The Role of Manifold Learning in Human Motion Analysis

Ahmed Elgammal and Chan Su Lee

Department of Computer Science,
Rutgers University, Piscataway, NJ, USA
{elgammal,chansu}@cs.rutgers.edu

Abstract. Human body is an articulated object with high degrees of freedom. Despite the high dimensionality of the configuration space, many human motion activities lie intrinsically on low dimensional manifolds. Although the intrinsic body configuration manifolds might be very low in dimensionality, the resulting appearance manifolds are challenging to model given various aspects that affects the appearance such as the shape and appearance of the person performing the motion, or variation in the view point, or illumination. Our objective is to learn representations for the shape and the appearance of moving (dynamic) objects that support tasks such as synthesis, pose recovery, reconstruction, and tracking. We studied various approaches for representing global deformation manifolds that preserve their geometric structure. Given such representations, we can learn generative models for dynamic shape and appearance. We also address the fundamental question of separating style and content on nonlinear manifolds representing dynamic objects. We learn factorized generative models that explicitly decompose the intrinsic body configuration (content) as a function of time from the appearance/shape (style factors) of the person performing the action as time-invariant parameters. We show results on pose recovery, body tracking, gait recognition, as well as facial expression tracking and recognition.

1 Introduction

Human body is an articulated object with high degrees of freedom. Human body moves through the three-dimensional world and such motion is constrained by body dynamics and projected by lenses to form the visual input we capture through our cameras. Therefore, the changes (deformation) in appearance (texture, contours, edges, etc.) in the visual input (image sequences) corresponding to performing certain actions, such as facial expression or gesturing, are well constrained by the 3D body structure and the dynamics of the action being performed. Such constraints are explicitly exploited to recover the body configuration and motion in model-based approaches [32, 28, 13, 64, 62, 23, 34, 72] 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. [61, 62], or probabilistically as a maximum likelihood problem, e.g. [72]. Such approaches achieve significant success when the search problem is constrained as in tracking context. However, initialization remains the most challenging problem, which can be partially alleviated by sampling approaches. The dimensionality of the initialization problem increases as we incorporate models for variations between individuals in physical body style, models for variations in action style, or models for clothing, etc. 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 [18, 12, 86, 33, 6, 27, 87, 22, 73, 24]. In such case constancy of the local appearance of individual body parts is exploited. Alternative paradigms are appearance-based and motion-based approaches where the focus is to track and recognize human activities without full recovery of the 3D body pose [58, 54, 57, 59, 55, 74, 63, 7, 17].

Recently, there have been research for recovering body posture directly from the visual input by posing the problem as a learning problem through searching a prelabelled database of body posture [51, 36, 70] or through learning regression models from input to output [29, 9, 66, 67, 65, 14, 60]. All these approaches pose the problem as a machine learning problem where the objective is to learn input-output mapping from input-output pairs of training data. Such approaches have great potential for solving the initialization problem for model-based vision. However, these approaches are challenged by the existence of wide range of variability in the input domain.

Role of Manifold:

Despite the high dimensionality of the configuration space, many human motion activities lie intrinsically on low dimensional manifolds. This is true if we consider the body kinematics as well as if we consider the observed motion through image sequences. Let us consider the observed motion. For example, the shape of the human silhouette walking or performing a gesture is an example of a dynamic shape where the shape deforms over time based on the action performed. These deformations are constrained by the physical body constraints and the temporal constraints posed by the action being performed. If we consider these silhouettes through the walking cycle as points in a high dimensional visual input space, then, given the spatial and the temporal constraints, it is expected that these points will lay on a low dimensional manifold. Intuitively, the gait is a 1-dimensional manifold which is embedded in a high dimensional visual space. This was also shown in [8]. Such manifold can be twisted, self-intersect in such high dimensional visual space.

Similarly, the appearance of a face performing facial expressions is an example of dynamic appearance that lies on a low dimensional manifold in the visual input space. In fact if we consider certain classes of motion such as gait, or a single gesture, or a single facial expressions and if we factor out all other sources of variability, each of such motions lies on a one-dimensional manifolds, i.e., a trajectory in the visual input space. Such manifolds are nonlinear and non-Euclidean.

Therefore, researchers have tried to exploit the manifold structure as a constraint in tasks such as tracking and activity recognition in an implicit way. Learning nonlinear deformation manifolds is typically performed in the visual input space or through intermediate representations. For example, Exemplar-based approaches such as [77] implicitly model nonlinear manifolds through points (exemplars) along the manifold. Such exemplars are represented in the visual input space. HMM models provide a probabilistic piecewise linear approximation which can be used to learn nonlinear manifolds as in [11] and in [9].

Although the intrinsic body configuration manifolds might be very low in dimensionality, the resulting appearance manifolds are challenging to model given various aspects that affect the appearance such as the shape and appearance of the person performing the motion, or variation in the view point, or illumination. Such variability makes the task of learning visual manifold very challenging because we are dealing with data points that lies on multiple manifolds on the same time: body configuration manifold, view manifold, shape manifold, illumination manifold, etc.

Linear, Bilinear and Multi-linear Models:

Can we decompose the configuration using linear models? Linear models, such as PCA [31], have been widely used in appearance modeling to discover subspaces for variations. For example, PCA has been used extensively for face recognition such as in [52, 1, 15, 47] and to model the appearance and view manifolds for 3D object recognition as in [53]. Such subspace analysis can be further extended to decompose multiple orthogonal factors using bilinear models and multi-linear tensor analysis [76, 80]. The pioneering work of Tenenbaum and Freeman [76] formulated the separation of style and content using a bilinear model framework [48]. In that work, a bilinear model was used to decompose face appearance into two factors: head pose and different people as style and content interchangeably. They presented a computational framework for model fitting using SVD. Bilinear models have been used earlier in other contexts [48, 49]. In [80] multi-linear tensor analysis was used to decompose face images into orthogonal factors controlling the appearance of the face, including geometry (people), expressions, head pose, and illumination. They employed high order singular value decomposition (HOSVD) [37] to fit multi-linear models. Tensor representation of image data was used in [71] for video compression and in [79, 84] for motion analysis and synthesis. N-mode analysis of higher-order tensors was originally proposed and developed in [78, 35, 48] and others. Another extension is algebraic solution for subspace clustering through generalized-PCA [83, 82]

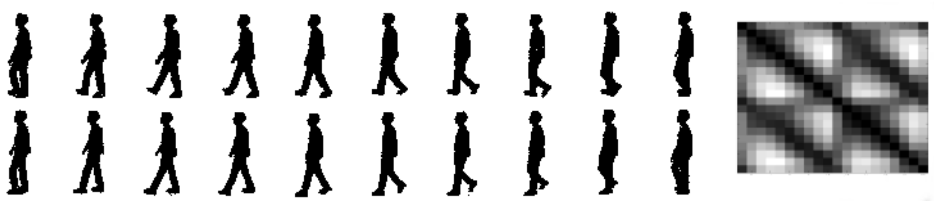


Fig. 1. Twenty sample frames from a walking cycle from a side view. Each row represents half a cycle. Notice the similarity between the two half cycles. The right part shows the similarity matrix: each row and column corresponds to one sample. Darker means closer distance and brighter means larger distances. The two dark lines parallel to the diagonal show the similarity between the two half cycles

In our case, the object is dynamic. So, can we decompose the configuration from the shape (appearance) using linear embedding? For our case, the shape temporally undergoes deformations and self-occlusion which result in the points lying on a nonlinear, twisted manifold. This can be illustrated if we consider the walking cycle in Figure 1. The two shapes in the middle of the two rows correspond to the farthest points in the walking cycle kinematically and are supposedly the farthest points on the manifold in terms of the geodesic distance along the manifold. In the Euclidean visual input space these two points are very close to each other as can be noticed from the distance plot on the right of Figure 1. Because of such nonlinearity, PCA will not be able to discover the underlying manifold. Simply, linear models will not be able to interpolate intermediate poses. For the same reason, multidimensional scaling (MDS) [16] also fails to recover such manifold.

Nonlinear Dimensionality Reduction and Decomposition of Orthogonal Factors:

Recently some promising frameworks for nonlinear dimensionality reduction have been introduced, e.g. [75, 68, 2, 10, 38, 85, 50]. Such approaches can achieve embedding of nonlinear manifolds through changing the metric from the original space to the embedding space based on local structure of the manifold. While there are various such approaches, they mainly fall into two categories: Spectral-embedding approaches and Statistical approaches. Spectral embedding includes approaches such as isometric feature mapping (Isomap) [75], Local linear embedding (LLE) [68], Laplacian eigenmaps [2], and Manifold Charting [10]. Spectral-embedding approaches, in general, construct an affinity matrix between data points using data dependent kernels, which reflect local manifold structure. Embedding is then achieved through solving an eigen-value problem on such matrix. It was shown in [3, 26] that these approaches are all instances of kernel-based learning, in particular kernel principle component analysis KPCA [69]. In [4] an approach for embedding out-of-sample points to complement such approaches. Along the same line, our work [19, 21] introduced a general framework for mapping between input and embedding spaces.

All these nonlinear embedding frameworks were shown to be able to embed nonlinear manifolds into low-dimensional Euclidean spaces for toy examples as well as for real images. Such approaches are able to embed image ensembles nonlinearly into low dimensional spaces where various orthogonal perceptual aspects can be shown to correspond to certain directions or clusters in the embedding spaces. In this sense, such nonlinear dimensionality reduction frameworks present an alternative solution to the decomposition problems. However, the application of such approaches is limited to embedding of a single manifold.

Biological Motivation:

While the role of manifold representations is still unclear in perception, it is clear that images of the same objects lie on a low dimensional manifold in the visual space defined by the retinal array. On the other hand, neurophysiologist have found that neural population activity firing is typically a function of small number of variables which implies that population activity also lie on low dimensional manifolds [30].

2 Learning Simple Motion Manifold

2.1 Case Study: The Gait Manifold

In order to achieve a low dimensional embedding of the gait manifold, nonlinear dimensionality reduction techniques such as LLE [68], Isomap [75], and others can be used. Most these techniques result in qualitatively similar manifold embedding. As a result of nonlinear dimensionality reduction we can reach an embedding of the gait manifold in a low dimension Euclidean space [19]. Figure 2 illustrates the resulting embedded manifold for a side view of the walker [1]. 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.

One important 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 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 to be a figure eight shape where each cycle of the eight (half eight) lies in a different plane. Each half of the “eight” figure corresponds to half a walking cycle. The cross point represents the body pose where it is totally 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 results in an embedding similar to that shown in top left of Figure 2 where the two half cycles lies over each other. Different people are expected to have different manifolds. However, such manifolds are all topologically equivalent. This can be noticed in Figure 8-c. Such property will be exploited later in the chapter to learn unified representations from multiple manifolds.

2.2 Learning the Visual Manifold: Generative Model

Given that we can achieve a low dimensional embedding of the visual manifold of dynamic shape data, such as the gait data shown above, the question is how to use this embedding to learn representations of moving (dynamic) objects that supports tasks

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

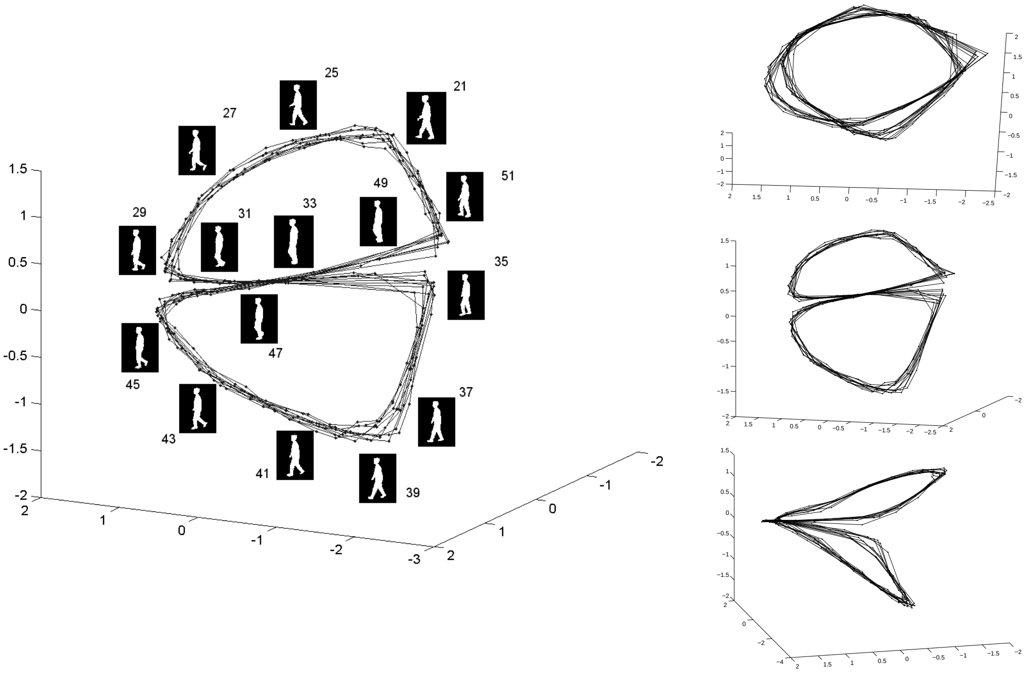


Fig. 2. Embedded gait manifold for a side view of the walker. Left: sample frames from a walking cycle along the manifold with the frame numbers shown to indicate the order. Ten walking cycles are shown. Right: three different views of the manifold.

such as synthesis, pose recovery, reconstruction and tracking. In the simplest form, assuming no other source of variability besides the intrinsic motion, we can think of a view-based generative model of the form

$$y_t = T_\alpha \gamma(x_t; a) \tag{1}$$

where the shape (appearance), y_t, at time t is an instance driven from a generative model where the function γ is a mapping function that maps body configuration x_t at time t into the image space. The body configuration x_t is constrained to the explicitly modeled motion manifold. i.e., the mapping function γ maps from a representation of the body configuration space into the image space given mapping parameters a that are independent from the configuration. T_α represents a global geometric transformation on the appearance instance.

The manifold in the embedding space can be modeled explicitly in a function form or implicitly by points along the embedded manifold (embedded exemplars). The embedded manifold can be also modelled probabilistically using Hidden Markov Models and EM. Clearly, learning manifold representations in a low-dimensional embedding space is advantageous over learning them in the visual input space. However, our em-

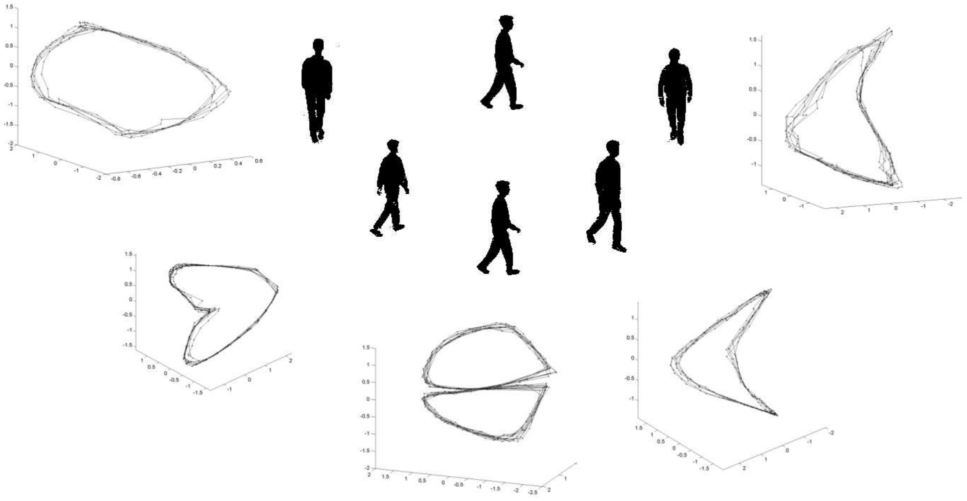


Fig. 3. Embedded manifolds for 5 different views of the walkers. Frontal view manifold is the right most one and back view manifold is the leftmost one. We choose the view of the manifold that best illustrates its shape in the 3D embedding space

phasize is on learning the mapping between the embedding space and the visual input space.

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 to the embedding space, i.e., mapping from R^d to R^e. However, learning such mapping is not feasible since the visual input is very high-dimensional so learning such mapping will require 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, i.e., in a generative manner, with a mechanism to directly solve for the inverse mapping. Another fundamental reason to learn the mapping in this direction is the inherent ambiguity in 2D data. Therefore, mapping from visual data to the manifold representation is not necessary a function. While learning a mapping from the manifold to the visual data is a function.

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 [56]. 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 Radial basis function (RBF) interpolation framework. The use of RBF for image synthesis and analysis has been pioneered by [56, 5] where RBF networks were used to learn nonlinear mappings between image space and a supervised parameter space. In our work we use RBF interpolation framework in a novel way to learn mapping from unsupervised learned parameter space to the input space. 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 as the centers for the basis functions.

Let the set of representative input instances (shape or appearance) be $\mathsf{Y} = \{y_i \in R^d \quad i = 1, \cdots, N\}$ and let their corresponding points in the embedding space be $\mathsf{X} = \{x_i \in R^e, \quad i = 1, \cdots, N\}$ where e is the dimensionality of the embedding space (e.g. $e = 3$ in the case of gait). 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

$$f^k(x) = p^k(x) + \sum_{i=1}^{N} w_i^k \phi(|x - x_i|), \tag{2}$$

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 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

$$f(x) = B \cdot \psi(x), \tag{3}$$

where B is a $d \times (N{+}e{+}1)$ dimensional matrix with the k-th row $[w_1^k \cdots w_N^k \quad c^{k^T}]$ and the vector $\psi(x)$ is $[\phi(|x - x_1|) \cdots \phi(|x - x_N|) \quad 1 \quad x^\top]^\top$. 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

$$\sum_{i=1}^{N} w_i p_j(x_i) = 0, j = 1, \cdots, m \tag{4}$$

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

$$\begin{pmatrix} A & P \\ P^\top & 0 \end{pmatrix} B^\top = \begin{pmatrix} Y \\ 0_{(e+1)\times d} \end{pmatrix}, \tag{5}$$

where $A_{ij} = \phi(|x_j - x_i|), \quad i, j = 1 \cdots N$, P is a matrix with i-th row $[1 \quad x_i^\top]$, and 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. Similarly, mapping can be learned using arbitrary centers in the embedding space (not necessarily at data points) [56, 19].

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

2.3 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 might need to answer

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

In both cases we need to obtain a solution for

$$x^* = \underset{x}{\operatorname{argmin}} ||y - B\psi(x)|| \tag{6}$$

where for the second question the answer is constrained to be on the embedded manifold. In the cases where the manifold is only one dimensional, (for example in the gait case, as will be shown) only one dimensional search is sufficient to recover the manifold point closest to the input. However, we show here how to obtain a closed-form solution for x^*.

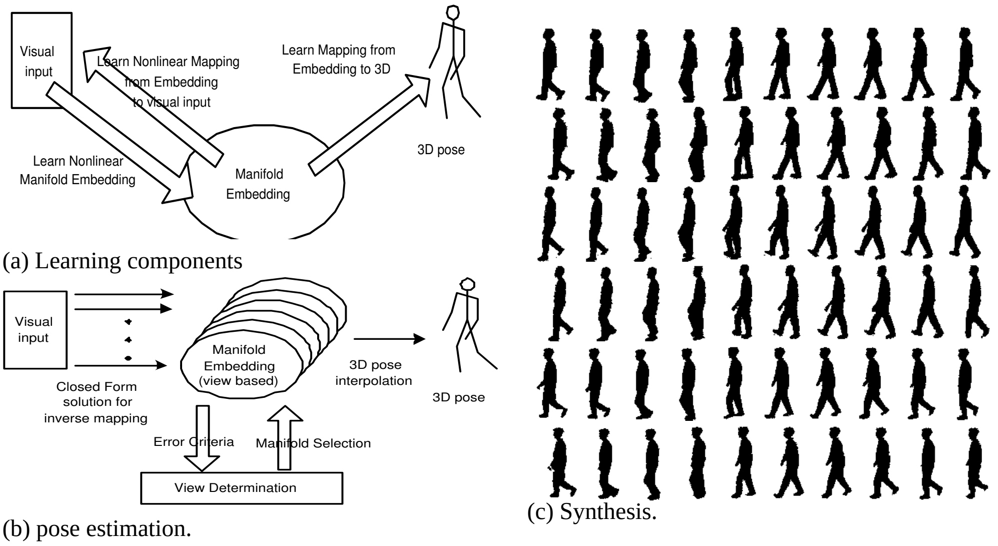
Each input yields a set of d nonlinear equations in e unknowns (or d nonlinear equations in one e-dimensional unknown). Therefore a solution for x^* can be obtained by least square solution for the over-constrained nonlinear system in 6. However, because of the linear polynomial part in the interpolation 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 $B = USV^\top$ and, therefore, vector $\psi(x)$ can be recovered simply as

$$\psi(x) = V\tilde{S}U^T y \tag{7}$$

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

2.4 Synthesis, Recovery and Reconstruction:

Given the learned model, we can synthesis new shapes along the manifold. Figure 4-c shows an example of shape synthesis and interpolation. Given a learned generative model in the form of Equation 3, we can synthesize new shapes through the walking cycle. In these examples only 10 samples were used to embed the manifold for half a cycle on a unit circle in 2D and to learn the model. Silhouettes at intermediate body configurations were synthesized (at the middle point between each two centers) using the learned model. The learned model can successfully interpolate shapes at intermediate configurations (never seen in the learning) using only two-dimensional embedding. The figure shows results for three different peoples.


Visual input
Learn Nonlinear Mapping from Embedding to visual input
Learn Mapping from Embedding to 3D
Learn Nonlinear Manifold Embedding
Manifold Embedding
3D pose
Visual input
Manifold Embedding (view based)
3D pose interpolation
3D pose
Closed Form solution for inverse mapping
Error Criteria
Manifold Selection
View Determination

(a) Learning components

(b) pose estimation.

(c) Synthesis.

Fig. 4. (a,b) Block diagram for the learning framework and 3D pose estimation. (c) Shape synthesis for three different people. First, third and fifth rows: samples used in learning. Second, fourth, sixth rows: interpolated shapes at intermediate configurations (never seen in the learning)

Given a visual input (silhouette), and the learned model, we can recover the intrinsic body configuration, recover the view point, and reconstruct the input and detect any spatial or temporal outliers. In other words, we can simultaneously solve for the pose, view point, and reconstruct the input. A block diagram for recovering 3D pose and view point given learned manifold models are shown in Figure 4. The framework [20] is based on learning three components as shown in Figure 4-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 as described in section 2.1. 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 as described in section 2.2.
3. Manifold-to-pose: a nonlinear mapping from the embedding space into the 3D body pose space.

Given an input shape, the embedding coordinate, i.e., the body configuration can be recovered in closed-form as was shown in section 2.3. Therefore, the model can be used for pose recovery as well as reconstruction of noisy inputs. Figure 5 shows examples of the reconstruction given corrupted silhouettes as input. In this example, the manifold representation and the mapping were learned from one person data and tested on other people date. Given a corrupted input, after solving for the global geometric transformation, the input is projected to the embedding space using the closed-form inverse mapping approximation in section 2.3. The nearest embedded manifold point represents the intrinsic body configuration. A reconstruction of the input can achieved by projecting back to the input space using the direct mapping in Equation 3. As can be noticed from the figure, the reconstructed silhouettes preserve the correct body pose in each case which shows that solving for the inverse mapping yields correct points on the manifold. Notice that no mapping is learned from the input space to the embedded space. Figure 6 shows examples of 3D pose recovery obtained in closed-form for different people from different view. The training has be done using only one subject data from five view points. All the results in Figure 6 are for subjects not used in the training. This shows that the model generalized very well.

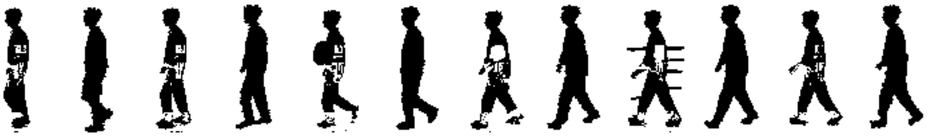


Fig. 5. Example pose-preserving reconstruction results. Six noisy and corrupted silhouettes and their reconstructions next to them.

3 Adding more Variability: Factoring out the Style

The generative model introduced in Equation 1 generates the visual input as a function of a latent variable representing body configuration constrained to a motion manifold. Obviously body configuration is not the only factor controlling the visual appearance of humans in images. Any input image is a function of many aspects such as person body structure, appearance, view point, illumination, etc. Therefore, it is obvious that the visual manifolds of different people doing the same activity will be different. So, how to handle all these variabilities. Let's assume the simple case first, a single view point and we deal with human silhouettes so we do not have any variability due to illumination or appearance. Let the only source of variability be variation in people silhouette shapes. The problem now is how to extend the generative model in Equation 1 to include a variable describing people shape variability. For example, given several sequences of walking silhouettes, as in Fig. 7, with different people walking, how to decompose the intrinsic body configuration through the action from the appearance (or shape) of the person performing the action. we aim to learn a decomposable generative model that explicitly decomposes the following two factors:

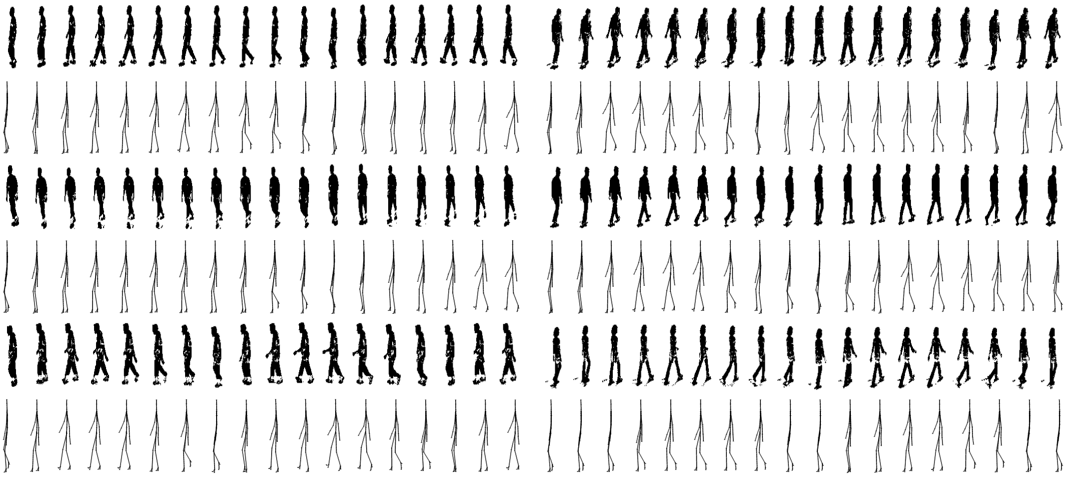


Fig. 6. 3D reconstruction for 4 people from different views: person 70 views 1,2; person 86 views 1,2; person 76 view 4; person 79 view 4

- Content (body pose): A representation of the intrinsic body configuration through the motion as a function of time that is invariant to the person, i.e., the content characterizes the motion or the activity.
- Style (people) : Time-invariant person parameters that characterize the person appearance (shape).

On the other hand, given an observation of certain person at a certain body pose and given the learned generative model we aim to be able to solve for both the body configuration representation (content) and the person parameter (style). In our case the content is a continuous domain while style is represented by the discrete style classes which exist in the training data where we can interpolate intermediate styles and/or intermediate contents.

This can be formulated as a view-based generative model in the form

$$y_t^s = \gamma(x_t^c; a, b^s) \tag{8}$$

where the image, y_t^s, at time t and of style s is an instance driven from a generative model where the function $\gamma(\cdot)$ is a mapping function that maps from a representation of body configuration x_t^c (content) at time t into the image space given mapping parameters a and style dependent parameter b^s that is time invariant[2]. A framework was introduced in [21] to learn a decomposable generative model that explicitly decomposes the intrinsic body configuration (content) as a function of time from the appearance (style) of the person performing the action as time-invariant parameter. The framework is based on decomposing the style parameters in the space of nonlinear functions that maps between a learned unified nonlinear embedding of multiple content manifolds and the visual input space.

[2] We use the superscript s, c to indicate which variables depend on style or content respectively.

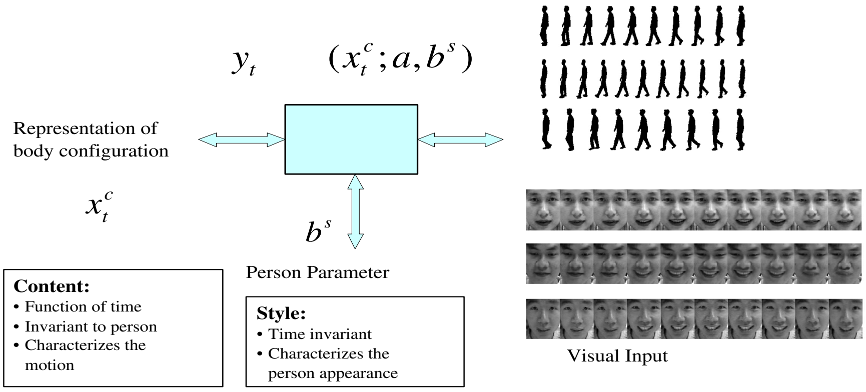


Fig. 7. Style and content factors: Content: gait motion or facial expression. Style: different silhouette shapes or face appearance.

Suppose that we can learn a unified, style-invariant, nonlinearly embedded representation of the motion manifold $\mathcal{M}$ in a low dimensional Euclidean embedding space, $\mathbb{R}^e$, then we can learn a set of style-dependent nonlinear mapping functions from the embedding space into the input space, i.e., functions $\gamma_s(x_t^c) : \mathbb{R}^e \rightarrow \mathbb{R}^d$ that maps from embedding space with dimensionality e into the input space (observation) with dimensionality d for style class s. Since we consider nonlinear manifolds and the embedding is nonlinear, the use of nonlinear mapping is necessary. We consider mapping functions in the form

$$y_t^s = \gamma_s(x_t) = C^s \cdot \psi(x_t^c) \tag{9}$$

where C^s is a $d \times N$ linear mapping and $\psi(\cdot) : \mathbb{R}^e \rightarrow \mathbb{R}^N$ is a nonlinear mapping where N basis functions are used to model the manifold in the embedding space, i.e.,

$$\psi(\cdot) = [\psi_1(\cdot), \cdots, \psi_N(\cdot)]^T$$

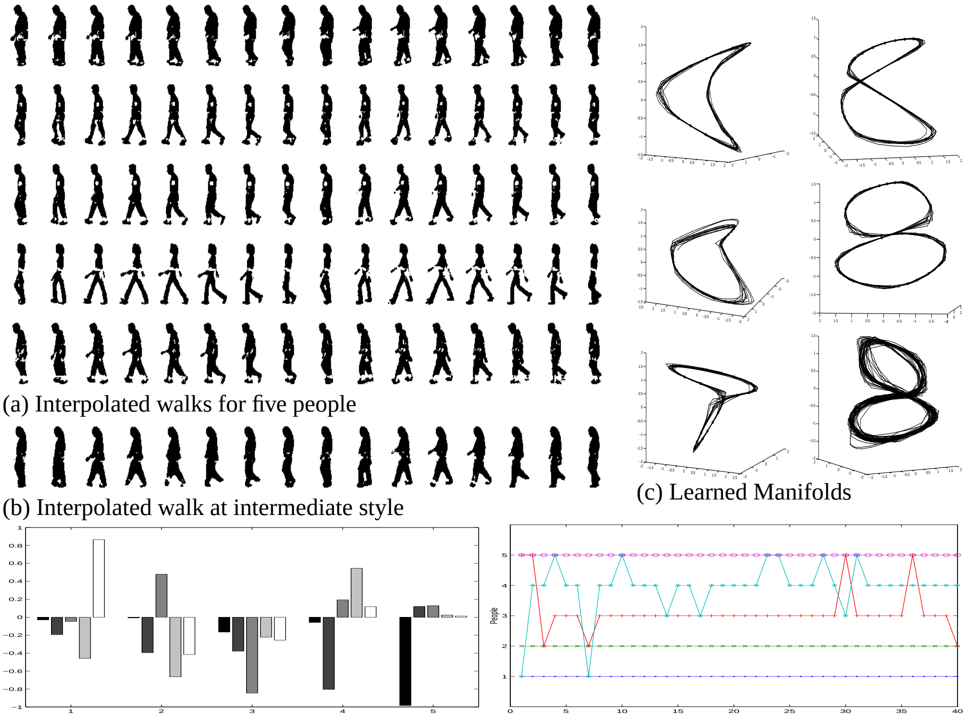
Given learned models of the form of Equation 9, the style can be decomposed in the linear mapping coefficient space using bilinear model in a way similar to [76, 80]. Therefore, input instance y_t can be written as asymmetric bilinear model in the linear mapping space as

$$y_t = \mathcal{A} \times_3 b^s \times_2 \psi(x_t^c) \tag{10}$$

where $\mathcal{A}$ is a third order tensor (3-way array) with dimensionality $d \times N \times J$, b^s is a style vector with dimensionality J, and $\times_n$ denotes mode-n tensor product. Given the role for style and content defined above, the previous equation can be written as

$$y_t = \mathcal{A} \times_3 b^{\mathit{people}} \times_2 \psi(x_t^{\mathit{pose}}) \tag{11}$$

Figure,8 shows examples for decomposing styles for gait. The learned generative model is used to interpolate walking sequences at new styles as well as to solve for the style parameters and body pose. In this experiment we used five sequences for five different people [3] each containing about 300 frames which are noisy. The learned manifolds are shown in Figure 8-b which shows a different manifold for each person. The learned unified manifold is also shown in Figure 8-e. Figure 8 shows interpolate walking sequences for the five people generated by the learned model. The figure also shows the learned style vectors. We evaluated style classifications using 40 frames for each person and the result is shown in the figure with correct classification rate of 92%. We also used the learned model to interpolate walks in new styles. The last row in the figure shows interpolation between person 1 and person 4.



(a) Interpolated walks for five people

(b) Interpolated walk at intermediate style

(c) Learned Manifolds

(d) Style parameters

(e) Style classification

Fig. 8. (a) interpolated walks for five people. (b) Interpolated walk at intermediate style between person 1 and 4. (c) Learned manifolds for the five people and the unified manifold (bottom right). (d) Estimated style parameters given the unified manifold. (e) Style classification for test data of 40 frames for 5 people.

4 Style Adaptive Tracking: Bayesian Tracking on a Manifold

Given the explicit manifold model and the generative model learned in section 3, we can formulate contour tracking within a Bayesian tracking framework. We can achieve

[3] The data are from CMU Mobogait database

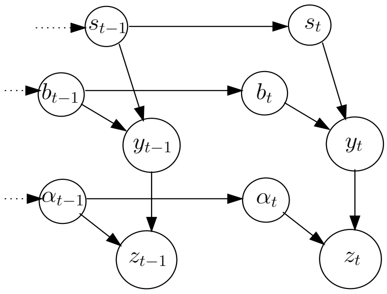


Fig. 9. Graphic model for decomposed generative model

style adaptive contour tracking on cluttered environments where the generative model can be used to as an observation model to generate contours of different people shape styles and different poses. The tracking is performed on three conceptually independent spaces: body configuration space, shape style space and geometric transformation space. Therefore, object state combines heterogeneous representations. The manifold provides a constraint on the motion, which reduces the system dynamics of the global nonlinear deformation into a linear dynamic system. The challenge will be how to represent and handle multiple spaces without falling into exponential increase of the state space dimensionality. Also, how to do tracking in a shape space which can be high dimensional?

Figure 9 shows a graphical model illustrating the relation between different variables. The shape at each time step is an instance driven from a generative model. Let $z_t \in R^d$ be the shape of the object at time instance t represented as a point in a d-dimensional space. This instance of the shape is driven from a model in the form

$$z_t = T_{\alpha_t}\gamma(b_t; s_t), \tag{12}$$

where the $\gamma(\cdot)$ is a nonlinear mapping function that maps from a representation of the body configuration b_t into the observation space given a mapping parameter s_t that characterizes the person shape in a way independent from the configuration and specific for the person being tracked. T_{α_t} represents a geometric transformation on the shape instance. Given this generative model, we can fully describe observation instance z_t by state parameters α_t, b_t, and s_t. The mapping $\gamma(b_t; s_t)$ is a nonlinear mapping from the body configuration state b_t as

$$y_t = \mathcal{A} \times s^t \times \psi(b_t), \tag{13}$$

where $\psi(b_t)$ is a kernel induced space, $\mathcal{A}$ is a third order tensor, s^k is a shape style vector and $\times$ is appropriate tensor product.

The tracking problem is then an inference problem where at time t we need to infer the body configuration representation b_t and the person specific parameter s_t and the geometric transformation T_{α_t} given the observation z_t. The Bayesian tracking framework enables a recursive update of the posterior $P(X_t|Z^t)$ over the object state X_t

given all observation $Z^t = Z_1, Z_2, .., Z_t$ up to time t:

$$P(X_t|Z^t) \propto P(Z_t|X_t) \int_{X_{t-1}} P(X_t|X_{t-1}) P(X_{t-1}|Z^{t-1}) \qquad (14)$$

In our generative model, the state X_t is $[\alpha_t, b_t, s_t]$, which uniquely describes the state of the tracking object. Observation Z_t is the captured image instance at time t.

The state X_t is decomposed into three sub-states α_t, b_t, s_t. These three random variables are conceptually independent since we can combine any body configuration with any person shape style with any geometrical transformation to synthesize a new contour. However, they are dependent given the observation Z_t. It is hard to estimate joint posterior distribution $P(\alpha_t, b_t, s_t|Z_t)$ for its high dimensionality. The objective of the density estimation is to estimate states α_t, b_t, s_t for a given observation. The decomposable feature of our generative model enables us to estimate each state by a marginal density distribution $P(\alpha_t|Z^t)$, $P(b_t|Z^t)$, and $P(s_t|Z^t)$. We approximate marginal density estimation of one state variable along representative values of the other state variables. For example, in order to estimate marginal density of $P(b_t|Z^t)$, we estimate $P(b_t|\alpha_t^*, s_t^*, Z^t)$, where α_t^*, s_t^* are representative values such as maximum posteriori estimates.

Modeling body configuration space: Given a set of training data for multiple people, a unified mean manifold embedding can be obtained as was explained in section 3. The mean manifold can be parameterized by a one-dimensional parameter $\beta_t \in R$ and a spline fitting function $f : R \rightarrow R^3$, which satisfies $b_t = f(\beta_t)$, to map from the parameter space into the three dimensional embedding space.

Modeling style shape space: Shape style space is parameterized by a linear combination of basis of the style space. A generative model in the form of Equation 13 is fitted to the training data. Ultimately the style parameter s should be independent of the configuration and therefore should be time invariant and can be estimated at initialization. However, we don't know the person style initially and , therefore, the style needs to fit to the correct person style gradually during the tracking. So, we formulated style as time variant factor that should stabilize after some frames from initialization. The dimension of the style vector depends on the number of people used for training and can be high dimensional.

We represent new style as a convex linear combination of style classes learned from the training data. The tracking of the high dimensional style vector s_t itself will be hard as it can fit local minima easily. A new style vector s is represented by linear weighting of each of the style classes s^k, $k = 1, \cdots, K$ using linear weight λ^k:

$$s = \sum_{k=1}^{K} \lambda^k s^k, \qquad \sum_{k=1}^{K} \lambda^k = 1, \qquad (15)$$

where K is the number of style classes used to represent new styles. The overall generative model can be expressed as

$$z_t = T_{\alpha_t} \left(\mathcal{A} \times \left[\sum_{k=1}^{K} \lambda_t^k s^k \right] \times \psi(f(\beta_t)) \right). \qquad (16)$$

Tracking problem using this generative model is the estimation of parameter α_t, β_t, and λ_t at each new frame given the observation z_t. Tracking can be done using a particle filter as was shown in [42, 43]. Figures 10 and 11 show style adaptive tracking results for two subjects. In the first case, the person style is in the training set while in the second case the person was not seen before in the training. In both cases, the style parameter started at the mean style and adapted correctly to the person shape. It is clear that the estimated body configuration shows linear dynamics and the particles are showing a gaussian distribution on the manifold.

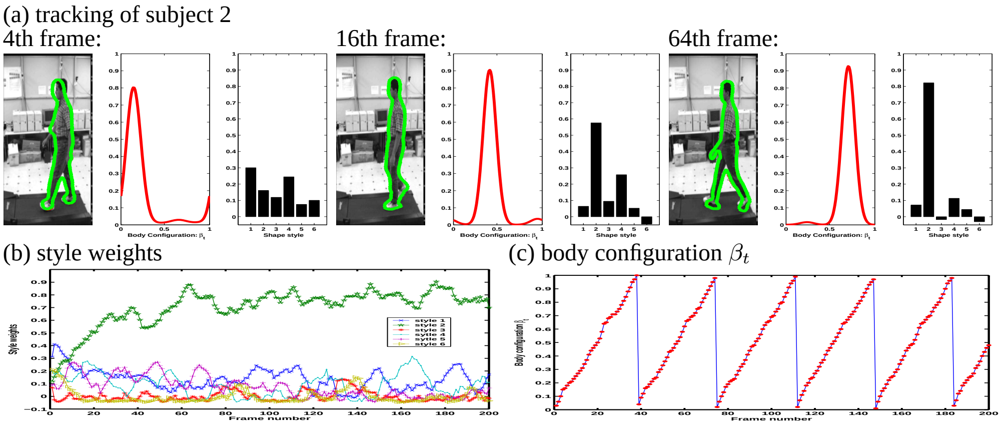

(a) tracking of subject 2
4th frame: 16th frame: 64th frame:
(b) style weights (c) body configuration β_t

Fig. 10. Tracking for known person

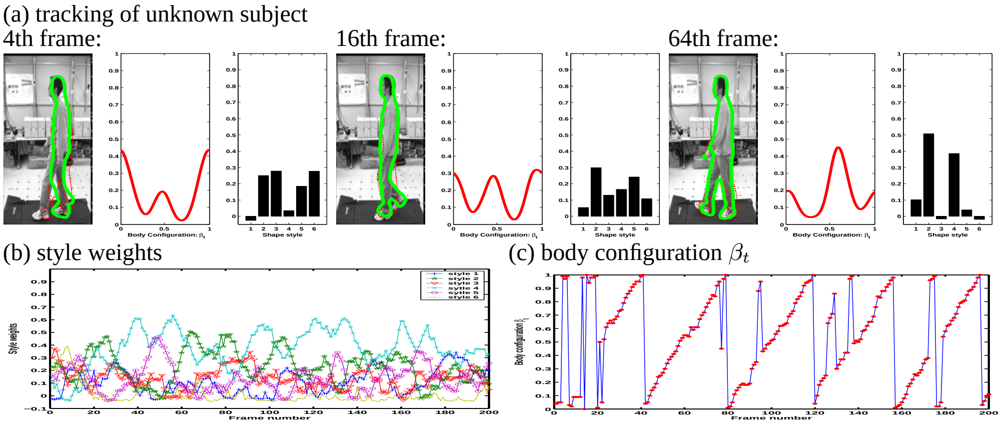

(a) tracking of unknown subject
4th frame: 16th frame: 64th frame:
(b) style weights (c) body configuration β_t

Fig. 11. Tracking for unknown person

5 Adding More Variability: Decomposable Generative Model

In section 3 it was shown how to separate a style factor when learning a generative model for data lying on a manifold. Here we generalize this concept to decompose several style factors. For example, consider the walking motion observed from multiple view points (as silhouettes). The resulting data lie on multiple subspaces and/or multiple manifolds. There is the underling motion manifold, which is one dimensional for the gait motion. There is the view manifold and the space of different people's shapes. Another example we consider is facial expressions. Consider face data of different people performing different facial dynamic expressions such as sad, smile, surprise, etc. The resulting face data posses several dimensionality of variability: the dynamic motion, the expression type and the person face. So, how to model such data in a generative manner. We follow the same framework of explicitly modeling the underlying motion manifold and over that we decompose various style factors.

We can think of the image appearance (similar argument for shape) of a dynamic object as instances driven from such generative model. Let $y_t \in R^d$ be the appearance of the object at time instance t represented as a point in a d-dimensional space. This instance of the appearance is driven from a model in the form

$$y_t = T_\alpha \gamma(x_t; a_1, a_2, \cdots, a_n) \tag{17}$$

where the appearance, y_t, at time t is an instance driven from a generative model where the function γ is a mapping function that maps body configuration x_t at time t into the image space. i.e., the mapping function γ maps from a representation of the body configuration space into the image space given mapping parameters $a_1, \cdots, a_n$ each representing a set of conceptually orthogonal factors. Such factors are independent of the body configuration and can be time variant or invariant. The general form for the mapping function γ that we use is

$$\gamma(x_t; a_1, a_2, \cdots, a_n) = \mathcal{C} \times_1 a_1 \times \cdots \times_n a_n \cdot \psi(x_t) \tag{18}$$

where $\psi(x)$ is a nonlinear kernel map from a representation of the body configuration to a kernel induced space and each a_i is a vector representing a parameterization of orthogonal factor i, $\mathcal{C}$ is a core tensor, $\times_i$ is *mode-i* tensor product as defined in [37, 81].

For example for the gait case, a generative model for walking silhouettes for different people from different view points will be in the form

$$y_t = \gamma(x_t; v, s) = \mathcal{C} \times v \times s \times \psi(x) \tag{19}$$

where v is a parameterization of the view, which is independent of the body configuration but can change over time, and s is a parameterization of the shape style of the person performing the walk which is independent of the body configuration and time invariant. The body configuration x_t evolves along a representation of the manifold that is homeomorphic to the actual gait manifold.

Another example is modeling the manifolds of facial expression motions. Given dynamic facial expression such as sad, surprise, happy, etc., where each expression start from neutral and evolve to a peak expression; each of these motions evolves along a one dimensional manifold. However, the manifold will be different for each person and for each expression. Therefore, we can use a generative model to generate different people faces and different expressions using a model in the form be in the form

$$y_t = \gamma(x_t; e, f) = \mathcal{A} \times e \times f \times \psi(x_t) \tag{20}$$

where e is an expression vector (happy, sad, etc.) that is invariant of time and invariant of the person face, i.e., it only describes the expression type. Similarly, f is a face vector describing the person face appearance which is invariant of time and invariant of the expression type. The motion content is described by x which denotes the motion phase of the expression, i.e., starts from neutral and evolves to a peak expression depending on the expression vector, e.

The model in Equation 18 is a generalization over the model in equations 1 and 8. However, such generalization is not obvious. In section 3 LLE was used to obtain manifold embeddings, and then a mean manifold is computed as a unified representation through nonlinear warping of manifold points. However, since the manifolds twists very differently given each factor (different people or different views, etc.) it is not possible to achieve a unified configuration manifold representation independent of other factors. These limitations motivate the use of a conceptual unified representation of the configuration manifold that is independent of all other factors. Such unified representation would allow the model in Equation 18 to generalize to decompose as many factors as desired. In the model in Equation 18, the relation between body configuration and the input is nonlinear where other factors are approximated linearly through multilinear analysis. The use of nonlinear mapping is essential since the embedding of the configuration manifold is nonlinearly related to the input.

The question is what conceptual representation of the manifold we can use. For example, for the gait case, since the gait is one dimensional closed manifold embedded in the input space, it is homeomorphic to a unit circle embedded in 2D. In general, all closed 1 D manifold is topologically homeomorphic to unit circles. We can think of it as a circle twisted and stretched in the space based on the shape and the appearance of the person under consideration or based on the view. So we can use such unit circle as a unified representation of all gait cycles for all people for all views. Given that all the manifolds under consideration are homeomorphic to unit circle, the actual data is used to learn nonlinear warping between the conceptual representation and the actual data manifold. Since each manifold will have its own mapping, we need to have a mechanism to parameterize such mappings and decompose all these mappings to parameterize variables for views, different people, etc.

Given an image sequences $y_t^a, t = 1, \cdots, T$ where a denotes a particular class setting for all the factors $a_1, \cdots, a_n$ (e.g., a particular person s and view v) representing a whole motion cycle and given a unit circle embedding of such data as $x_t^a \in R^2$ we can learn a nonlinear mapping in the form

$$y_t^a = B^a \psi(x_t^a) \tag{21}$$

Given such mapping the decomposition in Equation 1 can be achieved using tensor analysis of the coefficient space such that the coefficient B^a are obtained from a multilinear [81] model

$$B^a = \mathcal{C} \times_1 a_1 \times \cdots \times_n a_n$$

Given a training data and a model fitted in the form of Equation 18 it is desired to use such model to recover the body configuration and each of the orthogonal factors involved, such as view point and person shape style given a single test image or given a full or a part of a motion cycle. Therefore, we are interested in achieving an efficient solution to a nonlinear optimization problem in which we search for x^*, a_i^* which minimize the error in reconstruction

$$E(x, a_1, \cdots, a_n) = || \; y - \mathcal{C} \times_1 a_1 \times \cdots \times_n a_n \times \psi(x) \; || \qquad (22)$$

or a robust version of the error. In [41] an efficient algorithms were introduced to recover these parameters in the case of a single image input or a sequence of images using deterministic annealing.

5.1 Dynamic Shape Example: Decomposing View and Style on Gait Manifold

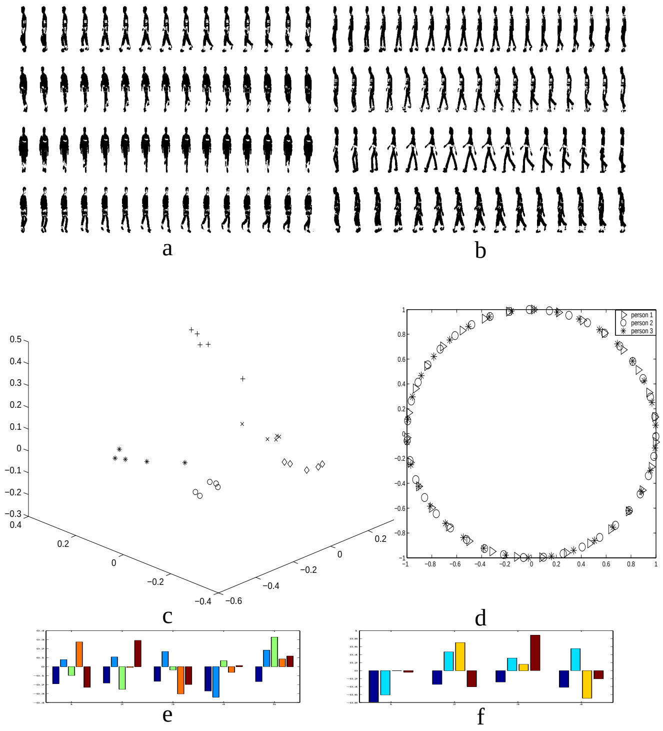


Fig. 12. a,b) Example of training data. Each sequence shows a half cycle only. a) four different views used for person 1 b) side views of people 2,3,4,5. c) style subspace: each person cycles have the same label. d) unit circle embedding for three cycles. e) Mean style vectors for each person cluster. f) View vectors

In this section we show an example of learning the nonlinear manifold of gait as an example of a dynamic shape. We used CMU Mobo gait data set [25] which contains walking people from multiple synchronized views[4]. For training we selected five people, five cycles each from four different views. i.e., total number of cycles for training is 100=5 people × 5 cycles × 4 views. Note that cycles of different people and cycles of the same person are not of the same length. Figure 12-a,b show examples of the sequences (only half cycles are shown because of limited space).

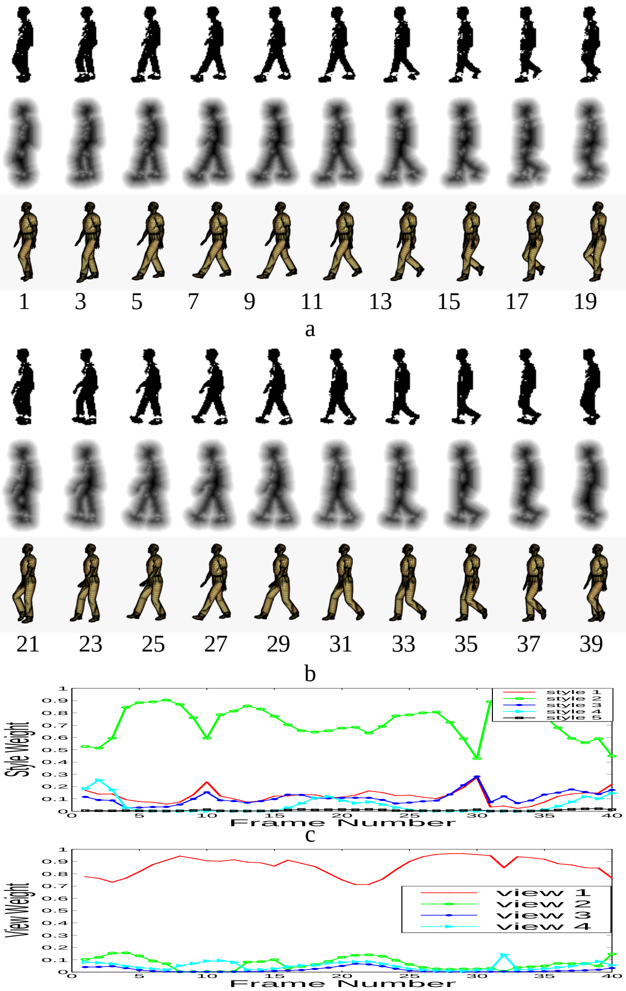


Fig. 13. a,b) example pose recovery. from top to bottom: input shapes, implicit function, recovered 3D pose. c) Style weights. d) View weights.

[4] CMU Mobo gait data set [25] contains 25 people, about 8 to 11 walking cycles each captured from six different view points. The walkers were using a treadmill.

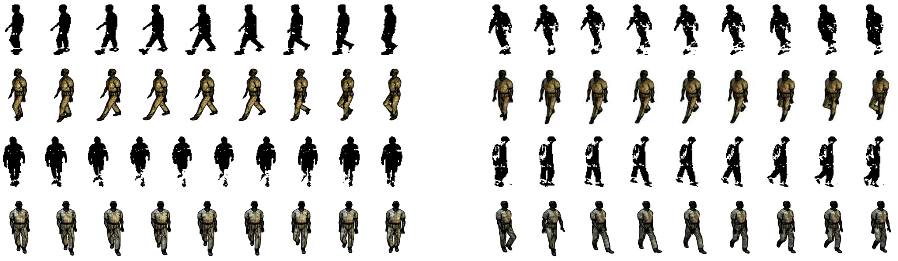


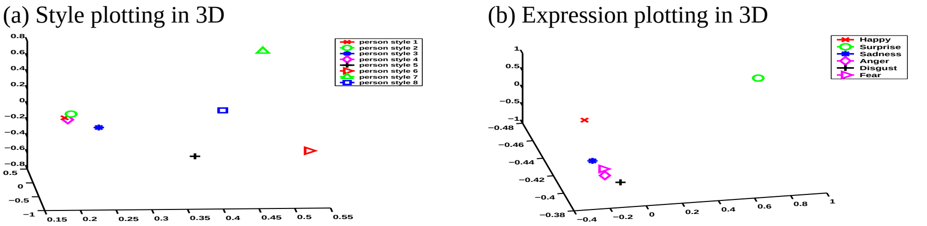
Fig. 14. Examples of pose recovery and view classification for four different people from four views.

The data is used to fit the model as described in Equation 19. Images are normalized to 60×100, i.e., $d = 6000$. Each cycle is considered to be a style by itself, i.e., there are 25 styles and 4 views. Figure 12-d shows example of model-based aligned unit circle embedding of three cycles. Figure 12-c shows the obtained style subspace where each of the 25 points corresponding to one of the 25 cycles used. Important thing to notice is that the style vectors are clustered in the subspace such that each person style vectors (corresponding to different cycles of the same person) are clustered together which indicate that the model can find the similarity in the shape style between different cycles of the same person. Figure 12-e shows the mean style vectors for each of the five clusters. Figure 12-f shows the four view vectors.

Figure 13 shows example of using the model to recover the pose, view and style. The figure shows samples of a one full cycle and the recovered body configuration at each frame. Notice that despite the subtle differences between the first and second halves of the cycle, the model can exploit such differences to recover the correct pose. The recovery of 3D joint angles is achieved by learning a mapping from the manifold embedding and 3D joint angle from motion captured data using GRBF in a way similar to Equation 21. Figure 13-c,d shows the recovered style weights (class probabilities) and view weights respectively for each frame of the cycle which shows correct person and view classification. Figure 14 shows examples recovery of the 3D pose and view class for four different people non of them was seen in training.

5.2 Dynamic Appearance Example: Facial Expression Analysis

We used the model to learn facial expressions manifolds for different people. We used CMU-AMP facial expression database where each subject has 75 frames of varying facial expressions. We choose four people and three expressions each (smile, anger, surprise) where corresponding frames are manually segmented from the whole sequence for training. The resulting training set contained 12 sequences of different lengths. All sequences are embedded to unit circles and aligned as described in section 5. A model in the form of Equation 20 is fitted to the data where we decompose two factors: person facial appearance style factor and expression factor besides the body configuration which is nonlinearly embedded on a unit circle. Figure 15 shows the resulting person style vectors and expression vectors.


(a) Style plotting in 3D

(b) Expression plotting in 3D

Fig. 15. Facial expression analysis for Cohn-Kanade Dataset for 8 subjects with 6 expressions and their 3D space plotting

We used the learned model to recognize facial expression, and person identity at each frame of the whole sequence. Figure 16 shows an example of a whole sequence and the different expression probabilities obtained on a frame per frame basis. The figure also shows the final expression recognition after thresholding along manual expression labelling. The learned model was used to recognize facial expressions for sequences of people not used in the training. Figure 17 shows an example of a sequence of a person not used in the training. The model can successfully generalizes and recognize the three learned expression for this new subject.

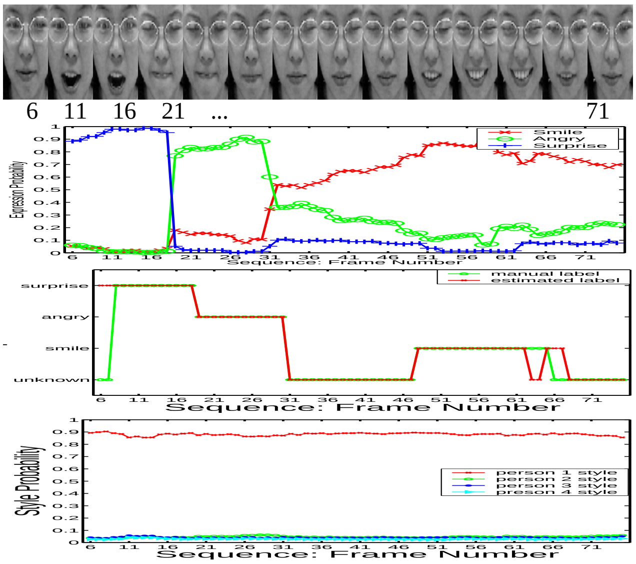


Fig. 16. From top to bottom: Samples of the input sequences; Expression probabilities; Expression classification; Style probabilities

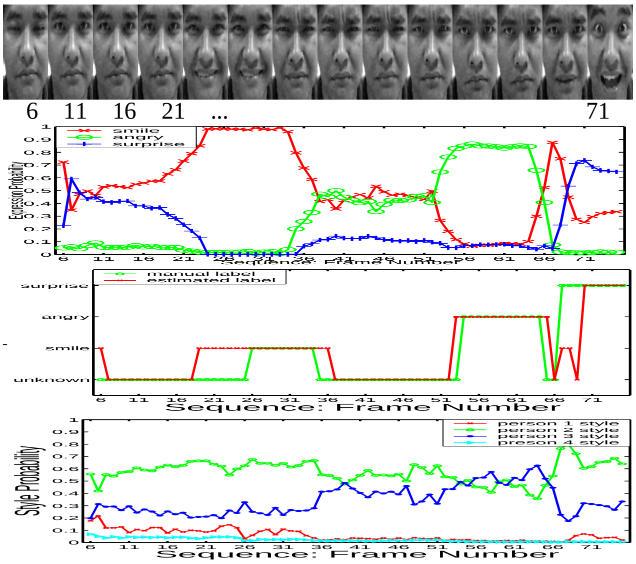


Fig. 17. Generalization to new people: expression recognition for a new person. From top to bottom: Samples of the input sequences; Expression probabilities; Expression classification; Style probabilities

6 Conclusion

In this chapter we focused on exploiting the underlying motion manifold for human motion analysis and synthesis. we introduced a framework for learning a landmark-free correspondence-free global representations of dynamic shape and dynamic appearance manifolds. The framework is based on using nonlinear dimensionality reduction to achieve an embedding of the global deformation manifold which preserves the geometric structure of the manifold. Given such embedding, a nonlinear mapping is learned from such embedded space into visual input space using RBF interpolation. Given this framework, any visual input is represented by a linear combination of nonlinear bases functions centered along the manifold in the embedded space. In a sense, the approach utilizes the implicit correspondences imposed by the global vector representation which are only valid locally on the manifold through explicit modeling of the manifold and RBF interpolation where closer points on the manifold will have higher contributions than far away points.

We also showed how approximate solution for the inverse mapping can be obtained in a closed form which facilitates recovery of the intrinsic body configuration. The framework was applied to learn a representation of the gait manifold as an example of a dynamic shape manifold. We showed how the learned representation can be used to interpolate intermediate body poses as well as in recovery and reconstruction of the input. We extended the approach to learn mappings from the embedded motion manifold to 3D joint angle representation which yields an approximate closed-form solution for 3D pose recovery.

We show how to learn a decomposable generative model that separates appearance variations from the intrinsics underlying dynamics manifold though introducing a framework for separation of style and content on a nonlinear manifold. The framework is based on decomposing the style parameters in the space of nonlinear functions that maps between a learned unified nonlinear embedding of multiple content manifolds and the visual input space. The framework yields an unsupervised procedure that handles dynamic, nonlinear manifolds. It also improves on past work on nonlinear dimensionality reduction by being able to handle multiple manifolds. The proposed framework was shown to be able to separate style and content on both the gait manifold and a simple facial expression manifold. As mention in [68], an interesting and important question is how to learn a parametric mapping between the observation and nonlinear embedding spaces. We partially addressed this question.

The use of a generative model is necessary since the mapping from the manifold representation to the input space will be well defined in contrast to a discriminative model where the mapping from the visual input to manifold representation is not necessarily a function. We introduced a framework to solve for various factors such as body configuration, view, and shape style. Since the framework is generative, it fits well in a Bayesian tracking framework and it provides separate low dimensional representations for each of the modelled factors. Moreover, a dynamic model for configuration is well defined since it is constrained to the 1D manifold representation. The framework also provides a way to initialize a tracker by inferring about body configuration, view point, body shape style from a single or a sequence of images.

The framework presented in this chapter was basically applied to one-dimensional motion manifolds such as gait and facial expressions. One-dimensional manifolds can be explicitly modeled in a straight forward way. However, there is no theoretical restriction that prevents the framework from dealing with more complicated manifolds. In this chapter we mainly modeled the motion manifold while all appearance variability are modeled using subspace analysis. Extension to modeling multiple manifolds simultaneously is very challenging. We investigated modeling both the motion and the view manifolds in [46]. The proposed framework has been applied to gait analysis and recognition in [39, 42, 44, 43]. It was also used in analysis and recognition of facial expressions in [40, 45].

7 Acknowledgment

This research is partially funded by NSF award IIS-0328991

References

1. P. N. Belhumeur, J. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. In *ECCV (1)*, pages 45–58, 1996.
2. M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003.
3. Y. Bengio, O. Delalleau, N. Le Roux, J.-F. Paiement, P. Vincent, and M. Ouimet. Learning eigenfunctions links spectral embedding and kernel pca. *Neural Comp.*, 16(10):2197–2219, 2004.

4. Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. L. Roux, and M. Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In *NIPS 16*, 2004.
5. D. Beymer and T. Poggio. Image representations for visual learning. *Science*, 272(5250), 1996.
6. M. J. Black and A. D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. In *ECCV (1)*, pages 329–342, 1996.
7. A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.
8. R. Bowden. Learning statistical models of human motion. In *IEEE Workshop on Human Modelling, Analysis and Synthesis*, 2000.
9. M. Brand. Shadow puppetry. In *International Conference on Computer Vision*, volume 2, page 1237, 1999.
10. 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.
11. C. Bregler and S. M. Omohundro. Nonlinear manifold learning for visual speech recognition. pages 494– 499, 1995.
12. L. W. Campbell and A. F. Bobick. Recognition of human body motion using phase space constraints. In *ICCV*, pages 624–630, 1995.
13. Z. Chen and H. Lee. Knowledge-guided visual perception of 3-d human gait from single image sequence. *IEEE SMC*, 22(2):336–342, 1992.
14. C. M. Christoudias and T. Darrell. On modelling nonlinear shape-and-texture appearance manifolds. In *Proc.of IEEE CVPR*, volume 2, pages 1067–1074, 2005.
15. T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models: Their training and application. *CVIU*, 61(1):38–59, 1995.
16. T. Cox and M. Cox. *Multidimentional scaling*. Chapman & Hall, 1994.
17. R. Cutler and L. Davis. Robust periodic motion and motion symmetry detection. In *Proc. IEEE CVPR*, 2000.
18. T. Darrell and A. Pentland. Space-time gesture. In *Proc IEEE CVPR*, 1993.
19. A. Elgammal. Nonlinear generative models for dynamic shape and dynamic appearance. In *Proc. of 2nd International Workshop on Generative-Model based vision. GMBV 2004*, July 2004.
20. A. Elgammal and C.-S. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, June-July 2004.
21. 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.
22. 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.
23. 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.
24. R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky. ‘Dynamism of a dog on a leash’ or behavior classification by eigen-decomposition of periodic motions. In *Proceedings of the ECCV’02*, pages 461–475, Copenhagen, May 2002. Springer-Verlag, LNCS 2350.
25. R. Gross and J. Shi. The cmu motion of body (mobo) database. Technical Report TR-01-18, Carnegie Mellon University, 2001.
26. J. Ham, D. D. Lee, S. Mika, and B. Sch&#246;lkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of ICML*, page 47, New York, NY, USA, 2004. ACM Press.

27. I. Haritaoglu, D. Harwood, and L. S. Davis. W4: Who ? when ? where? what? a real time system for detecting and tracking people. In *3rd International Conference on Face and Gesture Recognition*, 1998.
28. D. Hogg. Model-based vision: a program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.
29. Howe, Leventon, and W. Freeman. Bayesian reconstruction of 3d human motion from single-camera video. In *Proc. NIPS*, 1999.
30. H.S.Seung and D. D. Lee. The manifold ways of perception. *Science*, 290(5500):2268–2269, December 2000.
31. I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.
32. J.O'Rourke and Badler. Model-based image analysis of human motion using constraint propagation. *IEEE PAMI*, 2(6), 1980.
33. 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.
34. 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.
35. A. Kapteyn, H. Neudecker, and T. Wansbeek. An approach to n-model component analysis. *Psychometrika*, 51(2):269–275, 1986.
36. T. D. Kristen Grauman, Gregory Shakhnarovich. Inferring 3d structure with a statistical image-based shape model. In *ICCV*, 2003.
37. 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.
38. N. Lawrence. Gaussian process latent variable models for visualization of high dimensional data. In *NIPS*, 2003.
39. C.-S. Lee and A. Elgammal. Gait style and gait content: Bilinear models for gait recognition using gait re-sampling. In *6th International Conference on Automatic Face and Gestutre Recognition FG04*, 2004.
40. C.-S. Lee and A. Elgammal. Facial expression analysis using nonlinear decomposable generative models. In *AMFG*, pages 17–31, 2005.
41. C.-S. Lee and A. Elgammal. Homeomorphic manifold analysis: Learning decomposable generative models for human motion analysis. In *Workshop on Dynamical Vision*, 2005.
42. C.-S. Lee and A. Elgammal. Style adaptive bayesian tracking using explicit manifold learning. In *Proc. of British Machine Vision Conference*, pages 739–748, 2005.
43. C.-S. Lee and A. Elgammal. Gait tracking and recognition using person-dependent dynamic shape model. In *FGR*, volume 0, pages 553–559. IEEE Computer Society, 2006.
44. C.-S. Lee and A. M. Elgammal. Towards scalable view-invariant gait recognition: Multilinear analysis for gait. In *AVBPA*, pages 395–405, 2005.
45. C.-S. Lee and A. M. Elgammal. Nonlinear shape and appearance models for facial expression analysis and synthesis. In *ICPR (1)*, pages 497–502, 2006.
46. C.-S. Lee and A. M. Elgammal. Simultaneous inference of view and body pose using torus manifolds. In *ICPR (3)*, pages 489–494, 2006.
47. A. Levin and A. Shashua. Principal component analysis over continuous subspaces and intersection of half-spaces. In *ECCV, Copenhagen, Denmark*, pages 635–650, May 2002.
48. J. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, New York, New York, 1988.
49. D. Marimont and B. Wandell. Linear models of surface and illumination spectra. *J. Optical Society od America*, 9:1905–1913, 1992.

50. P. Mordohai and G. Medioni. Unsupervised dimensionality estimation and manifold learning in high-dimensional spaces by tensor voting. In *Proceedings of International Joint Conference on Artificial Intelligence*, 2005.
51. G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *European Conference on Computer Vision*, 2002.
52. M.Turk and A.Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.
53. H. Murase and S. Nayar. Visual learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.
54. R. C. Nelson and R. Polana. Qualitative recognition of motion using temporal texture. *CVGIP Image Understanding*, 56(1):78–89, 1992.
55. S. Niyogi and E. Adelson. Analyzing and recognition walking figures in xyt. In *Proc. IEEE CVPR*, pages 469–474, 1994.
56. T. Poggio and F. Girosi. Network for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.
57. R. Polana and R. Nelson. Low level recognition of human motion (or how to get your man without finding his body parts). In *IEEE Workshop on Non-Rigid and Articulated Motion*, pages 77–82, 1994.
58. R. Polana and R. C. Nelson. Qualitative detection of motion by a moving observer. *International Journal of Computer Vision*, 7(1):33–46, 1991.
59. R. Polana and R. C. Nelson. Detecting activities. *Journal of Visual Communication and Image Representation*, June 1994.
60. A. Rahimi, B. Recht, and T. Darrell. Learning appearane manifolds from video. In *Proc.of IEEE CVPR*, volume 1, pages 868–875, 2005.
61. J. M. Rehg and T. Kanade. Visual tracking of high DOF articulated structures: an application to human hand tracking. In *ECCV (2)*, pages 35–46, 1994.
62. J. M. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *ICCV*, pages 612–617, 1995.
63. J. Rittscher and A. Blake. Classification of human body motion. In *IEEE International Conferance on Compute Vision*, 1999.
64. K. Rohr. Towards model-based recognition of human movements in image sequence. *CVGIP*, 59(1):94–115, 1994.
65. R. Rosales, V. Athitsos, and S. Sclaroff. 3D hand pose reconstruction using specialized mappings. In *Proc. ICCV*, 2001.
66. R. Rosales and S. Sclaroff. Inferring body pose without tracking body parts. Technical Report 1999-017, 1, 1999.
67. 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.
68. S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Sciene*, 290(5500):2323–2326, 2000.
69. B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. The MIT Press, Cambridge, Massachusetts, 2002.
70. G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *ICCV*, 2003.
71. A. Shashua and A. Levin. Linear image coding of regression and classification using the tensor rank principle. In *Proc. of IEEE CVPR, Hawai*, 2001.
72. 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.
73. 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.

74. Y. Song, X. Feng, and P. Perona. Towards detection of human motion. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2000)*, pages 810–817, 2000.
75. J. Tenenbaum. Mapping a manifold of perceptual observations. In *Advances in Neural Information Processing*, volume 10, pages 682–688, 1998.
76. J. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12:1247–1283, 2000.
77. K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *ICCV*, pages 50–59, 2001.
78. L. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966.
79. M. A. O. Vasilescu. An algorithm for extracting human motion signatures. In *Proc. of IEEE CVPR, Hawai*, 2001.
80. M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensebles: Tensorfaces. In *Proc. of ECCV, Copenhagen, Danmark*, pages 447–460, 2002.
81. M. A. O. Vasilescu and D. Terzopoulos. Multilinear subspace analysis of image ensembles. 2003.
82. R. Vidal and R. Hartley. Motion segmentation with missing data using powerfactorization and gpca. In *Proceedings of IEEE CVPR*, volume 2, pages 310–316, 2004.
83. R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (gpca). In *Proceedings of IEEE CVPR*, volume 1, pages 621–628, 2003.
84. H. Wang and N. Ahuja. Rank-r approximation of tensors: Using image-as-matrix representation. In *Proceedings of IEEE CVPR,* volume 2, pages 346–353, 2005.
85. K. W. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by semidefinite programming. In *Proceedings of IEEE CVPR,* volume 2, pages 988–995, 2004.
86. 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.
87. Y. Yacoob and M. J. Black. Parameterized modeling and recognition of activities. *Computer Vision and Image Understanding: CVIU*, 73(2):232–247, 1999.