Separating Style and Content on a Nonlinear Manifold

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

Abstract

Bilinear and multi-linear models have been successful in decomposing static image ensembles into perceptually orthogonal sources of variations, e.g., separation of style and content. 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. The question that we address in this paper is how to separate style and content on manifolds representing dynamic objects. In this paper we 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 we present in this paper is based on decomposing the style parameters in the space of nonlinear functions which map between a learned unified nonlinear embedding of multiple content manifolds and the visual input space.

1. Problem Statement and Related Work

Linear, Bilinear and Multi-linear Models: Linear models, such as PCA [8], have been widely used in appearance modeling to discover subspaces for appearance variations. For example, PCA has been used extensively for face recognition such as in [13, 1, 6, 10] and to model the appearance and view manifolds for 3D object recognition as in [14]. Such subspace analysis can be further extended to decompose multiple orthogonal factors using bilinear models and multi-linear tensor analysis [19, 22]. The pioneering work of Tenenbaum and Freeman [19] formulated the separation of style and content using a bilinear model framework [11]. In this 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 [11, 12]. In [22] 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 n-mode SVD to fit multi-linear models. Tensor representation of image data was used in [17] for video compression and in [21] for motion analysis and synthesis. N-mode analysis of higher-order tensors was originally proposed and developed in [20, 9, 11] and others.

Nonlinear Manifolds and Decomposition: The applications of Multilinear tensor analysis as in [19, 22] to decompose variations into orthogonal factors are mainly for static image ensembles. The question we address in this paper is how to separate the style and content on a manifold representing a dynamic object. To illustrate our point we consider the human silhouette through the walking cycle (gait) such as shown in figure 1. If we consider such shapes as points in a high dimensional visual input space, then, given the physical body constraints and the temporal constraints imposed by the action being performed, it is expected that these points will lay on a low dimensional manifold. Intuitively, the gait is a 1-dimensional manifold which is embedded and twisted in a high dimensional visual space. Similarly, other human activities such as gesturing, are also low-dimensional manifolds [4, 2]. The question that we address in this paper is how to separate style and content on such manifolds. For example, given several sequences of walking silhouettes, as in figure 1, with different people walking, how to decompose the intrinsic body configuration through the action *(content)* from the appearance (or shape) of the person performing the action *(style)*.

Why don't we just use a bilinear model to decompose the style and content in our case where certain body poses can be denoted as content and different people as style? The answer is that in the case of dynamic (e.g. articulated) objects the resulting visual manifold is nonlinear. This can be illustrated if we consider the walking cycle example in figure 1. In this case, the shape temporally undergoes deformations and self-occlusion which result in the points lying on a nonlinear, twisted manifold. 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

Figure 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 a plot of the distance between the samples. The two dark line parallel to the diagonal shows the similarity between the two half cycles.

nonlinearity, PCA, bilinear, multilinear models will not be able to discover the underlying manifold and decompose orthogonal factors. Simply, linear models will not be able to interpolate intermediate poses and/or intermediate styles.

Another limitations of multilinear analysis, as presented in [19, 22], is that it is mainly a supervised procedure where the image ensemble need to be arranged into various style, content or orthogonal factor classes beforehand. Such requirement makes it hard if we try to use bilinear or multilinear models with image sequences to decompose orthogonal factors on a manifold. Typically, input sequences can be of different lengths, with different sampling rates, and with people performing the same activity with different dynamics. So we aim to have an unsupervised procedure with minimal human interaction.

Nonlinear Dimensionality Reduction and Decomposition of Orthogonal Factors: Recently some promising frameworks for nonlinear dimensionality reduction have been introduced including isometric feature mapping (Isomap) [18], Local linear embedding (LLE) [16]. Related nonlinear dimensionality reduction work also includes [3]. Both Isomap and LLE frameworks were shown to be able to embed nonlinear manifolds into low-dimensional Euclidean spaces for toy examples as well as for real images. 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. As we will show, if we introduce multiple manifolds (corresponding to different styles) to such approaches, they tend to capture the intrinsic structure of each manifold separately without generalizing to capture inter-manifolds aspects. This is because, typically, intra-manifold distances are much smaller than inter-manifold distances. The framework we present in this paper uses nonlinear dimensionality reduction to achieve an embedding of each individual manifold. However, our framework extends such approaches to separate the inter-manifolds style parameter.

Contribution: We introduce a novel framework for separating style and content on manifolds representing dynamic objects. We 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 we present in this paper 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. Decomposable Generative Model

We are given a set of image sequences representing certain motion such as gesture, facial expression or activity. Each sequence is performed by one subject. Given such sequences we aim to learn a decomposable generative model that explicitly decomposes the following two factors:

- 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.

We learn a view-based generative model in the form

$$y_t^s = \gamma(x_t^c; a, b^s) \tag{1}$$

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[1].

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. In this paper we consider mapping functions of the form

$$y_t^s = \gamma_s(x_t) = C^s \cdot \psi(x_t^c) \tag{2}$$

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

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 2, the style can be decomposed in the linear mapping coefficient space using bilinear model in a way similar to [19, 22]. Therefore, input instance y_t can be written as asymmetric bilinear model in the linear mapping space as

$$y_t = \mathcal{A} \times_3 b^s \times_2 \psi(x_t^c) \tag{3}$$

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

$$y_t = \mathcal{A} \times_3 b^{people} \times_2 \psi(x_t^{pose}) \tag{4}$$

In the following sections we will describe the details for fitting such model. Section 3 describes how to obtain a unified nonlinear embedding of the motion manifold. Section 4 describes how to learn nonlinear mappings in the form of equation 2 and 3. Section 5 describes how to solve for both the content and the style given an observation.

3. Unified Embedding

3.1. Nonlinear Dimensionality Reduction

We adapt an LLE framework [16]. Given the assumption that each data point and its neighbors lie on a locally linear patch of the manifold [16], each point (shape or appearance instance) $y_i, i = 1, \cdots, N$ can be reconstructed based on a linear mapping $\sum_j w_{ij} y_j$ that weights its neighbors contributions using the weights w_{ij}. In our case, the neighborhood of each point is determined by its K nearest neighbors based on the distance in the input space (no temporal information was used to define such neighbors). The objective is to find such weights that minimize the global reconstruction error, $E(W) = \sum_i |y_i - \sum_j w_{ij} y_j|^2 \quad i, j = 1 \cdots N$, under certain constraints. Optimal solution for such optimization problem can be found by solving a least-squares problem as was shown in[16].

Since the recovered weights W reflect the intrinsic geometric structure of the manifold, an embedded manifold in a low dimensional space can be constructed using the same weights. This can be achieved by solving for a set of points $X = \{x_i \in R^e, i = 1 \cdots N\}$ in a low dimension space, $e \ll d$, that minimize $E(X) = \sum_i |x_i - \sum_j w_{ij} x_j|^2 \quad i, j = 1 \cdots N$, where in this case the weights are fixed. Solving such problem can be achieved by solving an eigenvector problem as was shown in [16].

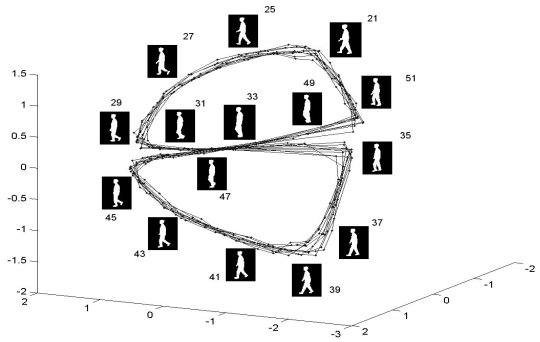


Figure 2. Embedded gait manifold for a side view of the walker. Sample frames from a walking cycle along the manifold with the frame numbers shown to indicate the order. Ten walking cycles are shown (300 frames).

Figure 2 shows an example of embedding a walking cycle with 300 frames from a side view. We use a three dimensional embedding since this is the least dimensional embedding that can discriminate the different body poses through the cycle. As can be noticed, the embedding can discriminate the two half cycles although the similarity (e.g., notice that frames 25 and 39 are embedded as the farthest points on the manifold). More results for embedding the gait manifold can be obtained from [7]. One point that need to be emphasized is that we do not use the temporal relation to achieve the embedding, since the goal is to obtain an embedding that preserves the geometry of the manifold. Temporal relation can be used to determine the neighborhood of each shape but that was found to lead to erroneous, artificial embedding.

3.2. Embedding Multiple Manifolds

Given sequences for multiple people, we need to obtain a unified embedding for the underlying body configuration manifold. Nonlinear dimensionality reduction approaches such as [18, 16, 3] are not able to embed multiple people manifolds simultaneously. Although such approaches try to capture the manifold geometry, typically, the distances between instances of the same person (within the same manifold) is much smaller than distances between corresponding points on different people's manifolds. Therefore, they tend to capture the intrinsic structure of each manifold separately without generalizing to capture inter-manifolds aspects. This is shown in figure 3-a where LLE is used to embed three people's manifolds where all the inputs are spatially registered. As a result, the embedding shows separate manifolds (e.g., in the left figure one manifold is degenerate to a point because the embedding is dominated by the manifold with largest intra-distance.) Even if we force LLE to include corresponding points on different manifolds to each point's neighbors, this result in artificial embedding that does not capture the manifold geometry. Another fun-

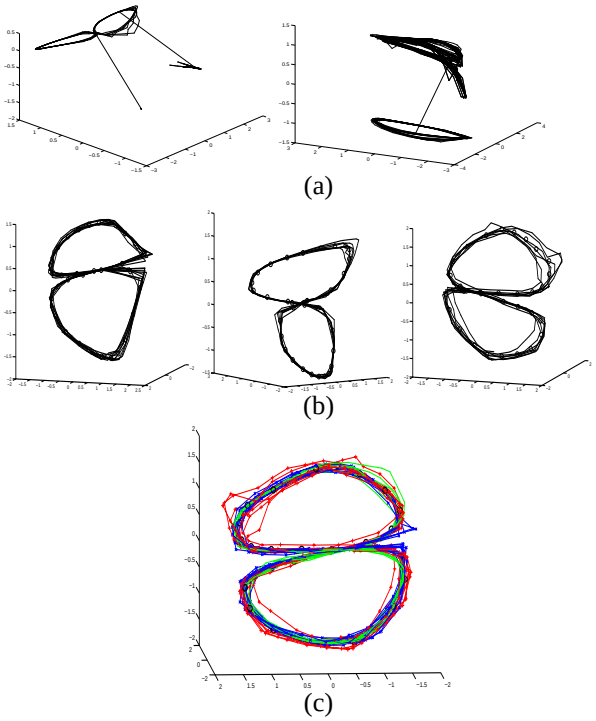


Figure 3. a) Embedding obtained by LLE for three people data with two different K values. Inter-manifold distance dominates the embedding. b) Separate embedding of three manifolds for three people data. c) Unified manifold embedding $\tilde{X}^k$

damental problem is that different people will have different manifolds because the appearance (shape) is different, which imposes different twists to the manifolds and therefore different geometry. This can be noticed in figure 7-b.

To achieve a unified embedding of a certain activity manifold from multiple people data, each person's manifold is embedded separately using LLE. Each manifold point is time mapped from 0 to 1. For the case of periodic motion, such as gait, each cycle on the manifold is time warped from 0 to 1 given a corresponding origin point on the manifold, denoted by t_o, where the cycles can be computed from the geodesic distances to the origin. Given the embedded manifold X^k for person k, a cubic spline $m^k(t)$ is fitted to the manifold as a function of time, i.e., $m^k(t) : t \rightarrow \mathbb{R}^e$ where $t = 0 \rightarrow 1$ is the time variable. The manifold for person k is sampled at N uniform time instances $m^k(t_i)$ where $i = 1 \cdots N$.

Given multiple manifolds a mean manifold $Z(t_i)$ is learned by warping $m^k(t_i)$ using non-rigid transformation using an approach similar to [5]. We solve for a mean manifold $Z(t_i)$ and a set of non-rigid transformations $f(.; \alpha_k)$ where the objective is to minimize the energy function

$$E(f) = \sum_k \sum_i \|Z(t_i) - f(m^k(t_i); \alpha_k)\|^2 + \lambda \|Lf\|^2$$

where λ is a regularization parameter and $\|Lf\|^2$ is a smoothness term. In particular thin-plate spline (TPS) is used for the nonrigid transformation. Given the transformation parameters α_k, the whole data sets are warped to obtain a unified imbedding $\tilde{X}^k$ for the k manifolds where

$$\tilde{X}^k = f(X^k; \alpha_k), k = 1 \cdots K.$$

Figure 3-b,c shows an example of three different manifolds and their warping into a unified manifold embedding.

4. Nonlinear Mapping

4.1. Learning Style Dependent Mappings

Let the sets of input image sequences be $\mathsf{Y}^k = \{y_i^k \in R^d \quad i = 1, \cdots, N_k\}$ and let their corresponding points on the unified embedding space be $\mathsf{X}^k = \{x_i^k \in R^e, \quad i = 1, \cdots, N_k\}$ where e is the dimensionality of the embedding space (e.g. $e = 3$ in the case of gait) and $k = 1 \cdots K$ is the person (style) index. Let the set of N centers representing the mean manifold be $Z = \{z_j \in R^e, j = 1, \cdots, N\}$. We can learn nonlinear mappings between the centers Z and each of the input sequence using generalized radial basis function interpolation GRBF [15], i.e., one mapping for each style class k.

Let's consider the case for k-th sequence. We will drop the index k when it is implied from the context for simplicity. We can solve for multiple interpolants $f^l : R^e \rightarrow R$ where l is l-th dimension (pixel) in the input space and f^l 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

$$f^l(x) = p^l(x) + \sum_{i=1}^{N} w_j^l \phi(|x - z_j|), \tag{5}$$

where $\phi(\cdot)$ is a real-valued basic function, w_j are real coefficients, $|\cdot|$ is the norm on R^e (the embedding space). Typical choices for the basis function include thin-plate spline ($\phi(u) = u^2 log(u)$), the multiquadric ($\phi(u) = \sqrt{u^2 + a^2}$), Gaussian ($\phi(u) = e^{-au^2}$), biharmonic ($\phi(u) = u$) and triharmonic ($\phi(u) = u^3$) splines. p^l is a linear polynomial with coefficients c^l, i.e., $p^l(x) = [1 \quad x^\top] \cdot c^l$. 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

$$f_k(x) = C^k \cdot \psi(x), \tag{6}$$

where C^k is a $d \times (N + e + 1)$ dimensional matrix with the l-th row $[w_1^l \cdots w_N^l \quad c^{l^\top}]$ and the vector $\psi(x)$ is $[\phi(|x - z_1|) \cdots \phi(|x - z_N|) \;\; 1 \;\; x^\top]^\top$. The matrix C^k represents the coefficients for d different nonlinear mappings for style class k, each from a low-dimension embedding space into real numbers. To insure orthogonality and to make the problem well posed, the following side condition constraints are imposed: $\sum_{i=1}^{N} w_i p_j(x_i) = 0, j = 1, \cdots, m$

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

$$\begin{pmatrix} A & P_x \\ P_t^\top & 0_{(e+1)\times(e+1)} \end{pmatrix}_k C^{k\top} = \begin{pmatrix} Y_k \\ 0_{(e+1)\times d} \end{pmatrix}, \quad (7)$$

where A, P_x, P_t are defined for the *k*-th style as: A is $N_k \times N$ matrix with $A_{ij} = \phi(|x_i^k - z_j|), \quad i = 1 \cdots N_k, j = 1 \cdots N$, P_x is a $N_k \times (e+1)$ matrix with *i*-th row $[1 \;\; x_i^{k\top}]$, P_t is a $N \times (e+1)$ matrix with *i*-th row $[1 \quad z_i^\top]$. Y_k is $(N_k \times d)$ matrix containing the input images for style k, i.e., $Y_k = [y_1^k \cdots y_{N_k}^k]^\top$. Solution for C^k is guaranteed under certain conditions on the basic functions used.

4.2. Separating Style

Given learned nonlinear mapping coefficients $C^1, C^2, \cdots, C^K$ for each person, the style parameters can be decomposed by fitting an asymmetric bilinear model [19] to the coefficient tensor. Let the coefficients be arranged as a $d \times M \times K$ tensor $\mathcal{C}$, where $M = (N+e+1)$. Therefore, we are looking for a decomposition in the form

$$\mathcal{C} = \mathcal{A}^c \times_3 B^s$$

where $\mathcal{A}^c$ is $d \times M \times J$ tensor containing content bases for the RBF coefficient space and $B^s = [b^1 \cdots b^K]$ is a $J \times K$ style coefficients. This decomposition can be achieved by arranging the mapping coefficients as a $dM \times K$ matrix as

$$C = \begin{pmatrix} c_1^1 & \cdots & c_1^K \\ \vdots & \ddots & \vdots \\ c_M^1 & \cdots & c_M^K \end{pmatrix} \quad (8)$$

where $[c_1^k, \cdots, c_M^k]$ are the columns for RBF coefficients C^k. Given the matrix C style vectors and contents bases can be obtained by singular value decomposition as $C = USV^T$ where the content bases are the columns of US and the style vectors are the rows of V.

5. Solving for Style and Content

Given a new input $y \in R^d$, it is required to find both the content, i.e., the corresponding embedding coordinates $x \in R^e$ on the manifold, and the person style parameters b^s. These parameters should minimize the reconstruction error defined as

$$E(x^c, b^s) = \|y - \mathcal{A} \times b^s \times \psi(x^c)\|^2$$

Solving for content: If the style vector, b^s, is known, we can solve for the content x^c. Note that, in our case, the content is a continuous variable in a nonlinearly embedded space. However, we show here how to obtain a closed-form solution for x^c.

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 $x^* = arg_x min \|y - B\psi(x)\|^2$ where $B = \mathcal{A} \times b^s$. 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 $B = \mathcal{A} \times b^s$. 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 $\psi(x) = V\acute{S}U^T y$ where $\acute{S}$ is the diagonal matrix obtained by taking the inverse of the nonzero singular values in the diagonal matrix S and setting the rest to zeros. Linear approximation for the embedding coordinate x^* can be obtained by taking the last e rows in the recovered vector $\psi(x)$.

Solving for style: If the embedding coordinate (content) is known, we can solve for style vector b^s. Given style classes $b^k, k = 1, \cdots, K$ learned from the training data and given the embedding coordinate x, the observation can be considered as drawn from a Gaussian mixture model centered at $\mathcal{A} \times b^k \times \psi(x)$ for each style class k. Therefore, observation probability $p(y|k, x)$ can be computed as

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

Style conditional class probabilities can be obtained as $p(k|x, y) = p(y|k, x)p(k|x)/p(y|x)$ where $p(y|x) = \sum_k p(y|x, k)p(k)$. A new style vector can then be obtained as a linear combination of the K class style vectors as $b^s = \sum_k w_k b^k$ where the weights w_k are set to be $p(k|x, y)$.

Given the two steps described above we can solve for both style b^s and content x^c in an EM-like iterative procedure where in the E-step we calculate the content x^c given the style parameters and in the M-step we calculate new style parameters b^s given the content. The initial content can be obtained using a mean style vector $\tilde{b}^s$.

6. Experimental Result

Representation Without loss of generality, for the experiments we show here, the following representations were used:

Shape Representation: We represent each shape instance as an implicit function $y(x)$ at each pixel x such that $y(x) = 0$ on the contour, $y(x) > 0$ inside the contour, and $y(x) < 0$ outside the contour. We use a signed-distance function for this purpose. Such representation imposes smoothness on the distance between shapes. Given such representation, the input shapes are points $y_i \in R^d, i = 1, \cdots, N$ where d is the same as the dimensionality of the input space and N

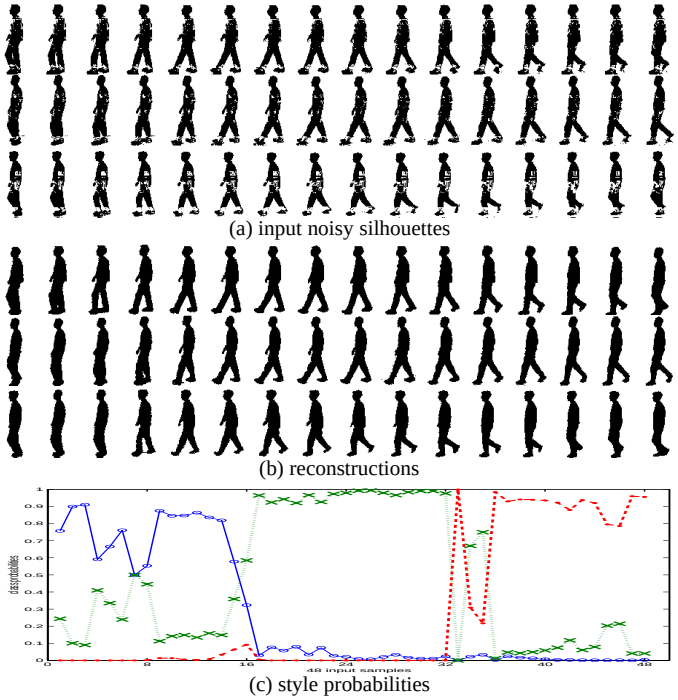
is the number of points. Implicit function representation is typically used in level-set methods.

Appearance Representation: Appearance is represented directly in a vector form, i.e., each instance of appearance is represented as points $y_i \in R^d, i = 1, \cdots, N$ where d is the dimensionality of the input space.

Experiment 1: In this experiment we use three people's silhouettes during a half walking cycle to separate the style (person shape) from the content (body pose). The input is three sequences containing 10, 11, 9 frames respectively. The input silhouettes are shown in figure 4-a. Note that the three sequences are not of equal length and the body poses are not necessarily in correspondence. Since the input size in this case is too small to be able to discover the manifold geometry using LLE, we arbitrary embed the data points on a circle as a topologically homomorphic manifold (as an approximation of the manifold of half a cycle) where each sequence is equally spaced along the circle. Embedding is shown in figure 4-b. We selected 8 RBF centers at 8 quadrics on the circle. The model is then fitted to the data in the form of equation 4 using TPS kernels. Figure 4-d shows the RBF coefficients for the three people (one in each row) where the last three columns are the polynomial coefficients. Figure 4-c shows the style coefficients for the three people and figure 4-e show the content bases.

Given the fitted model we can show some interesting results. First we can interpolate intermediate silhouettes for each of the three people's styles. This is shown in figure 4-f where 16 intermediate poses were rendered. Notice that the input contained only 9 to 11 data points for each person. A closer look at the rendered silhouettes shows that model can really interpolate intermediate silhouettes that were never seen as inputs (e.g., person 1 column 4 and person 3 columns 5, 15). We can also interpolate half walking cycles at new styles. This is shown in figure 4-f where intermediate styles and intermediate contents were used.

We can also use the learned model to reconstruct noisy and corrupted input instances in a way that preserve both the body pose and the person style. Given an input silhouette we solve for both the embedding coordinate and the style as was described in section 5 and use the model to reconstruct a corrected silhouette given the recovered pose and person parameters. Figure 5 shows such reconstruction where we used 48 noisy input silhouettes[2] were used (16 for each person shown at each row). The resulting people's probabilities are shown in figure 5-c and the resulting reconstructions are shown in figure 5-b in the same order. Notice that the reconstruction preserves both the correct body pose as well as the correct person shape. Only two errors can be spotted which are for inputs number 33,34 (last row, columns 2,3) where the probability for person 2 was higher than the person 3 and therefore the reconstruction preserved the second person style. Figure 6 shows another reconstruction example where the learned model was used to reconstruct corrupted inputs for person 3. The reconstruction preserve the person style as well as the body pose.



(a) input noisy silhouettes

(b) reconstructions

(c) style probabilities

Figure 5. Reconstruction example

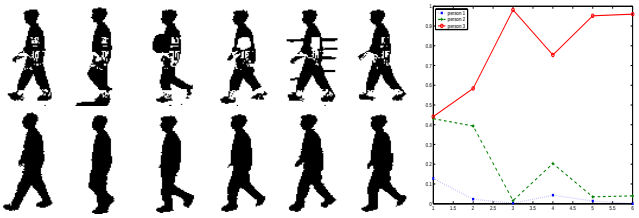


Figure 6. Pose and style preserving reconstruction. Right: style probabilities for each input

Experiment 2: 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 7-b which shows a different manifold for each person. The learned unified manifold is also shown in figure 7-e. Figure 7 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.

[2] All the silhouette data used in these experiments are from the CMU-Mobogait data set

[3] The data are from CMU Mobogait database

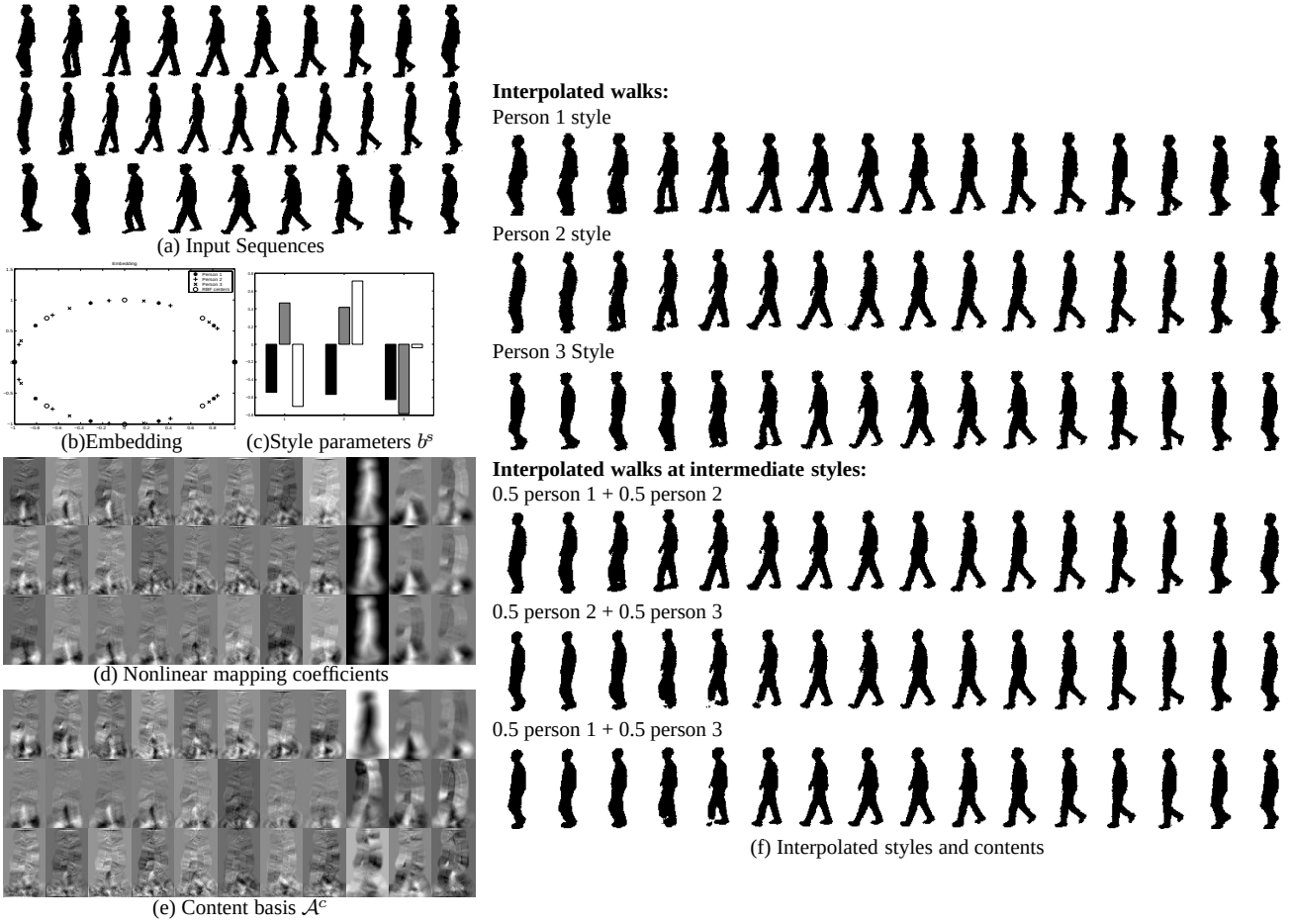


Figure 4. Learning style and content for a gait example

Experiment 3: Leaning a smile: In this experiment the proposed model was used to learn the manifold of a smile and separate the appearance (style) for 4 people[4]. The input sequences contain 27,31,29,27 frames respectively for the smile motion. All the input sequences were temporally scaled from 0 to 1 then LLE were used to obtain a one-dimensional embedding of the manifolds and a unified embedding is obtained as was described in section 3. The model was fitted using 8 equally spaced RBF centers along the mean manifold. The first four rows of figure 8 show interpolation of 10 intermediate faces at each of the learned styles. As can be noticed, the model is able to correctly interpolate the facial motion of the smile for the four people. It is hard to prove in this case that the model is actually interpolating new intermediate faces but we can easily show interpolating smiles at new styles. This is shown in the last three rows where the model is used to render smiles at intermediate styles.

7. Conclusion

We introduced a framework for separating style and content on manifolds representing dynamic objects. 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 in 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 [16], an interesting and important question is how to learn a parametric mapping between the observation and nonlinear embedding spaces. We partially addressed this question in this paper.

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

[4]The images are from the CMU facial expression data set

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

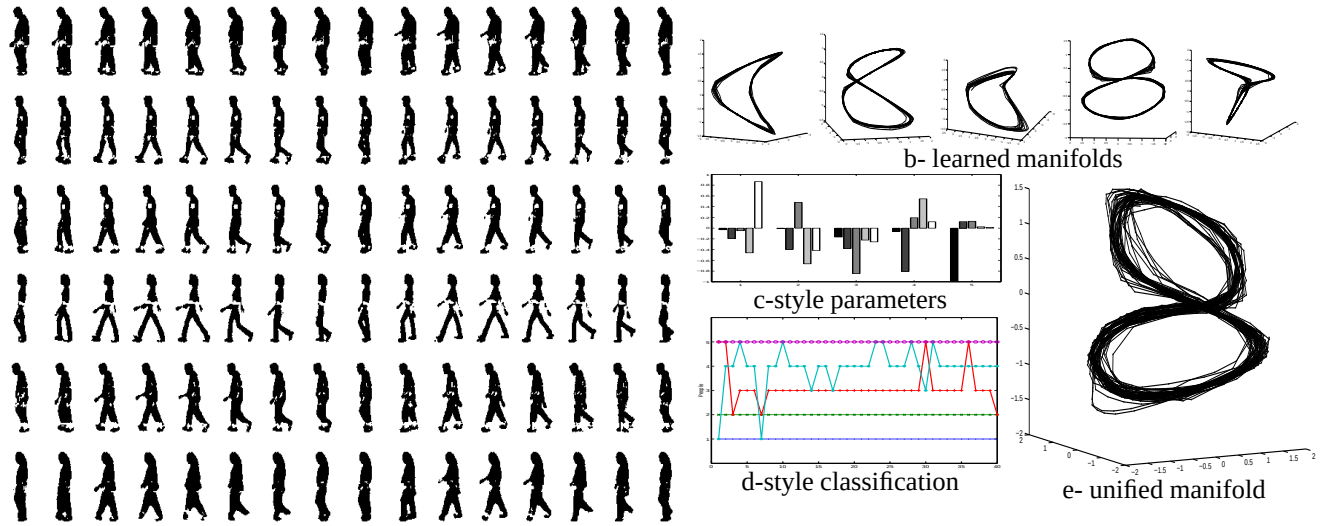


Figure 7. Left: interpolated walks. Last row is interpolated walk at intermediate style between row 1 and 4.

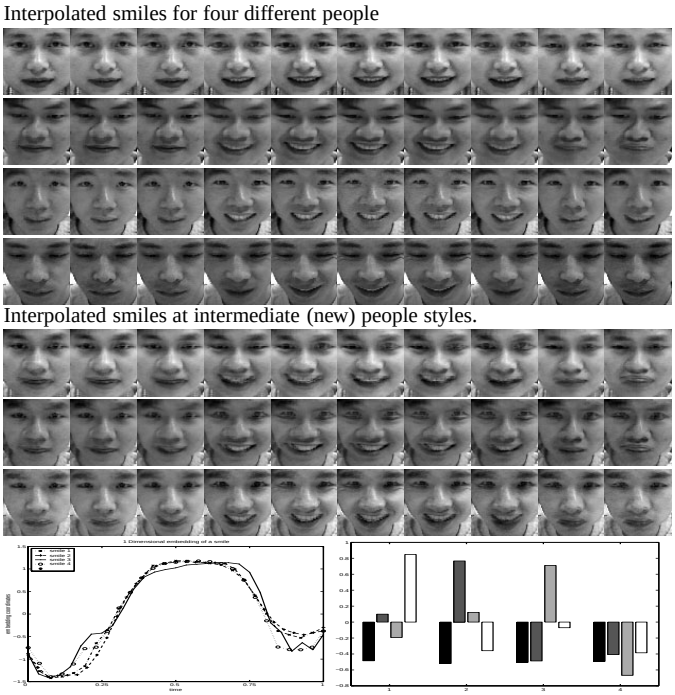


Figure 8. Learning a smile manifold. bottom: manifold embedding and style parameters

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

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

[5] H. Chui and A. Rangarajan. A new algorithm for non-rigid point matching. In *Proc. IEEE CVPR*, pages 44–51, 2000.

[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. of IEEE Conference on Computer Vision and Pattern Recognition*, June-July 2004.

[8] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[9] A. Kapteyn, H. Neudecker, and T. Wansbeek. An approach to n-model component analysis. *Psychometrika*, 51(2):269–275, 1986.

[10] 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.

[11] J. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, New York, New York, 1988.

[12] D. Marimont and B. Wandell. Linear models of surface and illumination spectra. *J. Optical Society od America*, 9:1905–1913, 1992.

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

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

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

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

[17] A. Shashua and A. Levin. Linear image coding of regression and classification using the tensor rank principle. In *Proc. of IEEE CVPR, Hawai*, 2001.

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

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

[20] L. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966.

[21] M. A. O. Vasilescu. An algorithm for extracting human motion signatures. In *Proc. of IEEE CVPR, Hawai*, 2001.

[22] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensebles: Tensorfaces. In *Proc. of ECCV, Copenhagen, Danmark*, 2002.