Style Adaptive Bayesian Tracking Using Explicit Manifold Learning

Chan-Su Lee and Ahmed Elgammal
Computer Science Department
Rutgers, The State University of New Jersey
Piscataway, NJ 08854, USA
Email:{chansu, elgammal}@cs.rutgers.edu

Abstract

Characteristics of the 2D contour shape deformation in human motion contain rich information and can be useful for human identification, gender classification, 3D pose reconstruction and so on. In this paper we introduce a new approach for contour tracking for human motion using an explicit modeling of the motion manifold and learning a decomposable generative model. We use nonlinear dimensionality reduction to embed the motion manifold in a low dimensional configuration space utilizing the constraints imposed by the human motion. Given such embedding, we learn an explicit representation of the manifold, which reduces the problem to a one-dimensional tracking problem and also facilitates linear dynamics on the manifold. We also utilize a generative model through learning a nonlinear mapping between the embedding space and the visual input space, which facilitates capturing global deformation characteristics. The contour tracking problem is formulated as states estimation in the decomposed generative model parameter within a Bayesian tracking framework. The result is a robust, adaptive gait tracking with shape style estimation.

1 Introduction

Vision-based human motion tracking and analysis systems have promising potentials for many applications such as visual surveillance in public area, activity recognition, and sport analysis. Human motion involves not only geometric transformations but also deformations in shape and appearance. Characteristics of the shape deformation in a person motion contain rich information such as body configuration, person identity, gender information, and even emotional states of the person. Human identification [4] and gender classification [5] are examples of the applications. There have been a lot of work on contour tracking such as active shape models (ASM) [3], active contours [10], and exemplar-based tracking [14]. It is still hard to get robust contour tracking and in the same time be able to adapt to different people being tracked with automatic initialization.

Modeling dynamics of shape and appearance is essential for tracking human motion. The observed human shape and appearance in video sequences goes through complicated global nonlinear deformation between frames. If we consider the global shape, there are two factors affecting the shape of the body contour through the motion: *global dynamics factor* and

BMVC 2005 doi:10.5244/C.19.55

person shape style factor, The dynamic factor is constrained because of dynamics of the motion and the physical characteristics of human body configuration [12, 1]. The person shape style is time-invariant factor characterizing distinguishable features in each person shape depending on body built(big, small, short, tall, etc.). These two factors can summarize rich characteristics of human motion and identity.

Our objective is to achieve trackers that can track global deformation in contours and can adapt to different people shapes automatically . There are several challenges to achieve this goal. First, modeling the human body shape space is hard, considering both the dynamics and the shape style. Such shapes lie on a nonlinear manifold. Also, in some cases there are topological changes in contour shapes through motion which makes establishing correspondences between contour points unfeasible. Second, how to simplify the dynamics of the global deformation. Can we learn a dynamic model for body configuration that is low in dimensionality and exhibits linear dynamics? For certain classes of motion like gait, facial expression and gestures, the deformation might lie on a low dimensional manifold if we consider a single person. Our previous work [7] introduced a framework to separate the motion from the style in a generative fashion where the motion is represented in a low dimensional representation.

In this paper we utilize a similar generative model within a Bayesian tracking formulation that is able to track contours in cluttered environment where the tracking is performed in three conceptually independent spaces: body configuration space, shape style space and geometric transformation space. Therefore, object state combines heterogeneous representations. 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? We present a new approach for tracking nonlinear global deformation in human motion based on Bayesian tracking with emphasis on the the gait motion.

Our contributions are as follows: First, by applying explicit nonlinear manifold learning and its parametric representation, we can find compact, low dimensional representation of body configuration which reduces the complexity of the dynamics of the motion into a linear system for tracking gait. Second, by utilizing conceptually independent decomposed state representation, we can achieve approximation of state posterior estimation in a marginalized way. Third, the adaptive tracking of person shape allows not only tracking of any new person contour but also shows potential for identification of the person during tracking. As a result, we achieve robust, adaptive gait tracking with style estimation from cluttered environment.

The paper organization is as follows. Section 2 summarizes the framework. Section 3 describes the state representation and learning. Section 4 describes the the tracking framework. Section 5 shows some experimental results.

2 Framework

We can think of the shape of a dynamic object as instances 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

$$z_t = T_{\alpha_t} \gamma(b_t; s_t), \tag{1}$$

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

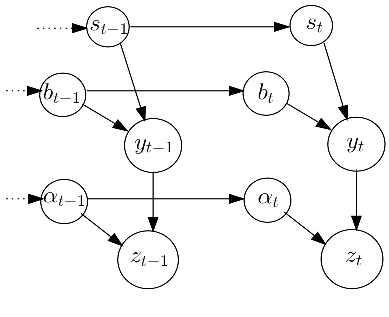


Figure 1: Graphic model for decomposed generative model

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. Figure 1 shows a graphical model illustrating the relation between these variables where y_t is a contour instance generated from model given body configuration b_t and shape style s_t and transformed in the image space through T_{α_t} to form the observed contour. The mapping $\gamma(b_t;s_t)$ is a nonlinear mapping from the body configuration state b_t as

$$y_t = \mathscr{A} \times s^t \times \psi(b_t), \tag{2}$$

where $\psi(b_t)$ is a kernel induced space, $\mathscr{A}$ is a third order tensor, s^k is a shape style vector and × is appropriate tensor product as defined in [11].

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:

$$P(X_t|Z^t) \propto P(Z_t|X_t) \int_{X_{t-1}} P(X_t|X_{t-1})P(X_{t-1}|Z^{t-1}) \tag{3}$$

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.

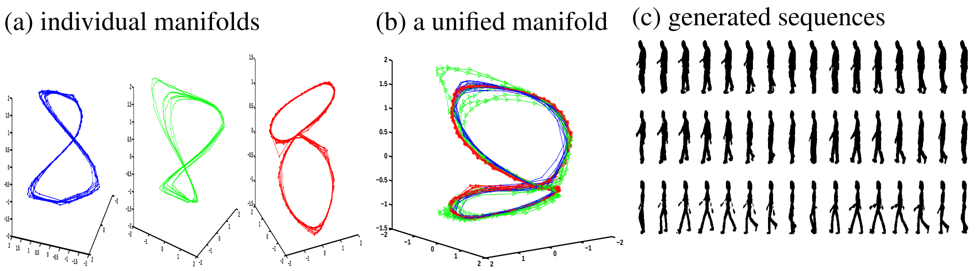

(a) individual manifolds (b) a unified manifold (c) generated sequences

Figure 2: Individual manifolds and their unified manifold

3 Learning Generative Model

Our objective is to establish generative models for the shape in the form of equation 2 where the intrinsic body configuration is decoupled from the shape style. Such model was introduced in [7] and applied to gait data and facial expression data. Given training sequences of different people performing the same motion (gait in our case), Locally Linear Embedding (LLE) [13] is applied to find low dimensional representation of body configuration for each person manifold. As a result of nonlinear dimensionality reduction, an embedding of the gait manifold can be obtained in a low dimensional Euclidean space where three is the least dimensional space where the two halves of the walking cycles can be discriminated. Figure 2-a shows low dimensional representation of side-view walking sequences for different people. Generally, 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 manifold can be used as intrinsic representation of the body configuration. The use of LLE to obtain intrinsic configuration for tracking was previously reported in [15].

Modeling body configuration space: Body configuration manifold is parameterized using a spline fitted to the embedded manifold representation. First, cycles are detected given an origin point on the manifold by computing geodesics along the manifold. Second, a mean-manifold for each person is obtained by averaging difference cycles. Obviously, each person will have a different manifold based on his spatio-temporal characteristics. Third, non-rigid transformation using an approach similar to [2] is performed to find a unified manifold representation as in Fig. 2-b . Correspondences between different subjects are accomplished by selecting a certain body pose as the origin point in different manifolds and equal sampling in the parameterized representation. Finally, we parameterized the unified mean manifold by spline fitting.

The unified mean-manifold can be parameterized by a one-dimensional parameter $\beta_t \in R$ and a spline fitting function $f : R \rightarrow R^3$ that satisfies $b_t = f(\beta_t)$ is used to map from the parameter space into the three dimensional embedding space. Fig. 2-c shows three sequences generated using the same equidistant body configuration parameter $[\beta_1, \beta_2, \cdots, \beta_{16}]$ along the unified mean manifold with different style. As a result of one dimensional representation of each cycle in unified manifold, aligned body pose can be generated in different shape style.

Modeling style shape space: Shape style space is parameterized by a linear combination of basis of the style space. First, RBF nonlinear mappings [7] are learned between the embedded body configuration b_t on the unified manifold and corresponding shape observations y_t^k for each person k. The mapping has the form $y_t^k = \gamma_k(b_t) = C^k \cdot \psi(b_t)$, where C^k is the mapping coefficients which depend on particular person shape and $\psi(\cdot)$ is a nonlinear mapping with N

RBF kernel functions to model the manifold in the embedding space. Given learned nonlinear mapping coefficients $C^1, C^2, \cdots, C^K$, for training people $1, \cdots, K$, the shape style parameters are decomposed by fitting an asymmetric bilinear to the coefficient space. As a result, we can generate contour instance y_t^k for particular person k at any body configuration b_t as

$$y_t^k = \mathscr{A} \times s^k \times \psi(b_t), \tag{4}$$

where $\mathscr{A}$ is a third order tensor, s^k is a shape style vector for person k.

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 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:

$$s = \sum_{k=1}^{K} \lambda^k s^k, \qquad \sum_{k=1}^{K} \lambda^k = 1, \tag{5}$$

where K is the number of style classes used to represent new styles.
Modeling geometric transformation space: Geometric transformation in human motion has constraints. We parameterize the geometric transformation by scaling S_x, S_y and translation T_x, T_y. New representation of any contour point can be found using homogeneous transformation α with paramters $[T_x, T_y, Sx, Sy]$ as global transformation parameters. The overall generative model can be expressed as

$$z_t = T_{\alpha_t}\left(\mathscr{A} \times \left[\sum_{k=1}^{K} \lambda_t^k s^k\right] \times \psi(f(\beta_t))\right). \tag{6}$$

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.

4 Bayesian Tracking

4.1 Marginalized Density Representation

Since it is hard to estimate joint distribution of high dimensional data, we represent each of the sub-state densities by marginalized ones. We approximate the marginal density of each sub-state using maximum a posteriori (MAP) of the other sub-states, i.e.,

$$P(\alpha_t|Z_t) \propto P(\alpha_t|b_t{}^*, s_t{}^*, Z_t), \;\; P(b_t|Z_t) \propto P(b_t|\alpha_t{}^*, s_t{}^*, Z_t), \;\; P(s_t|Z_t) \propto P(s_t|\alpha_t{}^*, b_t{}^*, Z_t), \tag{7}$$

where $\alpha_t{}^*$, $b_t{}^*$, and $s_t{}^*$ are maximum a posteriori estimate of each approximated marginal density. Maximum a posteriori of the joint density $X_t^* = [\alpha_t^*, b_t^*, s_t^*] = \arg\max_{X_t} P(X_t|Z^t)$ will maximize the posterior marginal densities $\alpha_t^* = \arg\max_{\alpha_t} P(\alpha_t|Z^t)$, $b_t^* = \arg\max_{b_t} P(b_t|Z^t)$, and $s_t^* = \arg\max_{s_t} P(s_t|Z^t)$ because of our decomposable generative model.

Dynamic Models	Predicted States
$P(\alpha_t\|\alpha_{t-1}) = N(H\alpha_{t-1}, \sigma_\alpha{}^2)$	$\hat{\alpha}_t^{(i)} \propto P(\alpha_t\|\alpha_{t-1})$
$P(b_t\|b_{t-1}) \propto P(\beta_t\|\beta_{t-1}) = N(\beta_t + \tilde{v}, \sigma_b{}^2)$	$\hat{\beta}_t^{(j)} \propto P(\beta_t\|\beta_{t-1})$
$P(s_t\|s_{t-1}) \propto P(\lambda_t\|\lambda_{t-1}) = N(s_{t-1}, \sigma_{s_t}^2)$	$\hat{\lambda}_t^{(k)} \propto P(\lambda_t\|\lambda_{t-1}) \propto P(s_t\|s_{t-1})$

Table 1: Dynamic models and predicted density of states

4.2 Dynamic Model

The dynamic model $P(X_t|X_{t-1}) = P(\alpha_t, b_t, s_t|\alpha_{t-1}, b_{t-1}, s_{t-1})$ predicts the state X_t at time t given the previous state X_{t-1}. In our generative model, the states are decomposed into three different factors. In the Bayesian network shown in Fig. 1, the transitions of each of the sub-states are independent. Therefore, we can describe whole state dynamic model by dynamic models of individual sub-states as

$$P(\alpha_t, b_t, s_t|\alpha_{t-1}, b_{t-1}, s_{t-1}) = P(\alpha_t|\alpha_{t-1})P(b_t|b_{t-1})P(s_t|s_{t-1}). \tag{8}$$

In case of body configuration, the manifold spline parameter β_t will change in a constant speed if the subject walks in a constant speed (because it corresponds to constant frame rate used in the learning). However, the resulting manifold point representing body configuration $b_t = f(\beta_t)$ will move along the manifold in different step sizes. Therefore, we use β_t to model the dynamics since it results in a one dimensional linear dynamic system. In general, the walking speed can change gradually. So, the body configuration in each new state will move from the current state with a filtered speed that is adaptively estimated during the tracking.

Dynamic model of style is approximated by a random walk as the style may change smoothly around specific person style. The global transformation α_t captures global contour motion in the image space. The left column in table 1 shows the dynamic models, where $\tilde{v}$ is filtered estimated body configuration manifold velocity, and H is a transition matrix of the geometric transformation state which is learned from the training data. A new sub-state densities can be predicted using each sub-state dynamic models as in table 1 right column .

4.3 Tracking Algorithm using Particle Filter

We represent state densities using particle filters since such densities can be non-Gaussian and the observation is nonlinear. In particle filter, the density of posterior $P(X_t|Z^t)$ is approximated by a set of N weighted samples $\{X_t^{(i)}, \pi_t^{(i)}\}$, where $\pi_t^{(i)}$ is the weight for particle $X_t^{(i)}$. Using the generative model in Eq. 6, the tracking problem is to estimate α_t, λ_t, and β_t for given observation Z^t. The marginalized posterior densities for α_t, β_t, and λ_t are approximated by particles.

$$\{\alpha_t^{(i)}, {}^{\alpha}\pi_t^{(i)}\}_{i=1}^{N_\alpha}, \{\beta_t^{(j)}, {}^{b}\pi_t^{(j)}\}_{j=1}^{N_b}, \{\lambda_t^{(k)}, {}^{s}\pi_t^{(k)}\}_{k=1}^{N_s}, \tag{9}$$

where N_α,N_b, and N_s are the numbers of particles used for each of the sub-states. Density representation $P(\beta|Z^t)$ is interchangeable with $P(b_t|Z^t)$ as it has one to one correspondence between $\beta_t^{(j)}$ and $b_t^{(j)}$. The same holds between λ_t and s_t for the style state.

We estimate states by sequential update of the marginalized sub-densities utilizing the predicted densities of the other sub-states. These densities are updated with current observation Z_t by updating weighting values of each sub-state particles approximation using the observation. We estimate global transformation α_t using predicted density $\grave{s}_t$, $\grave{b}_t$. Then body configuration b_t is estimated using estimate global transformation α_t, and predicted density $\grave{s}_t$. Finally style s_t is estimated with given estimated α_t, and b_t. The following table summarizes the state estimation procedure using time $t-1$ estimation.

1. Importance-sampling with resampling:

$t-1$ state density estimation: $\{\alpha_{t-1}^{(i)}, {}^{\alpha}\pi_{t-1}^{(i)}\}_{i=1}^{N_\alpha}$, $\{\beta_{t-1}^{(j)}, {}^{b}\pi_{t-1}^{(j)}\}_{j=1}^{N_b}$, $\{\lambda_{t-1}^{(k)}, {}^{s}\pi_{t-1}^{(k)}\}_{k=1}^{N_s}$.

Resampling: $\{\grave{\alpha}_{t-1}^{(i)}, 1/N_\alpha\}$, $\{\grave{\beta}_{t-1}^{(j)}, 1/N_b\}$, and $\{\grave{\lambda}_{t-1}^{(k)}, 1/N_s\}$.

2. Predictive update of state densities:

$\alpha_t^{(i)} = H\grave{\alpha}_{t-1}^{(i)} + N(0, \sigma_\alpha^2)$

$\beta_t^{(j)} = \grave{\beta}_{t-1}^{(j)} + \tilde{v}_t + N(0, \sigma_b{}^2)$, $\quad b_t^{(j)} = f(\beta_t^{(j)})$

$\lambda_t^{(k)} = \grave{\lambda}_{t-1}^{(k)} + N(0, \sigma_{s_{t-1}}^2)$, $\quad \lambda_t^{(k)} = \frac{\lambda_t^{(k)}}{\sum_{i=1}^{N_s} \lambda_i^{(k)}{}_t}$, $\quad s_t^{(k)} = \sum_{i=1}^{N_s} \lambda_i^{(k)}{}_t s_i$

3. Sequential update of state weights using current observation: Global transformation α_t with $\hat{b}_t$, $\hat{s}_t$:

$P(\alpha_t^{(i)} | \hat{b}_t^*, \hat{s}_t^*, Z_t) \propto P(Z_t | \alpha_t^{(i)}, \hat{b}_t^*, \hat{s}_t^*) P(\alpha_t^{(i)})$

${}^{\alpha}\pi_t^{(i)} = P(Z_t | \alpha_t^{(i)}, \hat{b}_t^*, \hat{s}_t^*)$, $\quad {}^{\alpha}\pi_t^{(i)} = \frac{{}^{\alpha}\pi_t^{(i)}}{\sum_{j=1}^{N\alpha} {}^{\alpha}\pi_t^{(j)}}$

Body pose b_t with α_t, $\hat{s}_t$:

$\alpha_t^* = \alpha_t^{(i^*)}$, where $i^* = \arg\max_i {}^{\alpha}\pi_t^{(i)}$

$P(b_t^{(j)} | \alpha_t^*, \hat{s}_t^*, Z_t) \propto P(Z_t | \alpha_t^*, b_t^{(j)}, \hat{s}_t^*) P(b_t^{(j)})\ {}^{b}\pi_t^{(j)} = P(Z_t | \alpha_t^*, b_t^{(j)}, \hat{s}_t^*)$, $\quad {}^{b}\pi_t^{(j)} = \frac{{}^{b}\pi_t^{(j)}}{\sum_{i=1}^{N_b} {}^{b}\pi_t^{(i)}}$

Style s_t with α_t, b_t:

$b_t^* = b_t^{(j^*)}$, where $j^* = \arg\max_j {}^{b}\pi_t^{(j)}$

$P(s_t^{(k)} | \alpha_t^*, b_t^*, Z_t) \propto P(Z_t | \alpha_t^*, b_t^*, s_t^{(k)}) P(s_t^{(k)})$

${}^{s}\pi_t^{(k)} = P(Z_t | \alpha_t^*, b_t^*, s_t^{(k)})$, $\quad {}^{s}\pi_t^{(k)} = \frac{{}^{s}\pi_t^{(k)}}{\sum_{i=1}^{N_s} {}^{s}\pi_t^{(i)}}$

4.4 Observation Model

In our multi-state representation, we update weights ${}^{\alpha}\pi_t^{(i)}$, ${}^{b}\pi_t^{(j)}$, and ${}^{s}\pi_t^{(k)}$ by marginalized likelihood $P(Z_t | \alpha_t^{(i)}, b_t, s_t^*)$, $P(Z_t | \alpha_t^*, b_t^{(j)}, s_t^*)$, and $P(Z_t | \alpha_t^*, b_t^*, s_t^{(k)})$ given observation Z_t. Each sub state captures different characteristics in the dynamic motion and affects different variations in the observation. For example, body poses are changed according to body configuration state. Different body configurations show significant changes in edge direction in the legs. However, in case of style, the variation is subtle and changes along the global contours.

Observation model measures state X_t by updating the weights $\pi_t^{(i)}$ in the particle filter by measuring the observation likelihood $P(Z_t | X_t^{(i)})$. We can estimate the likelihood by

$$P(Z_t | X_t) = P(Z_t | \alpha_t, b_t, s_t) \propto \exp\left(-\frac{d(Z_t, z_t)}{\sigma_t^2}\right) = \exp\left(-\frac{d(Z_t, T_{\alpha_t}\mathscr{A} \times s_t \times \psi(b_t))}{\sigma_t^2}\right), \quad (10)$$

where $d(\cdot)$ is distance measure, z_t is the contour from the generative model using α_t, b_t, s_t. It is very important to use proper distance measurement $d(\cdot)$ in order to get accurate update of weights and density for new observation. We use three distance measures: Chamfer distance, weighted Chamfer distance, and oriented Chamfer distance.

Representation: We represent shape contour by an implicit function similar to [6] where the contour is the zero level of such function. From each new frame Z_t, we extracted edge using Canny edge detector algorithm.

Weighted Chamfer Distance for Geometric Transformation Estimation: For geometric transformation estimation, the predicted body configuration, and style estimate from the previous frame are used. Therefore, we need to find similarity measurement which is robust to the deviation of body pose and style estimation and sensitive to global transformation. Typically the shape or the silhouette of upper body part in walking sequence are relatively invariant to the body pose and style. By giving different weight to different contour points in Chamfer matching, we can emphasize upper body part and de-emphasize lower body part in the distance measurement. Weighted chamfer distance can be computed as

$$D_w(T,F,W) = \frac{1}{N}\sum_{i}^{N}\min_{f_i \in F}\rho(t_i, f_i)w_i, \tag{11}$$

where t_i is i'th feature location, f_i template feature and w_i i'th feature weight. Practically, weighted chamfer distance achieved more robust estimation of the geometric transformation.
Oriented Chamfer Distance for Body Pose: Different body poses can be characterized by the orientation of legs. Therefore, oriented edge based similarity measurement is useful in case of the body configuration estimation. Oriented chamfer distance matching was used in [8]. We use a linear filter [9] to detect oriented edge efficiently after edge detection. After applying the linear filter to the contour and the observation, we applied chamfer matching for each oriented distance transform and oriented contour template. The final result is the sum of each of the oriented chamfer distance. For style estimation, simple Chamfer distance is used.

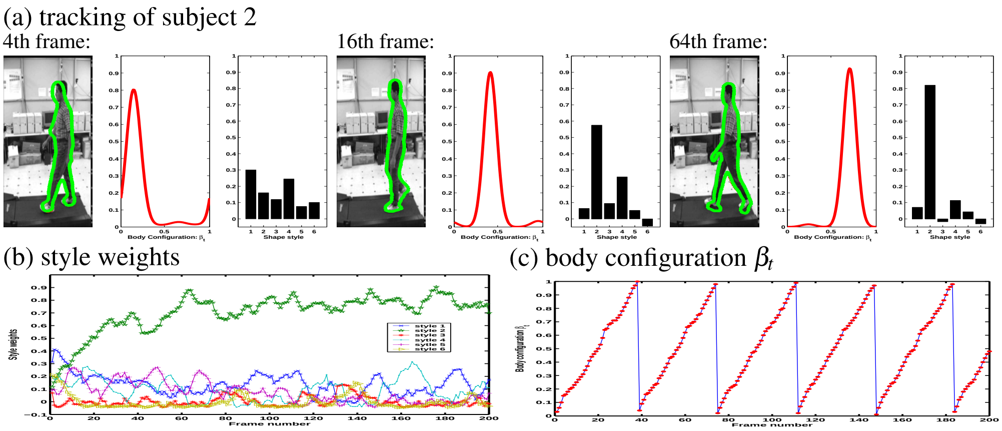
5 Experimental Result

We used CMU Mobo data set for learning the generative model, the dynamics, and for testing the tracker. Six subjects are used in learning the model. The tracking performance is evaluated for people used in training and for unknown people, which were not used in the learning. We initialize the tracker by giving a rough estimate of initial global transformation parameter. The body configuration is initialized by random particles along the manifold. In case of style, as we don't know the subject style from the initial frame, we initialize style by mean style, which means equal weights are applied for every style class.

Tracking for trained subjects :

For the training subjects, the tracker shows very good tracking results. It shows accurate tracking of body configuration parameter β_t and correct estimation of shape style s_t. Fig. 3 (a) shows several frames during tracking known subject. The left column shows tracking contours. The middle column shows the posterior body configuration density. The right column is the estimated style weights in each frame. Fig. 3 (b) shows tracking results for style weights. The figure shows that the style estimate converges to the subject's correct style and it becomes the major weighting factor after about 10 frames. The style weighting shows accurate identification of the subject as a result of tracking and it has many potential for human identification and others. Fig. 3 (c) shows estimated body configuration β value. Even though the two strides making each gait cycle are very similar and hard to differentiate in visual space, the body configuration parameter accurately find out the correct configuration. The figure shows that the body configuration correctly exhibits linear dynamics. As we have one to one mapping between body configuration on the manifold and 3D body pose, we can directly recover 3D body configuration using the estimated β or using manifold points b_t similar to [6].

Tracking for unknown subjects: Tracking for new subjects can be hard as we used small number of people for learning style in the generative model. It takes more frames to converge


(a) tracking of subject 2

4th frame:

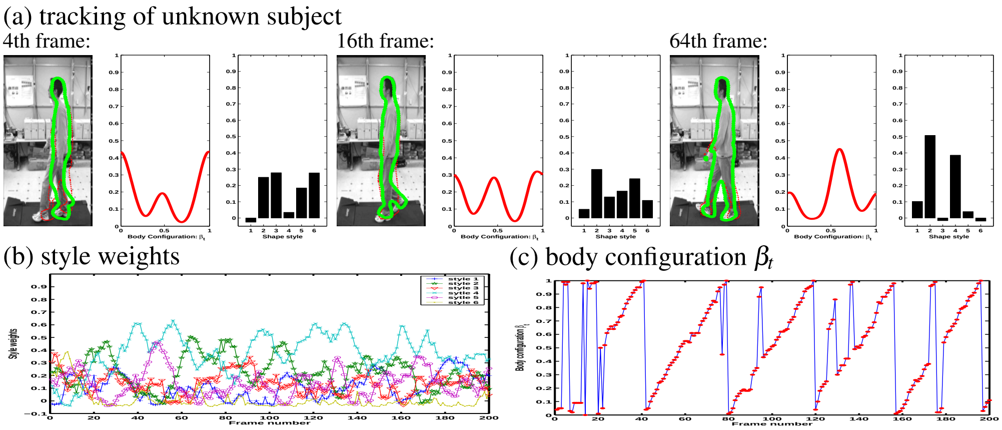
16th frame:

64th frame:

(b) style weights

(c) body configuration β_t

Figure 3: Tracking for known person


(a) tracking of unknown subject

4th frame:

16th frame:

64th frame:

(b) style weights

(c) body configuration β_t

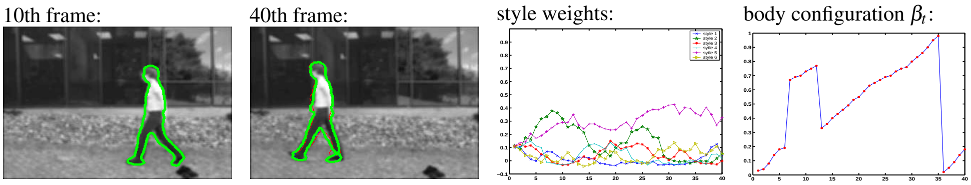
Figure 4: Tracking for unknown person

to accurate contour fitting as shown in Fig. 4 (b). However, after some frames it accurately fit to the subject contour even though we did not do any local deformation to fit the model to the subject. There is no one dominant style class weight during the tracking and sometimes the highest weight are switched depend on the observation. In case of the body configuration β_t, you can see sometimes it jumps about half cycle due to the similarity in the observation since the style is not accurate enough. Still, the result shows accurate estimation of body pose.

We also, applied tracking in normal walking situation using M. Black straight walking person image sequence even though we learned the generative model from treadmill walking data. Fig. 5 shows contour tracking results for 40 frames. Fig. 5 (c) shows estimated body configuration parameters. It confused in some intervals at the beginning but it recovers within the cycle.

6 Conclusion

We presented new framework for human motion tracking using a decomposable generative model. As a result of our tracking, we not only find accurate contour from cluttered envi-



10th frame:

40th frame:

style weights:

body configuration β_t:

Figure 5: Tracking straight walking

ronment without background learning or appearance model, but also get parameters for body configuration and shape style.

In current work, we assumed fixed view in learning the generative model and tracking human motion. We plan to extend the model to view variant situations. We used marginalized density approximation instead of full joint distribution of the state. Sampling based on Markov Chain Monte Carlo (MCMC) can be used for more accurate estimation of the marginal density. In our current system, we tried to overcome the problem in marginalization by finding robust distance measure in the observation and sequential estimation using predicted density.

References

[1] Richard Bowden. Learning statistical models of human motion. In *Proc. IEEE Workshop on Human Modeling, Analysis and Synthesis*, pages 10–17, 2000.

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

[3] T.F. Cootes, D. Cooper, C.J. Taylor, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[4] David Cunado, Mark S. Nixon, and John Carter. Automatic extraction and description of human gait models for recognition purposes. *Computer Vision and Image Understanding*, 90:1–41, 2003.

[5] James W. Davis and Hui Gao. An expressive three-mode principal components model for gender recognition. *Journal of Vision*, 4(5):362–377, 2004.

[6] Ahmed Elgammal and Chan-Su Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *Proc. IEEE CVPR*, pages 681–688, 2004.

[7] Ahmed Elgammal and Chan-Su Lee. Separating style and content on a nonlinear manifold. In *Proc. IEEE CVPR*, pages 478–485, 2004.

[8] D.M. Gavrila. Multi-feature hierarchical template matching using distance transforms. In *Proc. IEEE ICPR*, pages 439–444, 1998.

[9] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Addison Wesley, 1992.

[10] Michael Isard and Andrew Blake. Condensation–conditional density propagation for visual tracking. *Int.J.Computer Vision*, 29(1):5–28, 1998.

[11] Lieven De Lathauwer, Bart de Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM Journal On Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[12] D. Ormoneit, H. Sidenbladh, M. J. Black, T. Hastie, and D. J. Fleet. Learning and tracking human motion using functional analysis. In *Proc. IEEE Workshop on Human Modeling, Analysis and Synthesis*, pages 2–9, 2000.

[13] Sam Roweis and Lawrence Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[14] Kentaro Toyama and Andrew Blake. Probabilistic tracking in a metric space. In *Proc. IEEE ICCV*, pages 50–57, 2001.

[15] Qiang Wang, Guangyou Xu, and Haizhou Ai. Learning object intrinsic structure for robust visual tracking. In *Proc. of CVPR*, pages 227–233, 2003.