Homeomorphic Manifold Analysis: Learning Decomposable Generative Models for Human Motion Analysis

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

Abstract

If we consider the appearance of human motion such as gait, facial expression and gesturing, most of such activities result in nonlinear manifolds in the image space. Although the intrinsic body configuration manifolds might be very low in dimensionality, the resulting appearance manifold is challenging to model given various aspects that affects the appearance such as the view point, the person shape and appearance, etc. In this paper we learn decomposable generative models that explicitly decompose the intrinsic body configuration as a function of time from other conceptually orthogonal aspects that affects the appearance such as the view point, the person performing the action, etc. The frameworks is based on learning nonlinear mappings from a conceptual representation of the motion manifold that is homeomorphic to the actual manifold and decompose other sources of variation in the mapping coefficient space.

1 Introduction

Despite the high dimensionality of the configuration space, many human motion activities lie intrinsically on low dimensional manifolds. For example, the shape of the human silhouette through a walking cycle is an example of a dynamic shape where the shape deforms over time based on the action performed but it is also a function of the person body style and the view point. Gait is a 1-dimensional manifold embedded in the body configuration space and it is also a 1-dimensional manifold embedded in the visual input space. Similarly, the appearance of a face performing a facial expression is an example of dynamic appearance. Therefore, researchers have tried to exploit the manifold structure implicitly or explicitly in tasks such as tracking and activity recognition. Learning nonlinear deformation manifolds is typically performed in the visual input space or through intermediate representations. For example, Exemplar-based approaches such as [26] 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 of the manifold which can be used to learn nonlinear manifolds as in [5] and in [3].

Data vs. Concept Driven Manifold Embedding: Embedding manifolds to low dimensional spaces provides a way to explicitly model such manifolds. Learning motion manifolds can be achieved through linear subspace approximation (PCA) as in [9]. PCA have been widely used in appearance modeling to discover subspaces for appearance variations and modeling view manifolds as in [16, 15, 2, 6]. Linear subspace analysis can achieve a linear embedding of the motion manifold in a subspace. However, the dimensionality of the subspace depends on the variations in the data and not in the intrinsic dimensionality of the manifold. Nonlinear dimensionality reduction approaches can achieve much lower dimensionality embedding of nonlinear manifolds through changing the metric from the original space to the embedding space based on local structure of the manifold, e.g. [24, 20, 4]. Nonlinear dimensionality reduction has been recently exploited to model motion manifolds for tracking and 3D pose recovery [28, 8, 7, 23]. However, all these approaches (linear and nonlinear) are data-driven, i.e., the visual input is used to model motion manifolds. The resulting embedding is data-driven and therefore the resulting embedded manifolds of different people performing the same action will be quite different.

To explain our point, let us consider the gait case. Basically, the gait is a 1-dimensional closed loop, embedded in the visual input space, that twists differently depending on the view point, the body shape, self occlusion, clothing, etc. Therefore, embedded manifolds for different people walking from the same view point will be different. The same if we consider manifolds for different views of the same walking person. This was shown in [7, 8] where LLE [20] was used to obtain the embedding. These variations pose a challenge if we would like to use motion manifolds as constraints for the motion, for example in tracking or for body pose recovery. But, conceptually all these manifolds are the same. They are all topologically equivalent, i.e., homeomorphic to each other and we can establish a bijection between any pair of them. They are all also homeomorphic to the gait manifold in a kinematic 3D body configuration space. So, the question we try to address is: given conceptual knowledge about the topology of the manifold, how can we use such knowledge in modeling real motion manifolds

with different sources of variability such as different people, different views, etc. ?

Generative vs. Discriminative Models Several approaches have been introduced in the literature to directly infer 3D body pose as a learned function from the visual input [11, 3, 19, 18, 14, 22, 1]. Such approaches, as well as the one introduced here, have great potentials in solving the fundamental initialization problem for model-based vision as well as in recovering from tracker failures. However, almost all these approaches are discriminative approaches where the mapping is learned from the visual input to 3D or other intermediate representations. In contrast, in [7, 23] manifolds are learned in a generative fashion, i.e., learn mapping from a learned low dimensional manifold representation into the visual input. We argue that learning a generative mapping is advantageous for several reasons. Generative mapping provides means to synthesize the visual input and therefore fits well within a Bayesian tracking framework as an observation model. Mapping from the visual input to 3D poses or view points are not necessarily a function but mapping from a manifold representation to the visual input is a function given that the manifold representation doesn't self intersect which is guaranteed in case conceptual embedding is used, as in this paper.

Contribution: In this paper we consider such classes of human motion which lie on a one dimensional closed manifold such as gait and facial expressions. We introduce a framework to learn decomposable generative models for dynamic shape and dynamic appearance of objects where the motion is constrained to one dimensional closed manifolds while there are other sources of variability such as different views, different people, different classes of motion, etc., all of which are needed to be parameterized. The learned model supports tasks such as synthesis, body configuration recovery, recovery of other aspects such as view, person parameters, etc. As direct and important applications of the introduced framework, we consider the case of gait and also show results for facial expressions. We aim to learn a generative model that can generate walking silhouettes for different people from different view points. Given a single image or a sequence of images, we can use the model to solve for the body configuration, view and person shape style parameters. As a result we can directly infer 3D body pose, view point, and person shape style from the visual input. We also apply the model for facial expressions as an example of a dynamic appearance. In this case we learn a generative model that can generate different dynamic facial expressions for different people. The model can successfully be used to recognize expressions performed by different people never seen in the training.

2 Framework

Our objectives is to learn representations for the shape and/or the appearance of moving (dynamic) objects that supports tasks such as synthesis, pose recovery, view recovery, input reconstruction and tracking. Such learned representation will serve as decomposable generative models for dynamic appearance where 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

$$y_t = T_\alpha \gamma(x_t; a_1, a_2, \cdots, a_n) \quad (1)$$

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. T_α represents a global geometric transformation on the appearance instance. The general form for the mapping function γ that we use is

$$\gamma(x_t; a_1, a_2, \cdots, a_n) = \mathcal{C} \times_1 a_1 \times \cdots \times_n a_n \cdot \psi(x_t) \quad (2)$$

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 [12, 27].

The model in equation 2 is a generalization over the model introduced in [8] where only one factor can be decomposed. The main reason why the model in [8] is limited to decomposing a single factor is that the embedding used was data driven. In that work 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. Besides, in [8] there was no notion of optimal unified manifold representation. These limitations motivate the use of a natural conceptual unified representation of the configuration manifold that is independent of all other factors. Such unified representation would allow the model in equation 2 to generalize to decompose as many factors as desired. In the model in equation 2, 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.

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

$$y_t = \gamma(x_t; v, s) = \mathcal{C} \times v \times s \times \psi(x) \quad (3)$$

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 conceptual representation of the manifold that is homeomorphic to the actual gait manifold.

The question is what conceptual representation of the manifold we can use. 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

$$y_t^a = B^a \psi(x_t^a) \tag{4}$$

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 [27] 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 2 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^*, v^*, s^* which minimize the error in reconstruction

$$E(v, s, x) = || \, y - \mathcal{C} \times v \times s \times \psi(x) \, || \tag{5}$$

or a robust version of the error. We introduce and efficient algorithms to recover these parameters in the case of a single image input or a sequence of images.

3 Conceptual Embedding and Mapping

In this and next sections, for clarity of explanation and without loss of generality, we use the gait example to show the procedure, however, the same solution framework applies to other domains.

Conceptual Manifold Embedding: The input is a set of image sequences each represents a full cycle of the motion, e.g., a full walking cycle captured from different view points. Each image sequence is of certain person and certain view. We assume that the view does not change within any sequences. Each person can have multiple image sequences. The image sequences are not necessarily to be of the same length. We denote each sequence by $Y^{sv} = \{y_1^{sv} \cdots y_{N_{sv}}^{sv}\}$ where v denotes the view class index and s is style index. Let N_v and N_s denote the number of views and number of styles respectively, i.e., there are $N_s \times N_v$ sequences. Each sequence is temporally embedded at equidistance on a unit circle such that $x_i^{sv} = [cos(2\pi i/N_{sv} + \delta^{sv}) \;\; sin(2\pi i/N_{sv} + \delta^{sv})], i = 1 \cdots N_{sv}$ where the displacement parameter δ is used to align all the embedded sequences. Notice that by temporal embedding on a unit circle we do not preserve the metric in input space. Rather, we preserve the topology of the manifold.

Manifold Mapping: Given a set of distinctive representative and arbitrary points $\{z_i \in R^2, i = 1 \cdots N\}$ we can define an empirical kernel map[21] as $\psi_N(x) : \mathbb{R}^2 \rightarrow \mathbb{R}^N$ where

$$\psi_N(x) = [\phi(x, z_1), \cdots, \phi(x, z_N)]^\mathsf{T}, \tag{6}$$

given a kernel function $\phi(\cdot)$. For each input sequence Y^{sv} and its embedding X^{sv} we can learn a nonlinear mapping function $f^{sv}(x)$ that satisfies $f^{sv}(x_i) = y_i, i = 1 \cdots N_{sv}$ and minimizes a regularized risk criteria. From the representer theorem, such function admits a representation of the form

$$f(x) = \sum_{i=1}^{N} w_i \phi(x, z_i),$$

i.e., the whole mapping can be written as

$$f^{sv}(x) = B^{sv} \cdot \psi(x) \tag{7}$$

where B is a $d \times N$ coefficient matrix. If radial symmetric kernel function is used, we can think of equation 7 as a typical Generalized Radial basis function (GRBF) interpolation [17] where each row in the matrix B represents the interpolation coefficients for corresponding element in the input. i.e., we have d simultaneous interpolation functions each from 2D to 1D. The mapping coefficients can be obtained by solving the linear system

$$[y_1^{sv} \cdots y_{N_{sv}}^{sv}] = B^{sv}[\psi(x_1^{sv}) \cdots \psi(x_{N_{sv}}^{sv})]$$

Where the left hand side is a $d \times N_{sv}$ matrix formed by stacking the images of sequence sv column wise and the right hand side matrix is an $N \times N_{sv}$ matrix formed by stacking kernel mapped vectors

To align the sequences we use the model learned for a prototype cycle as a reference. Given a prototype cycle coefficients B^*, any new cycle embedding coordinate is

aligned to it by searching for the displacement parameter δ that minimizes the reconstruction error

$$E(\delta) = \sum_i \|y_i - B^* \cdot \psi(x_i(\delta))\|$$

Decomposition:

Multilinear tensor analysis decomposes multiple orthogonal factors as an extension of principal component analysis (PCA) (one orthogonal factor), and bilinear model (two orthogonal factors) [25]. Multilinear tensor analysis can be achieved by higher-order singular value decomposition (HOSVD), which is a generalization of SVD [12, 27].

Each of the coefficient matrices $B^{sv} = [b_1 b_2 \cdots b_N]$ can be represented as a coefficient vector b^{sv} by column stacking (stacking its columns above each other to form a vector) Therefore, b^{sv} is an $N_c = d \cdot N$ dimensional vector. All the coefficient vectors can then be arranged in an order-three gait coefficient tensor $\mathcal{B}$ with dimensionality $N_s \times N_v \times N_c$. The coefficient tensor is then decomposed as

$$\mathcal{B} = \tilde{\mathcal{A}} \times_1 \tilde{S} \times_2 \tilde{V} \times_3 \tilde{F}$$

where $\tilde{S}$ is the mode-1 basis of $\mathcal{B}$, which represents the orthogonal basis for the style space. Similarly, $\tilde{V}$ is the mode-2 basis representing the orthogonal basis of the view space and $\tilde{F}$ represents the basis for the mapping coefficient space. The dimensionality of these matrices are $N_s \times N_s$, $N_v \times N_v$, $N_c \times N_c$ for $\tilde{S}$,$\tilde{V}$ and $\tilde{F}$ respectively. $\mathcal{A}$ is a core tensor, with dimensionality $N_s \times N_v \times N_c$ which governs the interactions among different mode basis matrices.

Similar to PCA, it is desired to reduce the dimensionality for each of the orthogonal spaces to retain a subspace representation. This can be achieved by applying higher-order orthogonal iteration for dimensionality reduction [13, 27]. Final subspace representation is

$$\mathcal{B} = \mathcal{A} \times_1 S \times_2 V \times_3 F \tag{8}$$

where the reduced dimensionality for $\mathcal{A}$, S, V, and F are $n_s \times n_v \times n_c$, $N_s \times n_s$, $N_v \times n_v$, and $N_c \times n_c$ where n_s, n_v and n_c are the number of basis retained for each factor respectively. Using tensor multiplication we can obtain coefficient eigenmodes which is a new core tensor formed by $\mathcal{Z} = \mathcal{A} \times_3 F$ with dimension $n_s \times n_v \times N_c$.

Given this decomposition and given any n_s dimensional style vector s and any n_v dimensional view vector v we can generate coefficient matrix B^{sv} by unstacking the vector b^{sv} obtained by tensor product $b^{sv} = \mathcal{Z} \times_1 s \times_2 v$. Therefore we can generate any specific instant of the motion by specifying the body configuration parameter x_t through the kernel map defined in equation 6. Therefore, the whole model for generating image y_x^{sv} can be expressed as

$$y_t^{sv} = \textit{unstacking}(\mathcal{Z} \times_1 s \times_2 v) \cdot \psi(x_t)$$

This can be expressed abstractly also in the form of equation 3 by arranging the tensor $\mathcal{Z}$ into a order-four tensor $\mathcal{C}$ with dimensionality $d \times n_s \times n_v \times N$.

4 Parameter Estimation

Given a model fitted as described in the previous section and given a new image or a sequence of images, it is desired to efficiently solve for each of the orthogonal factors as well as body configuration. We discriminate here between two cases: 1: Input is a whole motion cycle. 2: Input is a single image. For the first case we can obtain a closed form analytical solution for each of orthogonal factors by aligning the input sequence manifold to the model conceptual manifold representation. For the second case we introduce an iterative solution.

4.1 Solving View and Style Given a Whole Sequences

Given a sequence of images representing a whole motion cycle, we can solve for the view, v, and shape style, s. First the sequence is embedded to a unit circle and aligned to the model as described in section 3. Then, mapping coefficients B is learned from the aligned embedding to the input. Given such coefficients, we need to find the optimal s and v factors which can generate such coefficients given the learned model. i.e., we need to find s and v which minimizes the error

$$E(s, v) = \|b - \mathcal{Z} \times_1 s \times_2 v\| \tag{9}$$

where b is the column stacking of B. If the style vector, s is known we can obtain a closed form solution for v. This can be achieved by evaluating the product $\mathcal{G} = \mathcal{Z} \times_1 s$ to obtain tensor $\mathcal{G}$. Solution for b can be obtained by solving the system $b = \mathcal{G} \times_2 v$ for v which can be written as a typical linear system by unfolding $\mathcal{G}$ as a matrix. Therefore estimate of v can be obtained by

$$v = (\mathcal{G}_2)^+ b \tag{10}$$

where $\mathcal{G}_2$ is the matrix obtained by mode-2 unfolding of $\mathcal{G}$ and $+$ denotes the psuedo inverse.

Similarly we can analytically solve for s if the view, v, is known by forming a tensor $\mathcal{H} = \mathcal{Z} \times_2 v$ and therefore

$$s = (\mathcal{H}_1)^+ b \tag{11}$$

where $\mathcal{H}_1$ is the matrix obtained by mode-1 unfolding of $\mathcal{H}$

Iterative estimation of v and s using equations 10 and 11 would lead to a local minima for the error in 9. Practically, it was found that starting with a mean style estimate $\tilde{s}$ we can obtain almost correct solution for v. Since the view classes are discrete, we can find the closest view class and use it to estimate s.

4.2 Solving for Body Configuration, View and Style From a Single Image

In this case the input is a single image and it is desired to estimate body configuration and each of the decomposable factors. For the gait case, given an input image y, we need to estimate body configuration, x , view v, and person shape style s which minimize the reconstruction error $E(x, v, s)$

$$E(x, v, s) = || \, y - \mathcal{C} \times v \times s \times \psi(x) \, || \qquad (12)$$

We can instead use a robust error metric and in both cases we end up with a nonlinear optimization problem.

We assume optimal style can be written as a linear combination of style classes in the training data. i.e., we need to solve for linear regression weights α such that $s = \sum_{k=1}^{K_s} \alpha_k s^k$ where each s^k is a mean of one of K_s style classes in the training data. Similarly for the view, we need to solve for weights β such that $v = \sum_{k=1}^{K_v} \beta_k v^k$ where each v^k is a mean of one of K_v view classes.

If the style and view factors are known, then equation 12 reduced to a nonlinear 1-dimensional search problem for, body configuration x on the unit circle that minimizes the error. On the other hand, if the body configuration and style factor are known, we can obtain view conditional class probabilities $p(v^k|y, x, s)$ which is proportional to observation likelihood $p(y \mid x, s, v^k)$. Such likelihood can be estimated assuming a Gaussian density centered around $\mathcal{C} \times v^k \times s \times \psi(x)$, i.e.,

$$p(y \mid x, s, v^k) \approx \mathcal{N}(\mathcal{C} \times v^k \times s \times \psi(x), \Sigma^{v^k}).$$

Given view class probabilities we can set the weights to $\beta_k = p(v^k \mid y, x, s)$. Similarly, if the body configuration and view factor are known, we can obtain style weights by evaluating image likelihood given each style class s^k assuming a Gaussian density centered at $C \times v \times s^k \times \psi(x)$.

This setting favors an iterative procedures for solving for x, v, s. However, wrong estimation of any of the factors would lead to wrong estimation of the others and leads to a local minima. For example wrong estimation of the view factor would lead to a totally wrong estimate of body configuration and therefore wrong estimate for shape style. To avoid this we use a deterministic annealing like procedure where in the beginning the view weights and style weights are forced to be close to uniform weights to avoid hard decisions about view and style classes. The weights are gradually become discriminative thereafter. To achieve this, we use a variable view and style class variances which are uniform to all classes and are defined as $\Sigma^v = T_v \sigma_v^2 I$ and $\Sigma^s = T_s \sigma_s^2 I$ respectively. The parameters T_v and T_s start with large values and are gradually reduced and in each step and a new body configuration estimate is computed.

We summarize the solution framework as follows

Input: image y, view class means v^k, style class means s^k, core tensor $\mathcal{C}$

Initialization :

- initialize T_v and T_s
- initialize α and β to uniform weights
- Compute initial $s = \sum_{k=1}^{K_s} \alpha_k s^k$
- Compute initial $v = \sum_{k=1}^{K_v} \beta_k v^k$

Iterate :

- Compute coefficient $B = \mathcal{C} \times s \times v$
- Estimate body configuration: 1-D search for x that minimizes $E(x) = ||y - B\psi(x)||$
- estimate new view factor
  - Compute $p(y|x, s, v^k)$
  - Update view weights $\beta_k = p(v^k|y, x, s)$
  - Estimate new v as $v = \sum_{k=1}^{K_v} \beta_k v^k$
- Update coefficient $B = \mathcal{C} \times s \times v$
- Estimate body configuration: 1-D search for x that minimizes $E(x) = ||y - B\psi(x)||$
- estimate new style factor
  - Compute $p(y|x, s^k, v)$
  - Update style weights $\alpha_k = p(s^k|y, x, v)$
  - Estimate new s as $s = \sum_{k=1}^{K_s} \alpha_k s^k$
- reduce T_v, T_s

One important aspect that need to be mentioned for the special case of gait is that there is a high similarity between silhouette shapes in each of the half cycles for certain views. In fact, if orthographic projection is used, side view silhouettes will look identical in both halves of the walking cycle. But since perspective imaging is actually used, there is slight differences in silhouette shapes between the two half cycles which are enough to discriminate body configuration throughout the cycle. However, such similarity can cause a confusion in estimating x, s, v. This motivates a modification of the above algorithm for the spacial case of gait where we use dual hypotheses for body configuration and view and style factors. At initialization we solve for body configuration x given the mean style and mean view factors then we initializes dual body configuration hypotheses as x and its antipodal point on the circle which we call $\tilde{x}$. The iterations above proceed with two sets of estimates (x, s, v) and $(\tilde{x}, \tilde{s}, \tilde{v})$. The two sets typically either converge to the same solution or they diverge to two antipodal body configurations where one of them will lead to less error.

5 Experimental Results

5.1 Dynamic Shape Example: Gait Analysis

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 [10] which contains walking people from multiple synchronized views [1]. 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 $\times$ 5 cycles $\times$ 4 views. Note that cycles of different people and cycles of the same person are not of the same length. Figure 1-a,b show examples of the sequences (only half cycles are shown because of limited space).

The data is used to fit the model as described in section 3. Images are normalized to 60×100, i.e., $d = 6000$.

[1] CMU Mobo gait data set [10] contains 25 people, about 8 to 11 walking cycles each captured from six different view points. The walkers were using a treadmill.

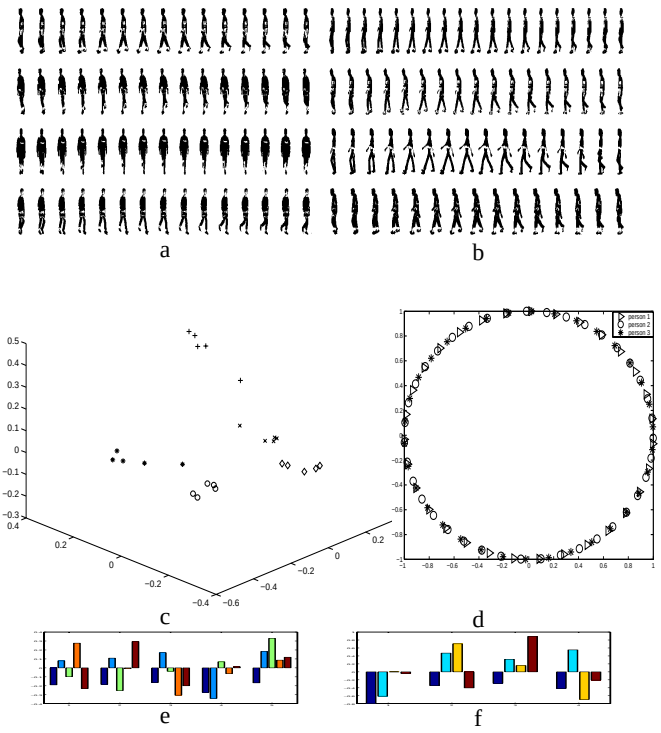

Figure 1. a,b) Example of training data. Each sequence shows a half cycle only. a) four different views used for person 1 b) side views of people 2,3,4,5. c) style subspace: each person cycles have the same label. d) unit circle embedding for three cycles. e) Mean style vectors for each person cluster.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

Each cycle is considered to be a style by itself, i.e., there are 25 styles and 4 views. Therefore, $N_s = 25$, $N_v = 4$. 18 equidistance points on the unit circle are used to obtain the kernel map space defined in equation 6, i.e., $N_c = 6000 \times 18$. After coefficient decomposition and dimensionality reduction as in equation 8 the dimensionality for $\mathcal{A}, S, V, F$ are $5 \times 4 \times 120$, 25×5, 4×4, $(18 \times 6000) \times 120$ respectively. Figure 1-d shows example of model-based aligned unit circle embedding of three cycles. Figure 1-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 1-e shows the mean style vectors for each of the five clusters. Figure 1-f shows the four view vectors.

Evaluation Experiment 1: In this experiment we used the learned model given the training data described above to evaluate the recovery of body configuration, view, and person shape style given test data of the same people in the training but with different cycles not used in the training. We used two new cycles for each of the five people from the four views, i.e., 40 cycles with a total of 1344 frames in

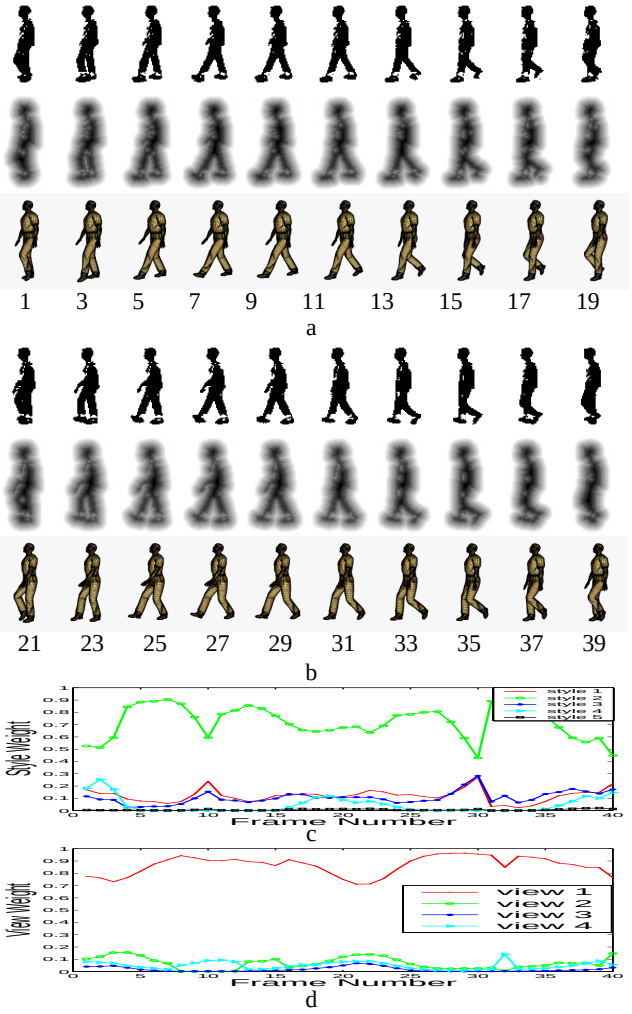

Figure 2. a,b) example pose recovery. from top to bottom: input shapes, implicit function, recovered 3D pose. c) Style weights. d) View weights.

all the test sequences. if we use a whole cycle for recovery of view and person parameter as described in 4.1 we obtain 100% view classification. For style classification we get 36 out of 40 correct classification using nearest style mean and 40 out of 40 using nearest neighbor. If we use single frames for recovery, as described in section 4.2, we get 7 frame errors among 1344 test frames for body configuration and style estimation, i.e., 99.5% accuracy with 100% correct view estimation [2]

[2] a body configuration is considered an error if the distance between correct and estimated embedding is more than $\pi/8$ which is about 2 to 4 frame distance in the original sequence.

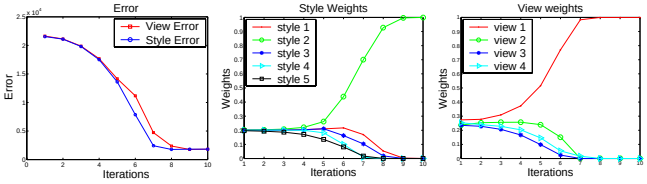

Figure 3. Iterations for frame 5 from above. Left: Error. Center: style weights. Right: View weights

Figure 4. Examples of pose recovery and view classification for three different people from three views.

Figure 2 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 4. Figure 2-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 3 visualizes the progress of the error, style weights, view weights through the iterations used to obtain the results for frame 5. As can be noticed, the weights start uniformly and then smoothly home to the correct style and view as the error is reduced and the correct body configuration is recovered.

Evaluation Experiment 2: In this experiment we used the learned model to evaluate the recovery of body configuration and view given test data of people which have not seen before in the training. We used 8 people sequences, 2 cycles each, from 4 views where none of these people were used in the training. Overall there are 2476 frames in the test sequences. The recovery of the parameters was done on a single frame basis as described in section 4.2. We obtained 111 errors in the recovery of the body configuration, i.e., body configuration accuracy is 95.52%. For view estimation we get 7 frame errors, i.e., view estimation accuracy 99.72%. This result shows that the model generalizes and we can recover the view and body configuration with very high accuracy for unseen people. Figure 4 shows examples recovery of the 3D pose and view class for four different people non of them was seen in training. More examples can be seen in the attached video clips.

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 3. A model in the form of equation 2 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.

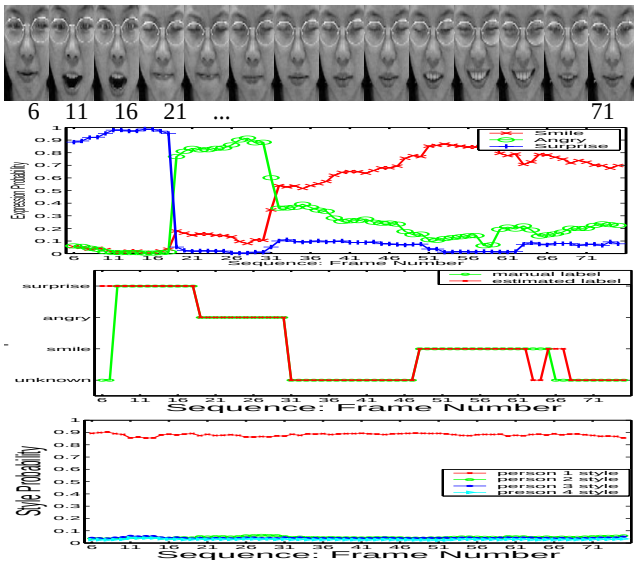

Figure 5. From top to bottom: Samples of the input sequences; Expression probabilities; Expression classification; Style probabilities

We used the learned model to recognize facial expression, and person identity at each frame of the whole sequence. Figure 5 shows an example of a whole sequence and the different expression probabilities obtained on a frame per frame basis using the algorithm described in section 4.2. The figure also shows the final expression recognition after thresholding along manual expression labeling. We used the learned model to recognize facial expressions for sequences of people not used in the training. Figure 6 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.

6 Conclusion

In this paper we presented a framework for learning a decomposable generative model for dynamic shape and dynamic appearance where the intrinsic motion lies on a closed 1D manifold which, in such case, is homeomorphic to a unit circle. Conceptual manifold embedding on a unit circle has many advantages. Fundamentally, this allows modeling any variations (twists) of the manifold given any number factors such as different people, different views, etc. since all resulting manifolds are still topologically equiva-

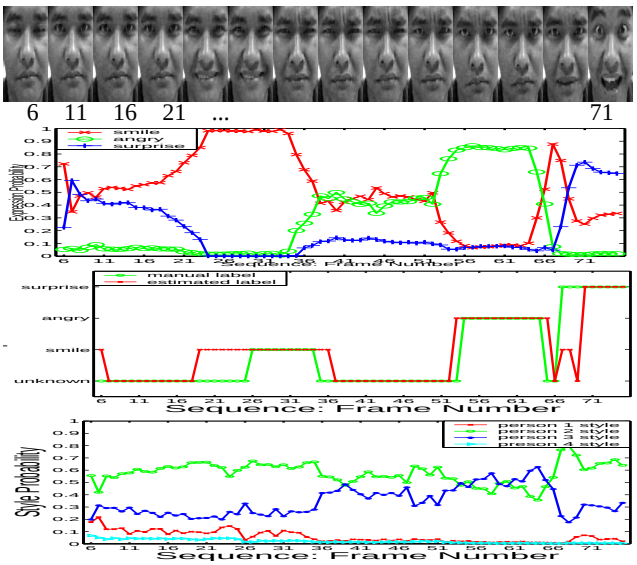

Figure 6. Generalization to new people: expression recognition for a new person. From top to bottom: Samples of the input sequences; Expression probabilities; Expression classification; Style probabilities

lent to the unit circle. This is not achievable if data-driven embedding is used. Another advantage of conceptual embedding is that we only need one cycle of data to learn the manifold while any data-driven embedding would require several cycles to achieve a reasonable embedding. For the case of gait we used temporal information to embed the data which, in this case, provides a straight forward dynamic model for tracking. The use of a generative model is tied to the use of conceptual embedding 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.

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

References

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

[2] 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.

[3] M. Brand. Shadow puppetry. In *Proc. ICCV*, pages 1237–1244, 1999.

[4] 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.

[5] C. Bregler and S. M. Omohundro. Nonlinear manifold learning for visual speech recognition. In *Proc. ICCV*, pages 494–499, 1995.

[6] 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.

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

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

[9] R. Fablet and M. J. Black. Automatic detection and tracking of human motion with a view-based representation. In *Proc. ECCV, LNCS 2350*, pages 476–491, 2002.

[10] R. Gross and J. Shi. The cmu motion of body (mobo) database. Technical Report TR-01-18, Carnegie Mellon University, 2001.

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

[12] 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.

[13] L. D. Lathauwer, B. de Moor, and J. Vandewalle. On the best rank-1 and rank-(r1, r2, ..., rn) approximation of higher-order tensors. *SIAM Journal On Matrix Analysis and Applications*, 21(4):1324–1342, 2000.

[14] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *Proc. ECCV*, pages 666–680, 2002.

[15] M.Turk and A.Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[16] H. Murase and S. Nayar. Visual learning and recognition of 3d objects from appearance. *IJCV*, 14:5–24, 1995.

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

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

[19] 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.

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

[21] B. Scholkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. Cambridge, Massachusetts: The MIT Press, 2002.

[22] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *Proc. ICCV*, pages 750–759, 2003.

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

[24] J. Tenenbaum. Mapping a manifold of perceptual observations. In *Proc. NIPS*, volume 10, pages 682–688, 1998.

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

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

[27] M. A. O. Vasilescu and D. Terzopoulos. Multilinear subspace analysis of image ensembles. In *Proc. CVPR*, pages 93–99, 2003.

[28] Q. Wang, G. Xu, and H. Ai. Learning object intrinsic structure for robust visual tracking. In *Proc. CVPR*, volume 2, pages 227–233, 2003.