Inferring 3D Body Pose from Silhouettes using Activity Manifold Learning

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

Abstract

We aim to infer 3D body pose directly from human silhouettes. Given a visual input (silhouette), the objective is to recover the intrinsic body configuration, recover the view point, reconstruct the input and detect any spatial or temporal outliers. In order to recover intrinsic body configuration (pose) from the visual input (silhouette), we explicitly learn view-based representations of activity manifolds as well as learn mapping functions between such central representations and both the visual input space and the 3D body pose space. The body pose can be recovered in a closed form in two steps by projecting the visual input to the learned representations of the activity manifold, i.e., finding the point on the learned manifold representation corresponding to the visual input, followed by interpolating 3D pose.

1. Introduction

Recovery of 3D body pose is a fundamental problem for human motion analysis in many applications such as motion capture, vision interface, visual surveillance, and gesture recognition. Human body is an articulated object that moves through the three-dimensional world. This motion is constrained by 3D body kinematics and dynamics as well as the dynamics of the activity being performed. Such constraints are explicitly exploited to recover the body configuration and motion in model-based approaches, such as [12, 10, 21, 20, 9, 14, 26], through explicitly specifying articulated models of the body parts, joint angles and their kinematics (or dynamics) as well as models for camera geometry and image formation. Recovering body configuration in these approaches involves searching high dimensional spaces (body configuration and geometric transformation) which is typically formulated deterministically as a nonlinear optimization problem, e.g. [20], or probabilistically as a maximum likelihood problem, e.g. [26]. Such approaches achieve significant success when the search problem is constrained as in a tracking context. However, initialization remains the most challenging problem which can be partially alleviated by sampling approaches. Partial recovery of body configuration can also be achieved through intermediate view-based representations (models) that may or may not be tied to specific body parts [6, 5, 31, 13, 27]. Alternatively, 3D body pose can be directly inferred from the visual input [11, 2, 23, 22, 16, 15, 25]. We call such approaches learning-based since their objective is to directly infer the 3D body pose as a function of the visual input. Such approaches have great potentials in solving the fundamental initialization problem for model-based vision.

The approach we present in this paper is inline with the learning-based approaches for pose recovery. In this paper we introduce a novel framework for inferring 3D body pose from silhouettes using a single monocular uncalibrated camera. The framework is based on explicitly learning view-based representations of the activity manifolds as well as learning mapping functions from such central representation to both the visual input space and the 3D body pose space. Given a visual input (silhouette) the body pose can be recovered in a closed form. The framework can simultaneously recover body configuration, the view point and reconstruct the input. We apply the framework for the gait as an example of a common human activity where we can successfully estimate the body pose for walking figures.

1.1. Related Work

In the last decade there have been extensive research in human motion analysis. We refer the reader to [8] for extensive survey of the broad subject. We focus our survey on related research on direct inference of 3D pose from images.

Inferring 3D pose from silhouettes can be achieved by learning mapping functions from the visual input to the pose space. However, learning such mapping between high dimensional spaces from examples is fundamentally an ill-posed problem. Therefore certain constraints have to be exploited. In [23, 22], learning specialized nonlinear mappings from Hu moment representation of the input shape and the pose space facilitated successful recovery of the pose directly from the visual input. In [2], the problem was constrained using nonlinear manifold learning where the pose is inferred by mapping sequences of the input to paths of the learned manifold. In [11] the reconstruction was based on 2D tracking of joints and a probabilistic model for human motion. In [15] 3D structure is inferred from

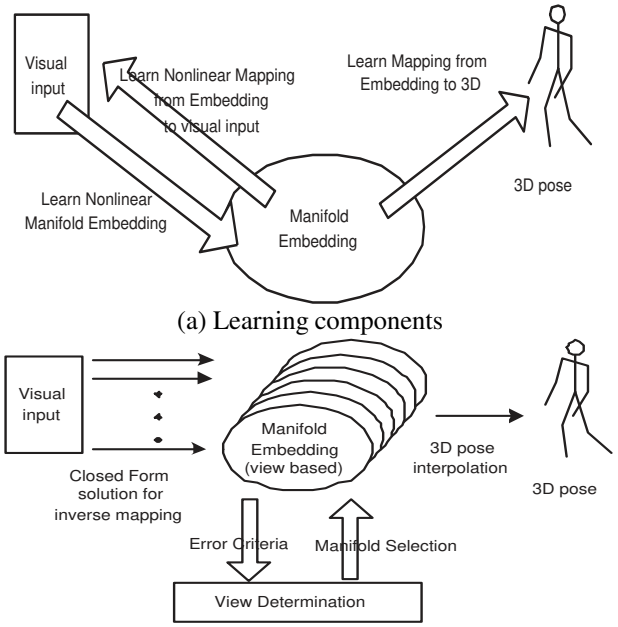

Figure 1. Block diagram for the framework. Top: Leaning components. Bottom: 3D pose estimation.

multi-view using a probabilistic model of multi-view silhouettes and key points on the object. Inferring pose can also be posed as a nearest neighbors search problem where the input is matched to a database of exemplars with known 3D pose. In [16] pose is recovered by matching the shape of the silhouette using shape context. In [25] locality sensitive hashing was used to efficiently match local models from the input to large exemplar sets.

The approach we use in this paper constrains the mapping problem through explicitly learning the activity manifold. Explicit manifold learning was previously used in [17] for modeling the appearance of rigid objects under changes in view points and illumination with linear PCA.

2. Framework

Given a visual input (silhouette), the objective is to recover the intrinsic body configuration, recover the view point, reconstruct the input and detect any spatial or temporal outliers. In other words, we aim to simultaneously solve for the pose, view point, and reconstruct the input.

If we consider the silhouette of a human performing certain activity or gesture, the shape of such silhouette deforms over time based on the activity performed. These deformations are constrained by the physical body constraints and the temporal constraints posed by the action being performed. For example, If we consider walking (gait), the human silhouettes through the walking cycle are points in a high dimensional visual input space. Given the spatial and the temporal constraints, it is expected that these points will lie on a low dimensional manifold. Intuitively, the gait is a 1-dimensional manifold which is embedded in a high dimensional visual space. Such manifold can be twisted, self-intersect in such high dimensional visual space. Similarly, if we consider other human activities such as gesturing, most of the gesture motion are also one-dimensional manifolds.

Given that such activity manifolds are low dimensional, the body pose can be recovered by projecting the visual input to a learned embedded representation of the activity manifold, i.e., finding the point on the learned manifold representation corresponding to the visual input. The questions are: how can we learn such representation of the manifold? and how to project from the input to such representation? The main challenge is that such manifolds are nonlinear since shape of the silhouette temporally undergoes deformations and self-occlusion which result in the points lying on a nonlinear, twisted manifold. Such nonlinear nature of the manifold makes the problem not obvious. Because of such nonlinearity, linear models such as PCA will not be able to discover the underlying manifold.

Learning nonlinear manifolds is typically performed in the visual input space or through intermediate representations. HMM models provide a probabilistic piecewise linear approximation of the manifold which can be used to learn nonlinear manifolds as in [4] and in [2]. Alternatively, Exemplar-based approaches such as [29] implicitly model nonlinear manifolds through points (exemplars) along the manifold. Such exemplars are represented in the visual input space. Recently some promising frameworks for nonlinear dimensionality reduction have been introduced including isometric feature mapping (Isomap) [28], Local linear embedding (LLE) [24]. Both Isomap and LLE frameworks were shown to be able to embed nonlinear manifolds into low-dimensional Euclidean spaces for toy examples as well as for real images. Recently, in [30], Isomap was used to enhance the tracking of parameterized contours within the Bayesian tracking framework. Related nonlinear dimensionality reduction work also includes [3].

In order to recover intrinsic body configuration (pose) from the visual input (silhouette) we explicitly learn view-based representations of the activity manifold as well as learn mapping functions between such representations and both the visual input space and the 3D body pose space. The framework is based on learning three components as shown in figure 1-a:

1. Learning Manifold Representation: using nonlinear dimensionality reduction we achieve an embedding of the global deformation manifold that preserves the geometric structure of the manifold. Given such embedding, the following two nonlinear mappings are learned.

2. Manifold-to-input mapping: a nonlinear mapping from the embedding space into visual input space.

3. Manifold-to-pose: a nonlinear mapping from the embedding space into the 3D body pose space.

We use Generalized Radial Basis Function (GRBF) interpolation framework for such nonlinear mapping. We show how approximate solution for the inverse mapping can be obtained by solving for the inverse of the manifold-to-input mapping in a closed form which facilitates the recovery of the intrinsic body configuration. Given a visual input (silhouette), its projections into view-based manifolds can be recovered in a closed form and therefore, the view can be determined using an embedding space error metric. Given the embedding coordinate, the 3D body pose can be directly interpolated using the learned manifold-to-pose mapping.

The following sections describe the details of the approach. Section 3 describes learning manifold representation. Section 4 describes learning nonlinear mapping from the manifold to the input and to the 3D pose space as well as the approach for pose recovery and view determination.

3. Learning Manifold Representation

3.1. Silhouette Representation

We use a global landmark-free correspondence-free representation of the visual input (silhouettes). There are two main motivations behind such representation: 1) Establishing correspondences between landmarks on the silhouettes is not not always feasible (has no meaning) because of the changes in topology over time (as in the gait case). Correspondences between landmarks (contours) are not always feasible because of self occlusion. 2) We aim to recover the pose from noisy and fragmented silhouettes. Landmark-based representations are typically sensitive to such effects.

We represent each shape instance as an implicit function $y(x)$ at each pixel x such that $y(x) = 0$ on the contour, $y(x) > 0$ inside the contour, and $y(x) < 0$ outside the contour. We use a signed-distance function such that

$$y(x) = \begin{cases} d_c(x) & x \text{ inside } c \\ 0 & x \text{ on } c \\ -d_c(x) & x \text{ outside } c \end{cases}$$

where the $d_c(x)$ is the distance to the closest point on the contour c with a positive sign inside the contour and a negative sign outside the contour. Such representation imposes smoothness on the distance between shapes. Given such representation, the input shapes are points $y_i \in R^d, i = 1, \cdots, N$ where all the input shapes are normalized and registered and d is the the dimensionality of the input space and N is the number of points. Implicit function representation is typically used in level-set methods [18].

3.2. Nonlinear Embedding

We adapt an LLE framework [24] to embed activity manifolds nonlinearly into a low dimensional space. Given the assumption that each data point and its neighbors lie on a locally linear patch of the manifold, each point can be reconstructed as a linear combinations of its local neighbors. The objective is to find the reconstruction weights that minimize the global reconstruction error. Optimal solution for such optimization problem can be found by solving a least-squares problem. Since the recovered weights reflect the intrinsic geometric structure of the manifold, an embedded manifold in a low dimensional space can be constructed using the same weights. This can be achieved by solving for a set of points, in a low dimensional space, that minimizes the reconstruction error where in this case the weights are fixed. Solving such problem can be achieved by solving an eigenvector problem. We refer the reader to [24] for details.

We applied the LLE to discover the geometric structure of the gait manifold as well as to establish a low dimensional embedding of such manifold. We used data sets of walking people from multiple views. Each data set consists of 300 frames and each containing about 8 to 11 walking cycles of the same person from certain view points[1]. In our case, the neighborhood of each point is determined by its K nearest neighbors based on the distance in the input space. One point that need to be emphasized is that we do not use the temporal relation to achieve the embedding, since the goal is to obtain an embedding that preserves the geometry of the manifold. Temporal relation can be used to determine the neighborhood of each shape but that was found to lead to erroneous artificial embedding. We also applied Isomap [28] framework on the same data to validate the results. Both Isomap and LLE resulted in qualitatively similar manifold embedding.

Figure 2 illustrates the resulting embedded manifold for a side view of the walker. Figure 3 illustrates the embedded manifolds for five different view points of the walker. For a given view point, the walking cycle evolves along a closed curve in the embedded space, i.e., only one degree of freedom controls the walking cycle which corresponds to the constrained body pose as a function of the time. Such conclusion is conforming with the intuition that the gait manifold is one dimensional. As mentioned earlier, temporal information was *not* used in the embedding. However, temporal information is used for visualization in figures 2 and 3. As apparent from the figures, embedding well preserves the temporal relation between input silhouettes.

Embedding Space Dimensionality: The question is: what is the least dimensional embedding space we can use to embed the walking cycle in a way that discriminate different poses through the whole cycle? The answer depends on the view point. The manifold twists in the embedding space given the different view points which impose different self occlusions. The least twisted manifold is the manifold for

[1] The data used are from the CMU Mobo gait data set which contains 25 people from six different view points. The walkers were using treadmill which might results in different dynamics from the natural walking

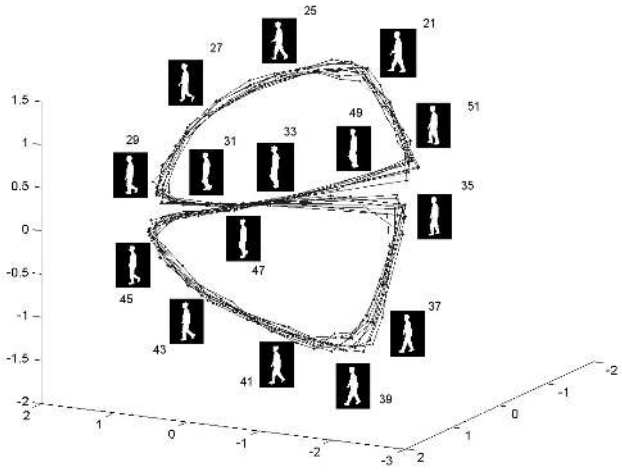

Figure 2. Embedded gait manifold for a side view of the walker. Sample frames from a walking cycle along the manifold with the frame numbers shown to indicate the order. Ten walking cycles are shown.

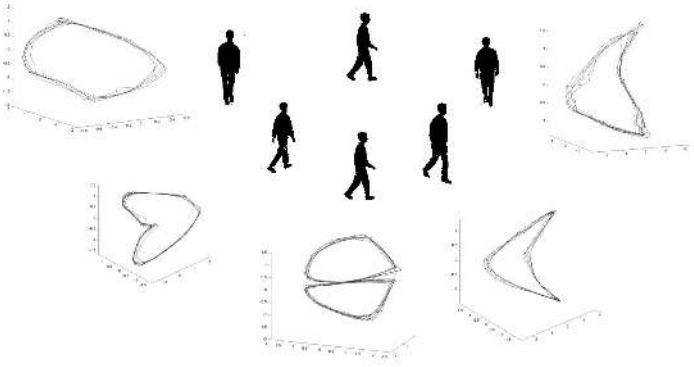

Figure 3. Embedded manifolds for 5 different views of the walkers. Frontal view manifold is the right most one and back view manifold is the leftmost one. We choose the view of the manifold that best illustrates its shape in the 3D embedding space

the back view as this is the least self occluding view (left most manifold in figure 3). In this case, the manifold can be embedded in a two dimensional space. For other views the curve starts to twist to be a three dimensional space curve. This is primarily because of the similarity imposed by the view point which attracts far away points on the manifold closer. The ultimate twist happens in the side view manifold where the curve twists and possibly self intersect to be a figure eight shape where each cycle of the eight (half eight) lies on a different plane. Each cycle of the eight figure corresponds to half a walking cycle. The closest point (cross point in case of intersection) represents the body pose where it is ambiguous, from the side view, to determine from the shape of the contour which leg is in front as can be noticed in Figure 2. Therefore, in a side view, three-dimensional embedding space is the least we can use to discriminate different poses. Embedding a side view cycle in a two-dimensional embedding space typically results in an embedding where the two half cycles lie over each other.

4. Learning Mapping

4.1. Learning Mapping: Manifold-to-Input

Given a visual input (silhouette), the objective is to recover the intrinsic body configuration by finding the point on the manifold in the embedding space corresponding to this input. Recovering such embedded representation will facilitate reconstruction of the input and detection of any spatial or temporal outliers.

Since the objective is to recover body configuration from the input, it might be obvious that we need to learn mapping from the input space, R^d, to the embedding space, R^e. However, learning such mapping is not feasible since the visual input is very high-dimensional so learning such mapping will require very large number of samples in order to be able to interpolate. *Instead,* we learn the mapping from the embedding space to the visual input space with a mechanism to directly solve for the inverse mapping.

It is well know that learning a smooth mapping from examples is an ill-posed problem unless the mapping is constrained since the mapping will be undefined in other parts of the space [19, 1]. We argue that, explicit modeling of the visual manifold represents a way to constrain any mapping between the visual input and any other space. Nonlinear embedding of the manifold, as was discussed in the previous section, represents a general framework to achieve this task. Constraining the mapping to the manifold is essential if we consider the existence of outliers (spatial and/or temporal) in the input space. This also facilitates learning mappings that can be used for interpolation between poses as we shall show. In what follows we explain our framework to recover the pose. In order to learn such nonlinear mapping we use Generalized Radial Basis Function (GRBF) interpolation framework [19]. Radial basis functions interpolation provides a framework for both implicitly modeling the embedded manifold as well as learning a mapping between the embedding space and the visual input space. In this case, the manifold is represented in the embedding space implicitly by selecting a set of representative points along the manifold.

Let the set of input instances (silhouettes) be $Y = \{y_i \in R^d \quad i = 1, \cdots, N\}$ and let their corresponding points in the embedding space be $X = \{x_i \in R^e, \quad i = 1, \cdots, N\}$ where e is the dimensionality of the embedding space (e.g. $e = 3$ in the case of gait). Let $\{t_j \in R^e, j = 1, \cdots, N_t\}$ be a set of N_t centers (not necessarily at data points) in the embedding space where such centers can be obtained using k-means clustering or EM algorithm. We can solve for multiple interpolants $f^k : R^e \rightarrow R$ where k is k-th dimension (pixel) in the input space and f^k is a radial basis

function interpolant, i.e., we learn nonlinear mappings from the embedding space to each individual pixel in the input space. Of particular interest are functions of the form

$$f^k(x) = p^k(x) + \sum_{i=1}^{N_t} w_i^k \phi(|x - t_i|), \quad (1)$$

that satisfies the interpolation condition

$$y_i^k = f^k(x_i)$$

where $\phi(\cdot)$ is a real-valued basic function, w_i are real coefficients, $|\cdot|$ is the norm on R^e (the embedding space). Typical choices for the basis function includes thin-plate spline ($\phi(u) = u^2 log(u)$), the multiquadric ($\phi(u) = \sqrt{u^2 + c^2}$), Gaussian ($\phi(u) = e^{-cu^2}$), biharmonic ($\phi(u) = u$) and triharmonic ($\phi(u) = u^3$) splines. $p^k(x)$ is a linear polynomial with coefficients c^k, i.e., $p^k(x) = [1 \quad x^\top] \cdot c^k$. This linear polynomial is essential to achieve approximate solution for the inverse mapping as will be shown.

The whole mapping can be written in a matrix form as

$$f(x) = B \cdot \psi(x), \quad (2)$$

where B is a $d \times (N_t + e + 1)$ dimensional matrix with the k-th row $[w_1^k \cdots w_{N_t}^k \quad c^{k^T}]$ and the vector $\psi(x)$ is

$$[\phi(|x - t_1|) \cdots \phi(|x - t_{N_t}|) \quad 1 \quad x^\top]^\top.$$

The matrix B represents the coefficients for d different nonlinear mappings, each from a low-dimension embedding space into real numbers. To insure orthogonality and to make the problem well posed, the following additional constraints are imposed

$$\sum_{i=1}^{N} w_i p_j(x_i) = 0, j = 1, \cdots, m$$

where p_j are the linear basis of p. Therefore the solution for B can be obtained by directly solving the linear systems

$$\begin{pmatrix} A & P_x \\ P_t^\top & 0_{(e+1)\times(e+1)} \end{pmatrix} B^\top = \begin{pmatrix} Y \\ 0_{(e+1)\times d} \end{pmatrix}, \quad (3)$$

where A is $N \times N_t$ matrix with $A_{ij} = \phi(|x_i - t_j|), \quad i = 1 \cdots N, j = 1 \cdots N_t$, P_x is a $N \times (e + 1)$ matrix with *i*-th row $[1 \quad x_i^\top]$, P_t is a $N_t \times (e+1)$ matrix with *i*-th row $[1 \quad t_i^\top]$. Y is $(N \times d)$ matrix containing the representative input images, i.e., $Y = [y_1 \cdots y_N]^\top$. Solution for B is guaranteed under certain conditions on the basic functions used [19].

Given such mapping, any input is represented by linear combination of nonlinear functions centered in the embedding space along the manifold. Equivalently, this can be interpreted as a form or basis images (coefficients) that are combined nonlinearly using kernel functions centered along the embedded manifold.

4.2. Solving For the Embedding Coordinates

Given a new input $y \in R^d$, it is required to find the corresponding embedding coordinates $x \in R^e$ by solving for the inverse mapping. There are two questions that we need to answer:

1. What is the coordinates of point $x^* \in R^e$ in the embedding space corresponding to such input?
2. What is the closest manifold point corresponding to such input?

To answer the first question we need to obtain a solution for

$$x^* = arg_x min \|y - B\psi(x)\|^2. \quad (4)$$

Each input yields a set of d nonlinear equations in e unknowns (or d nonlinear equations in one *e*-dimensional unknown). Therefore a solution for x^* can be obtained by least square solution for the over-constrained nonlinear system in 4. However, because of the linear polynomial part in the interpolation function, the vector $\psi(x)$ has a special form that facilitates a closed-form least square linear approximation and, therefore, avoid solving the nonlinear system. This can be achieved by obtaining the pseudo-inverse of B. Note that B has rank N since N distinctive RBF centers are used. Therefore, the pseudo-inverse can be obtained by decomposing B using SVD such that $B = USV^\top$ which can be performed offline. Therefore, vector $\psi(x)$ can be recovered simply as

$$\psi(x) = V\acute{S}U^T y$$

where $\acute{S}$ is the diagonal matrix obtained by taking the inverse of the nonzero singular values in the diagonal matrix S and setting the rest to zeros. Linear approximation for the embedding coordinate x can be obtained directly by taking the last e rows in the recovered vector $\psi(x)$.

The recovered point x is typically enough to recover the pose. However to enhance the result and constrain the solution, we need to answer the second question above, which can also be obtained efficiently. We need to find the point on the manifold closest to the projection x^*. For the gait case, the manifold is one dimensional, and therefore, only one dimensional search is sufficient to recover the manifold point closest to the input. To obtain such point, the embedded manifold is fitted with a cubic spline $m(t)$ as a function of the time variable $t \in [0, 1]$ where each cycle of the activity is temporally mapped from 0 to 1. Given such model, a one dimensional search is used to obtain t^* that minimizes $\|x - m(t)\|$. Reconstruction can be achieved by re-mapping the projected point using 2.

4.3. Determining View Point

Given the learned view-based manifolds $\mathcal{M}_v$ and the learned view-based mappings $B_v\psi_v(x)$ for each view v, determining the view point reduces to finding the manifold

that minimizes the inverse-mapping error of an input y or a sequence of inputs y_t. Given an input y and its projections x_v^* we chose the manifold that minimizes $\|x_v^* - m_v(t_v*)\|$. Figure 4 shows five view manifolds and the projection of a sequence to the five manifolds.

4.4. Learning Mapping: Manifold-to-3D

Similar to the mapping from the embedding space into the visual input, a mapping can be learned from the embedding space to the 3D body joint space. RBF interpolants in the form of equation 1 between the embedding space R^e and each degree of freedom of each body joint. We represent the body using 16 joints model and each joint is represented by its coordinates in a body centered global coordinate system. Representative points on the manifolds as well as their corresponding 3D body configurations are used in order to learn the mapping parameters as was shown in section 4.

4.5. Learning Multiple People Manifolds

The approach we described in this paper can be generalized to learn manifolds and mappings from multiple people data. However certain fundamental issues have to be addressed:

- How to learn unified representation of a certain activity manifolds from multiple people sequences.
- How to learn style-based nonlinear mappings from the unified manifold representation to each person silhouettes.
- Given an input, how to solve for both the person and the pose in the embedding space.

In [7] we presented a general framework to learn multiple people manifolds and to separate the content (body configuration) as time-dependent function from time-invariant style (person) parameters. This framework generalizes the learning procedure introduced in this paper.

5. Experimental Result

In the first experiment, we used a sequence from Georgia tech gait data with ground truth provided by motion capture data. the sequence contains 72 frames where we learn the model using the odd numbered frames and evaluated on the even numbered frames. The resulted 3D reconstruction is compared to the ground truth and is plotted in figure 9 for four of the sixteen joint angles. This experiment validates that our approach can interpolate 3D poses from unseen input silhouettes.

In order to show that the approach generalizes to different people, we used the CMU MoboGait database to train and evaluate the proposed approach. Each sequence of the database contains about 300 frames (8-11 walking cycles). The database contains 6 views of each walking person. We used five of them. The used views are shown in figure 4.

In each experiment, we used one person sequences to learn the manifolds of the five views and the mappings from the manifolds to the input sequences. The mappings from each of the manifolds to 3D body configuration were also learned. For the evaluation we use other people's sequences to evaluate the 3D reconstruction [2]. Figure 5 shows the view classification results for five evaluation sequences (five people) and five views. Overall correct classification rate is 93.05%. Obviously the view classification from a single frame can be erroneous because of self occlusion and therefore boosting several frames would lead to better results which is shown in figure 5-b where majority vote were used over sequence of each five frame view classification which results in a correct classification rate of 99.63%.

Figure 5 shows the 3D reconstruction for one person for each of the five views. Since the input sequences are synchronized, the reconstructed 3D poses from each view are supposed to be the same. The 3D reconstructions are always shown from the side view point. The reconstruction shows qualitatively correct reconstruction from all views. Unfortunately, there are no ground truth to evaluate the results of this experiment. Figure 8 show some 3D reconstruction results for four other people. As can be noticed, the input silhouettes are noisy.

Figure 7 shows 3D pose reconstructed from corrupted silhouette which are typical in surveillance applications due to errors in background subtraction, shadows, fragmentation, and carried objects. Reconstruction of the input silhouettes can be achieved by mapping back to the input space. Results related to input reconstruction were reported in [7]

6. Conclusion

In this paper we introduced a learning-based framework for inferring 3D body pose from silhouettes using a single monocular uncalibrated camera. The framework is based on explicitly learning view-based representations of the activity manifolds as well as learning mapping functions from such central representation to both the visual input space and the 3D body pose space. Given a visual input (silhouette) the body pose can be recovered in a closed form. We applied the framework for the gait as an example of a common human activity where we can successfully estimate the body pose for walking figures. The experiments showed that the model can be learned from one person data and successfully generalizes to recovering poses for other people from noisy data. Compared to previous approaches for inferring 3D body pose from visual input, we can point out certain advantageous and limitations. Our framework facilitates interpolation of intermediate 3D poses even if they are

[2] For the experiment we show here we use person 37 for the learning and evaluate on persons 15 in figure 5 and on 70, 86, 76, 79 in figure 8

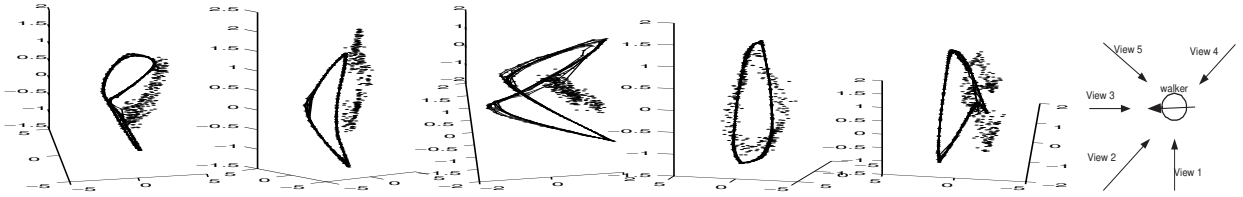

Figure 4. Five Manifolds for five view points and the projection of a sequences to each manifold.

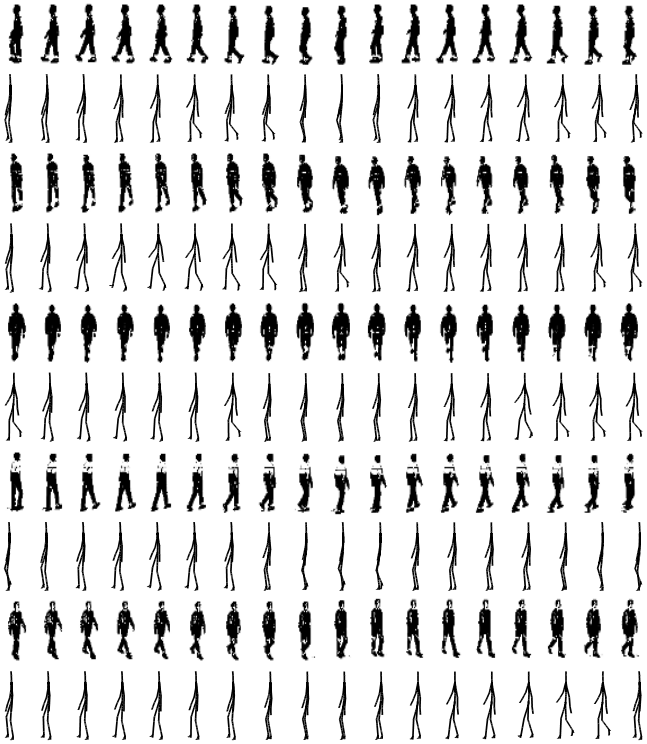

Figure 5. 3D reconstruction for five views.

not part of the training data. Unlike [23, 2], where mapping is learned directly between the input and pose space, our framework constrains the mapping to the learned manifold which facilitates robust pose recovery from noisy and corrupted inputs as well as for reconstruction of the input. Unlike [2], where sequences of the input is required to recover the pose, the framework facilitates recovery of the pose from single input instances as well as from sequences of input. Similar to [2], our approach is based on learning activity manifold and therefore its application is limited to

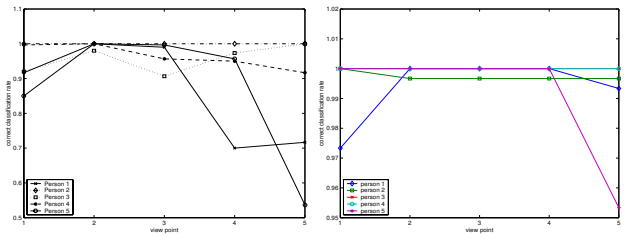

Figure 6. a- view classification from single frames b- view classification with boosting multiple frames

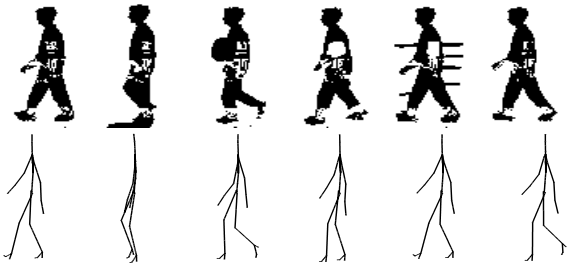

Figure 7. 3D reconstruction from corrupted inputs

recovery of poses for the learned activities only. In this paper we focus on the gait case. However the framework is general and can be applied to other activities by learning their manifolds.

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

References

[1] D. Beymer and T. Poggio. Image representations for visual learning. *Science*, 272(5250), 1996.

[2] M. Brand. Shadow puppetry. In *International Conference on Computer Vision*, volume 2, page 1237, 1999.

[3] M. Brand and K. Huang. A unifying theorem for spectral embedding and clustering. In *Proc. of the Ninth International Workshop on AI and Statistics*, 2003.

[4] C. Bregler and S. M. Omohundro. Nonlinear manifold learning for visual speech recognition. In *ICCV*, 1995.

[5] L. W. Campbell and A. F. Bobick. Recognition of human body motion using phase space constraints. In *ICCV*, pages 624–630, 1995.

[6] T. Darrell and A. Pentland. Space-time gesture. In *Proc IEEE CVPR*, 1993.

[7] A. Elgammal and C.-S. Lee. Separating style and content on a nonlinear manifold. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, June-July 2004.

[8] D. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, Jan 1999.

[9] D. Gavrila and L. Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1996.

[10] D. Hogg. Model-based vision: a program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.

[11] Howe, Leventon, and W. Freeman. Bayesian reconstruction of 3d human motion from single-camera video. In *Proc. NIPS*, 1999.

[12] J.O'Rourke and Badler. Model-based image analysis of human motion using constraint propagation. *IEEE PAMI*, 2(6), 1980.

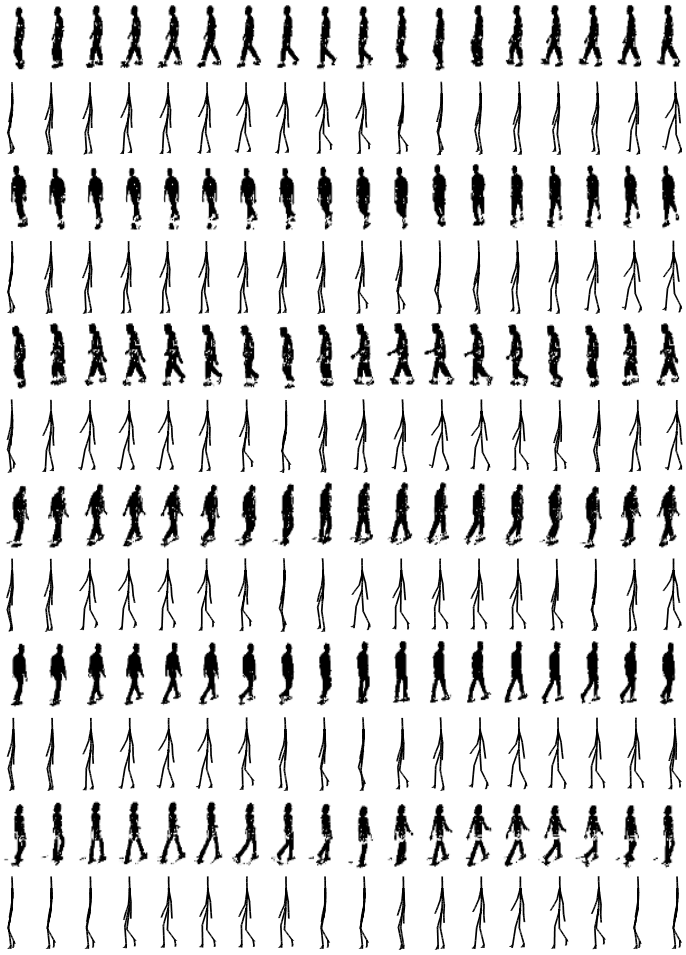

Figure 8. 3D reconstruction for 4 people from different views: From top to bottom: person 70 views 1,2; person 86 views 1,2; person 76 view 4; person 79 view 4

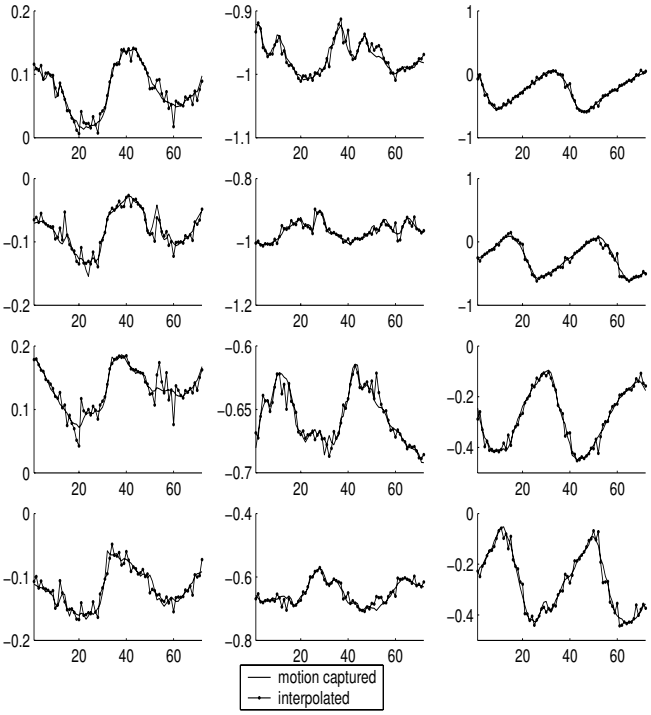

Figure 9. Evaluation of 3D reconstruction with ground truth for four joints (right foot, left foot, Lower right leg, lower left leg). Each row represents a joint angle x,y,z. (units in foot)

[13] S. X. Ju, M. J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. In *International Conference on Automatic Face and Gesture Recognition*, pages 38–44, Killington, Vermont, 1996.

[14] I. A. Kakadiaris and D. Metaxas. Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition, CVPR*, pages 81–87, Los Alamitos, California, U.S.A., 18–20 1996. IEEE Computer Society.

[15] T. D. Kristen Grauman, Gregory Shakhnarovich. Inferring 3d structure with a statistical image-based shape model. In *ICCV*, 2003.

[16] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *European Conference on Computer Vision*, 2002.

[17] S. Nayar, H. Murase, and S. Nene. Parametric appearance representation. In *Early Visual Learning*. Oxford University Press, February 1996.

[18] S. Osher and N. Paragios. *Geometric Level Set Methods*. Springer, 2003.

[19] T. Poggio and F. Girosi. Network for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.

[20] J. M. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *ICCV*, pages 612–617, 1995.

[21] K. Rohr. Towards model-based recognition of human movements in image sequence. *CVGIP*, 59(1):94–115, 1994.

[22] R. Rosales, V. Athitsos, and S. Sclaroff. 3D hand pose reconstruction using specialized mappings. In *Proc. ICCV*, 2001.

[23] R. Rosales and S. Sclaroff. Specialized mappings and the estimation of human body pose from a single image. In *Workshop on Human Motion*, pages 19–24, 2000.

[24] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Sciene*, 290(5500):2323–2326, 2000.

[25] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *ICCV*, 2003.

[26] 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.

[27] H. Sidenbladh, M. J. Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *Proc. ECCV 2002, LNCS 2350*, pages 784–800, 2002.

[28] J. Tenenbaum. Mapping a manifold of perceptual observations. In *Advances in Neural Information Processing*, volume 10, pages 682–688, 1998.

[29] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *ICCV*, pages 50–59, 2001.

[30] Q. Wang, G. Xu, and H. Ai. Learning object interinsic structure for robust visual tracking. In *CVPR*, volume 2, page 227, 2003.

[31] C. R. Wern, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfinder: Real-time tracking of human body. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 1997.