

Simultaneous Inference of View and Body Pose using Torus Manifolds

Chan-Su Lee and Ahmed Elgammal
Rutgers University
{chansu,elgammal}@cs.rutgers.edu

Abstract

Inferring 3D body pose as well as viewpoint from a single silhouette image is a challenging problem. We present a new generative model to represent shape deformations according to view and body configuration changes on a two dimensional manifold. We model the two continuous states by a product space (different configurations \times different views) embedded on a conceptual two dimensional torus manifold. We learn a nonlinear mapping between torus manifold embedding and visual input (silhouettes) using empirical kernel mapping. Since every view and body pose has a corresponding embedding point on the torus manifold, inferring view and body pose from a given image becomes estimating the embedding point from a given input. As the shape varies in different people even in the same view and body pose, we extend our model to be adaptive to different people by decomposing person dependent style factors. Experimental results with real data as well as synthetic data show simultaneous estimation of view and body configuration from given silhouettes from unknown people.

1. Introduction

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. Despite the high dimensionality of the body configuration space, many human motion 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. The observed motion, in terms of body shape contour, lies on a low dimensional manifold as well (visual manifold). However, the observed motion manifold changes given the viewpoint.

Recently, there are several approaches to model human motion with view variant observations. Separate view-based shape representations are used to model changes of shape in different views with separate dynamic mod-

els in [4, 2]. In [8], shape variations in different views as well as configuration variations due to motion dynamics are integrated with pictorial view-dependent shape models and Hidden Markov Models(HMMs). In [11, 5], view invariant representations of body configurations using visual hulls are computed from multiple calibrated camera. Canonical projection after camera calibration are also used for view invariant gait recognition and tracking [7]. Appearance manifold representations from dense view variant images are explored for object recognition in arbitrary views [9]. However in these appearance manifolds does not cover dynamic variations of appearances, which are essential in human motion analysis. Our work can be regarded as extending view manifold representation to include dynamic objects by modeling shape variation manifold due to configuration changes in addition to view variations.

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. 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 where walking and running are the most frequent motion as well as in sport analysis. We model view and configuration manifold on a product space (different configuration \times 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 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 as explained in Sec. 2.

Given a shape instance, we need to recover both the body configuration and view using the learned model. Inferring body pose and view is formulated as estimating embedding points from a given input as there is a corresponding em-

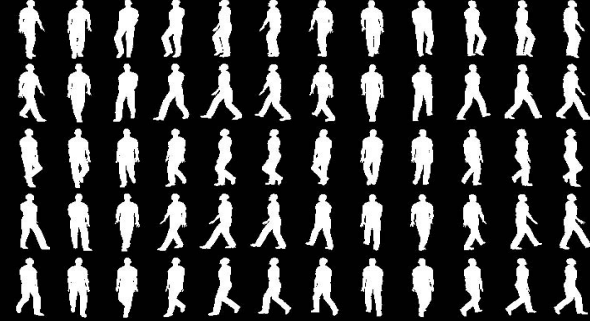


Figure 1. Example shape variations in view and body configuration change: Rows: body pose at $0, \frac{1}{5}T, \frac{2}{5}T, \frac{3}{5}T, \frac{4}{5}T$ (T : one cycle period). Cols: view $0, 30, 60, \dots, 330$.

bedding point for every body configuration and view on the continuous torus manifold. We can estimate an embedding point for a given image using approximation of inverse mapping from manifold points to visual inputs with constraints as explained in Sec. 3. Experiments with synthetic silhouette images show accurate estimation of body configuration and view from view variant walking sequences in Sec. 4.1. From real data, we have to count variations of silhouettes in different people. We model individual differences from the variations of mappings between the embedding and input silhouette shapes in Sec. 2.1. In addition to estimation of person style factor, we further tuned view and configuration estimation by two one dimensional search on the torus manifold based on sampling in Sec. 3.2. Experimental results with real data show simultaneous estimation of view and body configuration from given silhouettes as well as person style in Sec. 4.2.

2 Modeling View and Configuration: Torus Manifold Embedding

We use a torus manifold as a representation of the joint view and configuration state. A torus manifold, a two dimensional manifold embedded in a three dimensional Euclidean space with a single hole, is useful to represent periodic dynamic human motion observed from a viewing circle. The body configuration in periodic motion observed from a single view is a one dimensional manifold and homeomorphic to a circle. The view manifold given a fixed body configuration also can be modeled as a closed one dimensional manifold that is also homeomorphic to another circle. The view and the body configuration are independent. The joint manifold representing body configuration and view can be as a two dimensional manifold with two orthogonal coordinate axes: one for view and the other for body configuration.

The torus manifold can be constructed from a rectan-

gle, which can be represented by two orthogonal coordinate with range $[0, 1] \times [0, 1]$, by gluing both pairs of opposite edges together with no twists [6]. The view and body configuration manifold can be parameterized in the rectangle coordinate with two orthogonal axis on the torus manifold. Any manifold point in the torus can have two circles :one is in the plane of the torus and the other is perpendicular to it.

We represent any view and configuration as a point on the torus manifold with two independent parameters. Let the radius from the center of the hole to the center of the torus tube be R_c , and the radius of the tube be R_a , then a torus symmetric about the z axis can be described

$$(R_c - \sqrt{x^2 + y^2})^2 + z^2 = R_a. \quad (1)$$

It can be parameterized by μ, ν as

$$\begin{aligned} x &= (R_c + R_a \cos 2\pi\nu) \cos 2\pi\mu, \\ y &= (R_c + R_a \cos 2\pi\nu) \sin 2\pi\mu, \\ z &= R_a \sin 2\pi\nu, \end{aligned} \quad (2)$$

where $u, v \in [0, 1]$. Fig. 2 (a) shows coordinate of a torus manifold when $R_c = 2, R_a = 1$.

The torus can be used as a conceptual embedding for the joint view and configuration manifold. Given the torus manifold, we can learn a nonlinear mapping between points on the torus and the input sequences with continuous view and pose variations. By Eq. 2, we can represent any point on the torus manifold by a function g of the two variables μ and ν as $[x \ y \ z] = g(\mu, \nu)$. We define an empirical kernel map [10] as $\psi_N(\mathbf{x}) : \mathbb{R}^3 \rightarrow \mathbb{R}^{N_\mu \times N_\nu + N_p}$

$$\psi(\mathbf{x}) = [\phi(\mathbf{x}, \mathbf{z}_{11}), \phi(\mathbf{x}, \mathbf{z}_{12}), \dots, \phi(\mathbf{x}, \mathbf{z}_{N_\mu N_\nu}), 1, \mathbf{x}^T]^T \quad (3)$$

where \mathbf{z}_{ij} are representative points on the torus as kernel centers with N_μ is the number of kernel centers in μ axis and N_ν is the number of kernel center in ν axis. The total number of kernel is $N = N_\mu \times N_\nu$. N_p is the dimension of polynomial terms used for approximate solution of embedding points.

Given input shapes with all views and all poses \mathbf{y}_{vb} and their corresponding torus embedding \mathbf{x}_{vb} , we can learn a nonlinear mapping in the form

$$\mathbf{y}_{vb} = \mathbf{D} \cdot \psi(\mathbf{x}_{vb}) = \mathbf{D} \cdot \psi(g(\mu_v, \nu_b)). \quad (4)$$

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. Fig. 2 (c) (d) (e) show examples of generation of new sequence by different sampling of embedding points on the torus manifold. Corresponding sample trajectories are shown in Fig. 2(b).

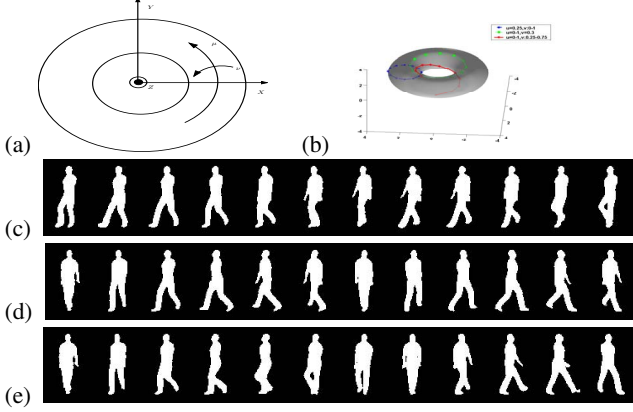


Figure 2. Torus embedding of continuous view and pose variant dynamic human motion: (a) 3D Cartesian coordinate and μ, ν axis. (b) Trajectory on torus manifold according to view and configuration change. (c) Synthesis of body pose variation in fixed view ($\mu = 0.25, \nu : 0 \rightarrow 1$). (d) Synthesis of view variations in fixed body configuration. $\mu : 0 \rightarrow 1, \nu = 0.3$. (e) Synthesis of view and body configuration variation. $\mu : 0 \rightarrow 1, \nu : 0.25 \rightarrow 0.75$

2.1 Multiple People View and Configuration Models

For the same view and body configuration, different people shows different silhouette shapes. In our previous work [3], we presented an approach for decomposing *style* variations in the space of nonlinear mapping coefficients from an embedded manifold to the observation space. In that work, a single view is considered where different people shapes are considered to be the style variability. However, the solution provided in [3] cannot be generalized to the case of different views since at each view, the body configuration manifold will be quite different. We extend the framework presented in [3] to model person-specific shape style given the embedding of both the view and motion manifolds described earlier.

Since we have a unique embedding of view and configuration on a torus manifold, the differences of silhouette among different people with the same view and configuration is reflected on the nonlinear mapping. Given N_p persons' data with different views and configurations, we learn N_p different person-specific mappings where for person p the mapping is

$$\mathbf{y}_{vb}^p = \mathbf{D}^p \cdot \psi(\mathbf{x}_{vb}) = \mathbf{D}^p \cdot \psi(g(\mu_v, \nu_b)). \quad (5)$$

After arranging each mapping coefficient as a vector \mathbf{d}^p by column stacking \mathbf{D}^p , we can arrange these coefficient vectors as a matrix

$$\mathbf{C} = [\mathbf{d}^1 \ \mathbf{d}^2 \ \dots \ \mathbf{d}^{N_p}] \quad (6)$$

where each column is the person-specific mapping coefficients. By decomposing the matrix \mathbf{C} using singular value

decomposition as $\mathbf{C} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ we can obtain the basis for the shape style space. Therefore, We can obtain person-specific mapping coefficient vector $\mathbf{d}^p = \mathbf{A}\mathbf{p}$ where $\mathbf{A} = \mathbf{U}\mathbf{S}$ and \mathbf{p} is person style vector. Now person dependent view and configuration model can be represented by

$$\mathbf{y}_{vb}^p = \mathbf{A} \times \mathbf{p} \times \psi(g(\mu_v, \nu_b)). \quad (7)$$

Where \mathcal{A} is the third order tensor with dimensions $N_\nu \times N_\mu \times N_p$ obtained from be restacking matrix \mathbf{A} accordingly.

3 Inferring View and 3D Body Pose

3.1 View and Configuration Estimation

We can estimate view and pose directly from the torus embedding coordinates. A solution for view and body configuration \mathbf{x}^* can be obtained by least square solution for the nonlinear system

$$\mathbf{x}^* = \arg \min_{\mathbf{D}} \|\mathbf{y} - \mathbf{D} \cdot \psi(\mathbf{x})\|^2. \quad (8)$$

We can find embedding coordinate \mathbf{x} from Eq. 4 using the pseudo-inverse of the mapping function.

$$\psi(\mathbf{x}) = \mathbf{D}^+ \mathbf{y} \quad (9)$$

By utilizing the linear polynomial term in the empirical kernel map, we can achieve a closed-form least square linear approximation by the pseudo-inverse of the coefficient matrix \mathbf{D}^+ . Vector $\psi(\mathbf{x})$ can be recovered by $\psi(\mathbf{x}_t) = \mathbf{D}^+ \mathbf{y}$. Linear approximation for the embedding coordinate \mathbf{x}^* can be obtained from the polynomial term $[1 \ \mathbf{x}^*]$ in $\psi(\mathbf{x}^*)$.

We can estimate the view configuration μ and body configuration ν directly from the estimated embedding point. For a given embedding coordinate x^*, y^* , and z^* , the view and body configuration ν^*, μ^* have the unique solution

$$\begin{aligned} \nu^* &= \frac{1}{2\pi} \tan^{-1} \left(\frac{z^*}{\sqrt{x^{*2} + y^{*2}} - R_c} \right), \\ \mu^* &= \frac{1}{2\pi} \tan^{-1} \left(\frac{y^*}{x^*} \right) \end{aligned} \quad (10)$$

with additional constraints of Eq. 1, which force the estimated coordinate stay on the surface of the torus. As a result, we achieve estimation of μ and ν in a closed-form approximation for a given input without any iterative estimation. Fig. 3 shows embedding of collected view and configuration samples used for training (a), estimated coordinate points \mathbf{x}^* without constraints (b), where some of the estimated coordinate points are out of the embedding manifold, and with constraints (c). Final estimation of view parameter and configuration parameter by Eq. 10 is shown in (d) and (e). It can be noticed that the estimated configuration parameter changes linearly in each walking cycle since the configuration is embedded at equidistant points on the torus embedding.

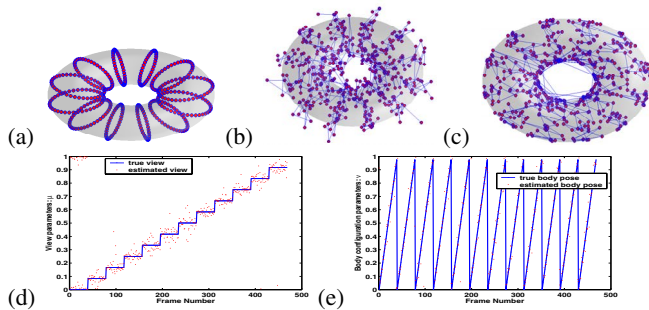


Figure 3. Embedding and Estimation of View and Configuration for Training Sequences: (a) Embedding for training sequences. (b) Approximate coordinate estimations. (c) Projection on the coordinate manifold. (d) True and estimated view parameters. (e) True and estimated body configuration parameters.

3.2 Estimating View and Configuration in Multiple People

We need to estimate person style parameters in addition to the view and configuration parameters for multiple people model in Sec. 2.1. We perform further two one dimensional searches of embedding parameter for robust estimation of view and configuration after approximate style estimation.

Person Style Estimation: If the person vector \mathbf{p} is known, we can solve for the view and body configuration using linear polynomial term. For known person with style parameter \mathbf{p}^{est} , we can compute the mapping coefficients $\mathbf{D}^{est} = \mathcal{A} \times \mathbf{p}^{est}$. So, Eq. 5 can be written $\mathbf{y}^{new} = \mathbf{D}^{est} \psi(\mathbf{x})$ and we can estimate view and configuration for a given input as described in Sec. 3.1.

On the other hand, if the view and body configuration are known, we can solve for person vector \mathbf{p}^{est} using the known embedding coordinate. For given person classes \mathbf{p}^k s, and view and body configuration \mathbf{x} the observation can be considered as drawn from a Gaussian mixture model centered at $\mathcal{A} \times \mathbf{p}^k \times \psi(\mathbf{x})$ for each person class k . The observation probability $p(\mathbf{y}|\mathbf{p}^k, \mathbf{x})$ can be computed as

$$p(\mathbf{y}|\mathbf{p}^k, \mathbf{x}) \propto \exp(-\|\mathbf{y} - \mathcal{A} \times \mathbf{p}^k \times \psi(\mathbf{x})\|^2 / (2\sigma^2)) \quad (11)$$

We can approximate a new person style vector as a linear combination of the person classes learned from the training data as $\mathbf{v}^{new} = \sum_k w_k \mathbf{v}^k$ where the weights w_k are set to be $p(\mathbf{p}^k|\mathbf{x}, \mathbf{y})$. As we can solve for person style vector by estimating weight for known view and body configuration and we can solve for view and body configuration for known person vector, we can estimate view and body configuration in an EM-like iterative procedure. The initial view and configuration estimation can be started using a mean person style vector $\bar{\mathbf{p}}$

View and Configuration Parameter Search: In order to achieve robust estimation of view and body configuration,

we search view and body configuration parameters based on sampling along the embedding manifold points. As the estimation of view and body configuration depends on the accuracy of style estimation based on least-square solutions in the nonlinear mapping in Sec. 3.1, the estimation of view and configuration are vulnerable to inaccurate style estimation. We can achieve fine tuning of view and configuration estimation by utilize the generative power in our model.

As our model can generate shapes corresponding to any view and configuration parameter with preserving nonlinear properties in shape variations, we can compare any given input silhouette with sample silhouettes generated for all the possible view and configuration parameters. But generating all the samples that cover the embedding torus manifold will be too expensive to be useful in real applications. However, if we assume one of the parameter is known, estimation of the other parameter can be one dimensional search along the orthogonal axis for the known parameter. For example, if we know the view parameter for given input, then we can sample one dimensional configuration parameter in the given view plane, which is a circle on our torus manifold.

$$\nu^{i*} = \arg \min_i \|\mathbf{y}^{input} - \mathcal{A} \times \mathbf{p}^i \times \psi(g(\mu^*, \nu^i))\|,$$

where $\nu^i = \frac{i}{N_s}$, $i = 1 \dots N_s$, N_s is sample number. Similarly, based on estimated configuration parameter and person parameter, we can estimate view with sampling along a view embedding circle constrained by estimated configuration parameter. We can repeat these iterations several times. Our experimental results in Sec. 4.2 show that even one iteration of the two one dimensional search of view and the configuration after style estimation shows large improvement of view and configuration parameter estimation.

3.3 3D Reconstruction

We can reconstruct 3D body pose directly from configuration parameter ν . To reconstruct 3D body pose, we learn RBF interpolation function between the body joint coordinates and body configuration parameter on the torus. We represent the 3D body pose using 18 joints' model and each joint is represented by its coordinates in a body centered global coordinate system. As a result, for any body configuration parameter ν , we can reconstruct 3D body pose as shown in the following experimental results.

4 Experimental Results

4.1 Synthetic Gait with Circular View Variation

Synthetic data are used to evaluate accuracy of the view and configuration estimations with known ground truth values. We collected walking sequence from 12 different views

with view interval $30^\circ = \frac{360}{12}$ using Poser[®] animation software based on motion capture data. 12 sequences are used for learning torus manifold embedding for view and configuration with $N_\mu = 12$, and $N_\nu = 14$. Fig. 1 shows some examples of normalized silhouette sequences used for learning nonlinear mapping between torus manifold embedding and visual input. Fig. 3 (a) shows embedding of image sequences for training data.

We collected three cycle gait sequences with circular view variations to test the accuracy of estimating intermediate views and configurations. During three walking cycles, the view was changed in a constant speed starting at 0° to 360° view. We compute true body pose embedding based on detected cycles. Each cycle embedded in equally separated points on the torus manifold. Fig. 4 shows input silhouette sequence samples (a), and generated new silhouettes (d) using estimated view (b) and body configuration (c). Estimated view parameter shows continues change of view. Compared to the true view parameter, the average error in view estimation μ is 0.0495, which is corresponding to 17.8° error. For the body configuration, when we assume constant speed walking and equidistant embedding based on segmented cycle, the average error in body pose parameter ν is 0.084, which is corresponding to around 3.4 frame difference in pose estimation when one cycle is 40 frames. Average reconstruction error of 3D position of all the joints (including end segment) was 2.16 inch. Fig. 4 (e) shows estimated embedding trajectory points on the torus manifold. When we performed the same experiment with dense view samples (10° interval, which 3 times more data than the previous ones), we can achieve more accurate estimations of view and body configuration parameters: average view error 13.3° , configuration 2.7 frame, and 3D reconstruction error 1.86 inch per joint location.

4.2 Estimation of View and Body Configuration from Real Data

For real data set, we collected walking sequences on a treadmill for five people with 11 different views around circle in the same camera height. Fig. 5 shows an example of normalized silhouette shape representation. To achieve consistent shape representation in different people, we performed normalization with silhouette correction by applying image filtering, cropping, manual hole filling and resizing for the extracted silhouette image. We parameterized shape contour using signed distance function for robust shape representation in learning and matching shape contour as in Fig. 5 (d).

Estimation of View and Body Pose in Intermediate Views: We tested estimation of view and body pose for new camera views which are not used for learning. We used 7 views for training and 4 other views of 140 frames are

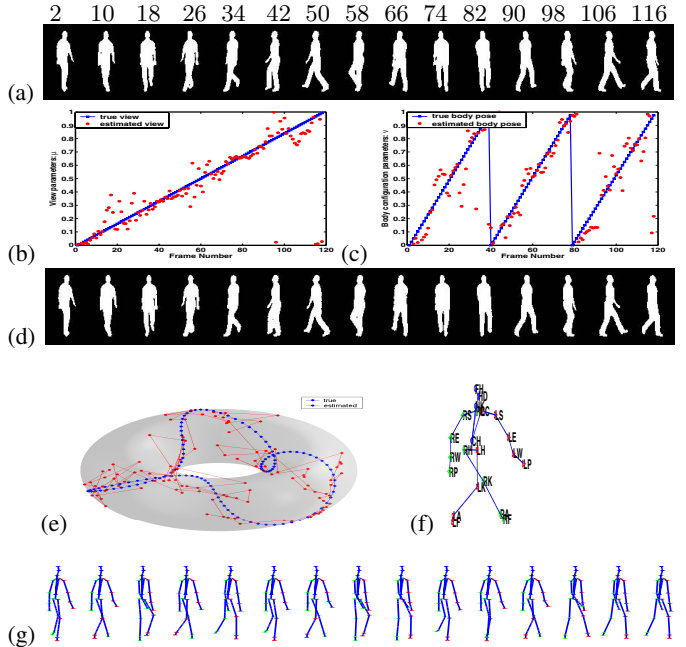


Figure 4. Estimation of view and body configuration for synthetic data: (a) Input silhouettes. (b) Estimated and true view parameters. (c) Estimated and true body pose parameters. (d) Reconstructed silhouettes based on view and body configuration estimation. (e) Estimated and true torus embedding trajectory. (f) 3D model used for reconstruction. (g) 3D reconstruction based on estimated body pose parameters.

used for testing intermediate view and configuration estimations. 140 frames from 4 different views in a cycle are used. Fig. 6 shows experimental results. The experimental result shows that accurate estimation of body configuration even though some of the view shows offset in the estimated view parameters.

View and Body Pose Estimation in Multiple People: We collected data for four people with seven different views to learn a generative model with four dimensional person vector ($N_p = 4$) as explained in Sec. 2.1. Fig. 7 (a-d) show after person style parameter estimation using closed form solution. You can see errors in view and configuration by comparing the reconstructed image and original input images. By additional two one-dimensional search for view and body configuration parameters as proposed in Sec. 3.2 improves estimation results in Fig. 7(e-i). The reconstructed silhouette shapes are closed to the input shape and the estimated view parameter (g) is close to the true view parameter $\mu = 0.125$.

5 Conclusions

We formulated inferring view and body pose estimation as estimating embedding point on torus manifold, where we

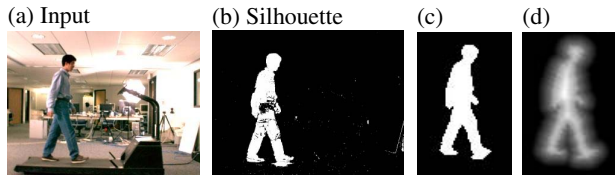


Figure 5. Captured data and its shape representation: (a) An example of captured image frame. (b) An extracted silhouette after background subtraction. (c) Normalized silhouette with corrections. (d) Shape representation based on signed distance.

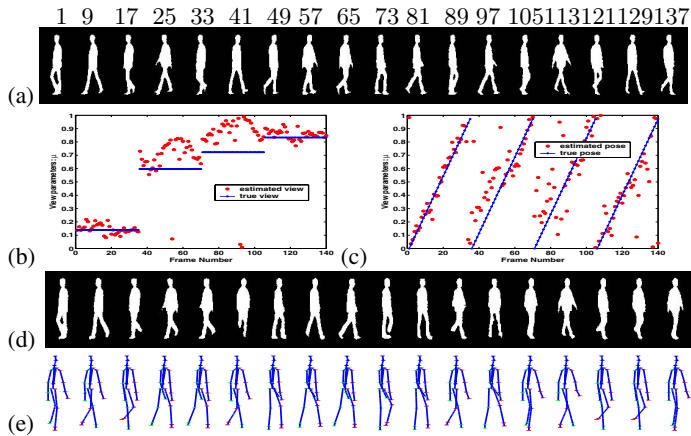


Figure 6. Estimation of view and body configuration for intermediate fixed view : (a) Input silhouettes. (b) Estimated and true view parameters. (c) Estimated and true body pose parameters. (d) Reconstructed silhouettes based on view and body configuration estimation. (e) reconstructed 3d pose

model cyclic view and configuration change on two dimensional manifold. In order to estimate view and configuration in real data from multiple people, we decomposed person dependant factors in the mapping space. Estimation of person factor and embedding in iterative way allows estimation of view and configuration from real new person data. The proposed simultaneous view and body pose inferring from a single image will be very useful for initialization for tracking human motion in arbitrary view and other recognition of human activities based on inferred body configuration.

Acknowledgement This research is partially funded by NSF award IIS-0328991

References

- [1] A. Elgammal. Nonlinear manifold learning for dynamic shape and dynamic appearance. In *Workshop Proc. of GMBV*, 2004.
- [2] 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.

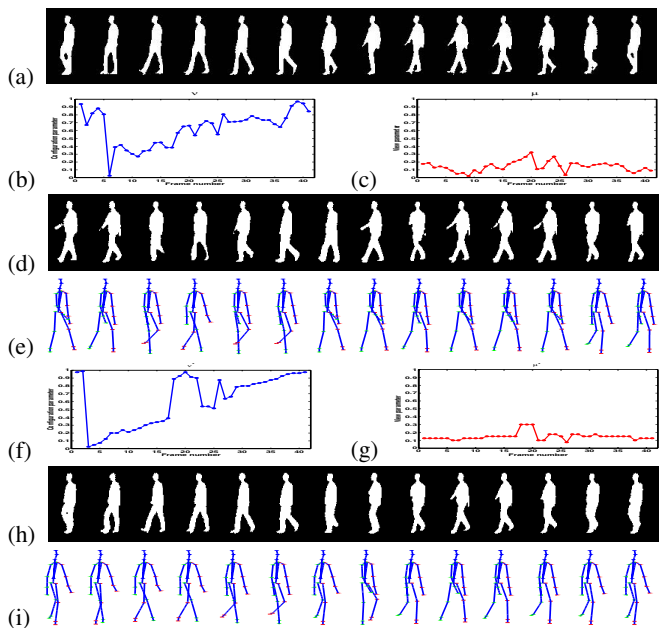


Figure 7. View and configuration estimation in multiple people: (a) Input silhouettes. (b) Estimation of view parameters. (c) Estimation of body configuration. (d) Shape synthesis based on best estimated shape and view parameter. (e) Inferred body configuration. (f) (g) (h) (i): Estimated view, estimated body configuration, synthesized shape and inferred body configuration after additional two dimensional search for view and body configuration.

- [3] A. Elgammal and C.-S. Lee. Separating style and content on a nonlinear manifold. In *Proc. CVPR*, volume 1, pages 478–485, 2004.
- [4] R. Fablet and M. J. Black. Automatic detection and tracking of human motion with a view-based representation. In *Proc. ECCV 2002, LNCS 2350*, pages 476–491, 2002.
- [5] K. Grauman, G. Shakhnarovich, and T. Darrell. Inferring 3d structure with a statistical image-based shape model. In *Proc. of ICCV*, volume 1, pages 641–647, 2003.
- [6] A. Gray. *Modern Differential Geometry of Curves and Surfaces with Mathematica*. CRC Press, 2nd edition, 1997.
- [7] A. Kale, A. K. R. Chowdhury, and R. Chellappa. Towards a view invariant gait recognition algorithm. In *Proc. on Advanced Video and Signal Based Surveillance*, pages 143–150, 2003.
- [8] X. Lan and D. P. Huttenlocher. A unified spatio-temporal articulated model for tracking. In *Proc. of CVPR*, volume 1, pages 722–729, 2004.
- [9] R. Pless and I. Simon. Using thousands of images of an object. In *Proc. of the Joint Conference on Information Science (CVPRIP)*, pages 684–687, 2002.
- [10] B. Scholkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.
- [11] G. Shakhnarovich, L. Lee, and T. Darrell. Integrated face and gait recognition from multiple views. In *Proc. of CVPR*, volume 1, pages 439–446, 2001.