and view have dense samples. GPLVM has some problems in this experiment due to ambiguity of body pose in different views.

Table 1. Average error (in inches) in normalized 3D body pose estimation

Approaches	Proposed	NN	Torus [7]	GPLVM [6]
Average Error	0.79	0.86	2.46	4.88

7. Conclusions

In this paper we introduced an approach for explicit modeling of body configuration and view with two separate low dimensional embedded representations. The body configuration is embedded from kinematic data, i.e., invariant of the view. The view is represented in a posture-invariant manner. As a result, we have a generative model that parameterizes the motion, the view, and the shape style. The model is appropriate for tracking and pose estimation of complex motion from un-calibrated stationary or moving camera. We showed several experimental results and quantitative evaluations for wide variety of motion including simple motion, such as gait and golf swings, to complex motions, such as aerobics and ballet dancing. The model can successfully self initialize, track, and recover the parameters for view and 3D configurations even with a moving camera. The results shows superior tracking of both configuration and view with only 30 particles used for each.

Acknowledgment: This research is partially funded by NSF CAREER award IIS-0546372.

References

[1] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevance vector regression. In *Proc. of CVPR*, volume 2, pages 882–888, 2004. 1

[2] C. M. Christoudias and T. Darrell. On modelling nonlinear shape-and-texture appearance manifolds. In *Proc. of CVPR*, volume 2, pages 1067–1074, 2005. 1

[3] A. Elgammal and C.-S. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *Proc. of CVPR*, volume 2, pages 681–688, 2004. 1

[4] A. Elgammal and C.-S. Lee. Separating style and content on a nonlinear manifold. In *Proc. of CVPR*, volume 1, pages 478–485, 2004. 3

[5] L. D. Lathauwer, B. de Moor, and J. Vandewalle. A multilinear singular value decomposiiton. *SIAM Journal On Matrix Analysis and Applications*, 21(4):1253–1278, 2000. 2, 3, 4

[6] N. D. Lawrence. Gaussian process models for visualisation of high dimensional data. In *Proc. of NIPS*, 2004. 2, 6, 7

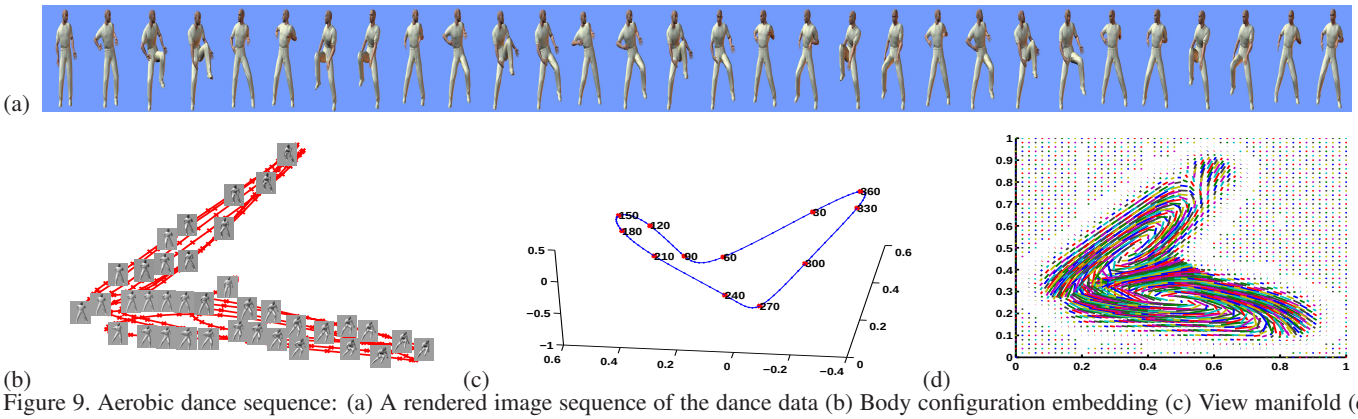

Figure 9. Aerobic dance sequence: (a) A rendered image sequence of the dance data (b) Body configuration embedding (c) View manifold (d) Velocity field.

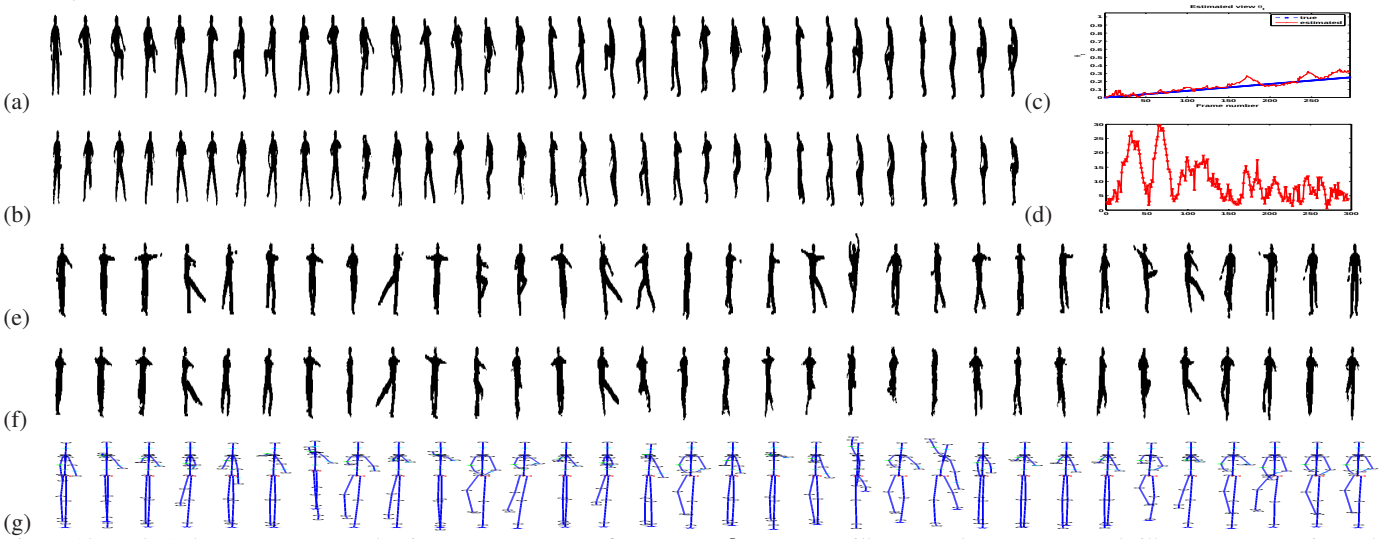

Figure 10. (a-d) A dance sequence evaluation - camera moves from $0 - 90^o$: (a) Input silhouettes (b) Reconstructed silhouettes. (c) Estimated view parameter (d) Average joints' location error in a each frame in *mm*. (e-g) A ballet motion: (e) A test input sequence. (f) Synthesis of silhouettes based on estimated body configuration and view. (g) Reconstruction of 3D body posture based *Shown in a body-centered coordinate from a frontal view without body rotation*.

[7] C.-S. Lee and A. Elgammal. Simultaneous inference of view and body pose using torus manifolds. In *Proc. of ICPR*, pages 489–494, 2006. 1, 6, 7

[8] V. I. Morariu and O. I. Camps. Modeling correspondences for multi-camera tracking using nonlinear manifold learning and target dynamics. In *Proc. of CVPR*, volume 1, pages 545–552, 2006. 1

[9] H. Murase and S. Nayar. Visual learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995. 1

[10] R. Rosales, V. Athitsos, and S. Sclaroff. 3d hand pose reconstruction using specialized mappings. In *Proc. of ICCV*, pages 378–387, 2001. 1

[11] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linar embedding. *Science*, 290(5500):2323–2326, 2000. 2

[12] B. Schoelkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002. 3

[13] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *Proc. of ECCV*, volume 2, pages 702–718, 2000. 1

[14] 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. 5

[15] C. Sminchisescu and A. Jepson. Generative modeling of continuous non-linearly embedded visual inference. In *ICML*, pages 140–147, 2004. 1

[16] J. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319 – 2323, 2000. 2

[17] J. B. Tenenbaum and W. T. Freeman. Separating style and content with biliear models. *Neural Computation*, 12:1247–1283, 2000. 3

[18] R. Urtasun, D. J. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In *Proc. of ICCV*, pages 403–410, 2005. 1

[19] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *Proc. of ECCV*, pages 447–460, 2002. 4

[20] J. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models. In *Proc. of NIPS*, pages 1441–1448. 2