

Dual Generative Models for Human Motion Estimation from an Uncalibrated Monocular Camera

Xin Zhang and Guoliang Fan *
School of Electrical and Computer Engineering
Oklahoma State University, Stillwater, OK
{xin.zhang,guoliang.fan}@okstate.edu

Abstract

We propose a new approach to estimate gait kinematics from image sequences taken by a monocular uncalibrated camera. This approach involves two generative models for gait representations in the kinematic and visual spaces, which induce two gait manifolds that characterize the gait variability in terms of the kinematics and visual appearance. A manifold topology enforcement scheme is introduced to incorporate the two gait manifolds. Moreover, a new particle filtering algorithm is proposed for dynamic gait tracking and estimation where a segmental jump-diffusion Markov Chain Monte Carlo (MCMC) technique is developed to accommodate the dynamic nature of the gait variability. The proposed algorithm is trained from CMU Mocap data and tested on the HumanEva dataset with promising results.

1. Introduction

Video-based human motion estimation has recently received great interest due to its wide applications, and it also has been promoted by recent advancements in the fields of computer vision and machine learning. It is still a challenging topic due to the variability, complexity and nonlinearity of human motion as well as the uncertainty and ambiguity of observed image sequences. In this work, we are interested in the estimation of human body configurations from image sequences taken by an uncalibrated monocular camera. Specifically, we focus the human gait that is useful for biometrics and biomechanical modeling applications. The existing approaches for video-based human motion estimation can be roughly classified into two categories, i.e., the discriminative and generative methods.

For discriminative approaches, the relationship between observations and kinematics is learnt directly from training data where prior constraints and inference methods play important roles. In [3], the image-pose relationship is learnt via the relevance vector machine (RVM). In [10], an exemplar database was developed that characterizes human poses by both image data and kinematic data, and the training and test data have to be same. One recent work in [14] introduced the body symmetric correlation as prior in the particle filter-based inference framework where the average joint estimation error is around 140mm.

Generative methods usually involve a prior motion model. In [13], a maximum *a posteriori* (MAP) estimation framework was incorporated into the Gaussian Process Latent Variable Model (GPLVM) where observations are 2D feature points extracted from images. In [7], a nonlinear tensor decomposition technique was proposed that combines both manifold learning and multilinear tensor decomposition. It provides a generative model for human motion based on which the human pose can be inferred from body silhouettes. In [5, 6], two low dimensional motion representations are learnt from gait kinematics and visual silhouettes, and a non-linear mapping between the two models was established for motion estimation. Most methods above cannot estimate unknown gait kinematics.

Here we will significantly extend the method in [7] in order to estimate unknown gait kinematics from image sequences captured by an uncalibrated camera. Our assumption is that *a new human gait can be extrapolated from a set of representative gaits*. We develop two generative models to represent human gaits in both the kinematic and visual spaces, which induce two *gait manifolds* to capture the gait variability in terms of kinematics and appearance. We develop a manifold topology enforcement technique to incorporate two gait manifolds. Moreover, a new particle filtering algorithm is proposed for dynamic gait tracking and estimation.

*This work is supported by the National Science Foundation (NSF) under Grant IIS-0347613.

2. Proposed Algorithm

The highlight of our research is the two gait generative models that are learnt in the kinematic space (gait motions) and visual space (gait silhouettes) respectively. In the following, we will discuss (1) how to obtain the two generative models; (2) how to establish the mapping relationship between two generative models; (3) how to support dynamic gait tracking and estimation by using two generative models and their relationship.

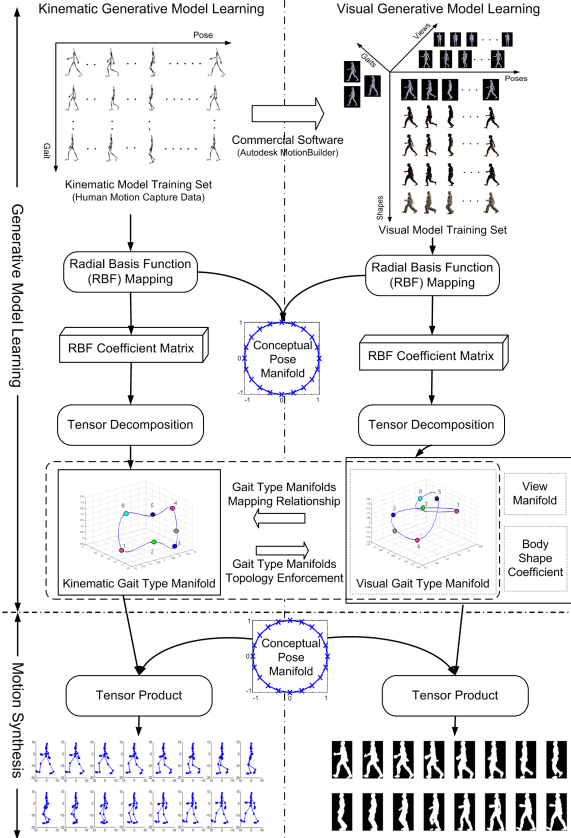


Figure 1. The two gait generative models that are learnt in the kinematic and visual spaces respectively. The two models are associated through the non-linear mapping function between two gait manifolds induced by the two generative models.

2.1 Dual Gait Generative Models

We will involve two gait generative models, namely the *kinematic gait generative model (KGGM)* and the *visual gait generative model (VGGM)*. Both of them are learnt from a large training data set by using non-linear tensor decomposition proposed in [7] that combines manifold learning with multi-linear analysis for human motion representation.

2.1.1 Kinematic Gait Generative Model (KGGM)

We represent a set of gait kinematics as a sequence of Euler angles of each joint estimated from motion capture data. The KGGM characterizes the kinematics of a gait by two independent factors, the pose (a certain stage in a complete walking cycle) and the gait type (the gait individuality), both of which can be specified along a low dimensional manifold. For a periodic gait motion, we can define a 2D circle to represent the generic *gait independent* pose manifold.

For gait i of N_p poses, its k -dimensional kinematics is denoted by $\mathbf{Z}^i = \{\mathbf{z}^{(i,q)} \in \mathbb{R}^k, q = 1, \dots, N_p\}$ and corresponding poses are represented by $\mathbf{P} = \{\mathbf{p}^q \in \mathbb{R}^2, q = 1, \dots, N_p\}$ that are uniformly sampled from the pose manifold. A generalized Radial Basis Function (RBF), $\psi(\mathbf{p}) = [\phi(\mathbf{p}, \mathbf{c}_1), \dots, \phi(\mathbf{p}, \mathbf{c}_{N_c})]^T$ where $\{\mathbf{c}_l | l = 1, \dots, N_c\}$ are the kernel centers along the pose manifold, is used for the non-linear mapping $\mathbb{R}^{N_c} \rightarrow \mathbb{R}^k$, and mapping coefficients are encapsulated into a matrix \mathbf{D}^i . All gait-dependent mapping matrices $\{\mathbf{D}^i | i = 1, \dots, N_g\}$ (N_g training gaits) can be stacked as a tensor and the high order singular value decomposition (HOSVD) is applied to get a coefficient vector $\boldsymbol{\kappa}^i$ for each gait type. The generative model is defined as

$$\mathbf{z}^{(i,q)} = \mathcal{A} \times_1 \boldsymbol{\kappa}^i \times_2 \psi(\mathbf{p}^q), \quad (1)$$

where \mathcal{A} is a 3-order *core tensor*. Given a gait type $\boldsymbol{\kappa}^i$, this model can synthesize the body configuration of any pose \mathbf{p}^q defined along the circular manifold.

2.1.2 Visual Gait Generative Model (VGGM)

Additionally, we invoke the VGGM to represent a gait silhouette by four independent factors, i.e., the pose, the gait type, the view (under which the human is observed), and the body shape (the shape individuality). Using the same pose manifold, a d -dimensional gait silhouette $\mathbf{y} \in \mathbb{R}^d$ can be generated by the VGGM as,

$$\mathbf{y}^{(k,j,i,q)} = \mathcal{C} \times_1 \mathbf{v}^k \times_2 \mathbf{s}^j \times_3 \boldsymbol{\nu}^i \times_4 \psi(\mathbf{p}^q), \quad (2)$$

where \mathcal{C} is the 5-order core tensor. \mathbf{v}^k , \mathbf{s}^j , $\boldsymbol{\nu}^i$ represent the view, shape, and gait type respectively. It is worth mentioning that the gait type variables $\boldsymbol{\kappa}^i$ in (1) and $\boldsymbol{\nu}^i$ in (2) specify a gait type in the kinematic and visual spaces respectively. We can also obtain a view manifold by using the spline fitting method [7] that can generate the coefficient vector of an arbitrary view. A new shape can be created by a linear convex combination of a set of shape coefficient vectors learnt from the VGGM training. The learning of VGGM requires a large set of gait animations created by commercial software MotionBuilder that involves several 3D character models and various gait motion data.

2.2 Mapping between Two Generative Models

One key issue is how to find a mapping relationship between the KGGM and VGGM, by which we can infer unknown gait kinematics from gait silhouettes. Here we advocate a concept of *gait manifold* that captures gait variability among individuals. We can obtain two 1-D gait manifolds from the two generative models by using the spline fitting method that was used for view manifold generation in [7]. However, unlike the view manifold that has a clear intrinsic structure (the view order), the underlying structure of gait types is unknown. We use the shortest path to link all gait type vectors that was found effective to explore the unknown intrinsic low-dimensional structure in the coefficient space [12].

Hence two (kinematic/visual) gait manifolds can be obtained from the KGGM and VGGM that capture the gait variability in the kinematics and visual spaces. However, they may not share the same topology (the order of gait types along the manifold) due to their different natures. Because our goal is to estimate an unknown gait in the kinematic space, we want that the visual gait manifold is compliant with the kinematic one. We propose a *manifold topology enforcement* technique to ensure the consistency between the two manifolds (Fig. 1). Afterwards, a non-linear mapping function between two manifolds is learnt using the RBF [4] that fills the gap between the KGGM and VGGM. Essentially, kinematics of a new gait is estimated by non-linear interpolation from training gaits along the kinematic gait manifold.

2.3 Dynamic Gait Estimation

The VGGM suits well the Bayesian framework, where four latent variables ($\mathbf{x}_t = [\mathbf{p}_t, \mathbf{v}_t, \mathbf{s}_t, \boldsymbol{\nu}_t]$) evolve according their own dynamics as shown in Fig. 2. Specifically, we use a constant speed dynamic model for *pose* (\mathbf{p}_t) along the pose manifold, a random walk for *view* (\mathbf{v}_t) along the view manifold, and a random walk for *shape* (\mathbf{s}_t) in the coefficient space. We define $p(\mathbf{y}_t | \mathbf{x}_t)$ that computes the likelihood of observation \mathbf{y}_t given \mathbf{x}_t and involves silhouette synthesis defined in (2). At each time step, the four variables are estimated sequentially and iteratively via a MCMC-based particle filter ($\mathbf{p}_t \rightarrow \mathbf{v}_t \rightarrow \mathbf{s}_t \rightarrow \boldsymbol{\nu}_t \rightarrow \mathbf{p}_{t+1}$). We need a special treatment for $\boldsymbol{\nu}_t$ that usually may vary near the *contact* pose (as defined in [1]) in a long gait sequence. Hence, we need a mixed dynamics for $\boldsymbol{\nu}_t$ that can exploit (e.g., jump) and explore (e.g., diffusion) along the gait manifold. Also, the segmental modeling [9] is needed where each segment corresponds a half walking cycle with a stable gait type. We propose a *segmental jump-diffusion MCMC algorithm* to estimate $\boldsymbol{\nu}_t$.

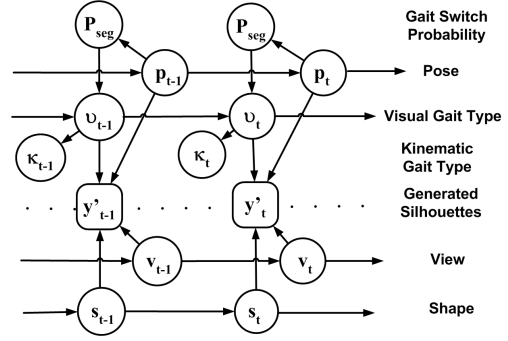


Figure 2. The graphic model for the inference.

We define a segmental prior to control the mixed dynamics of $\boldsymbol{\nu}$, $P_{seg} = \beta e^{-\tau_p^2/2\delta^2} \sim (0, 1]$ where τ_p is the distance from pose \mathbf{p} to *contact* pose along the pose manifold. To simplify the inference, we introduce a *scalar* gait variable ν that is defined along a circle and can be mapped to the *vector* gait variable $\boldsymbol{\nu}$ defined in the VGGM, i.e., $\nu \leftrightarrow \boldsymbol{\nu}$. ν is assumed to have M modes each of which is characterized by a Gaussian, i.e., $\{\mathcal{N}(\nu; \mu_r, \delta_r^2) | r = 1, \dots, M\}$. We use a mixed dynamics to infer ν or $\boldsymbol{\nu}$ that includes *jump* (the switch between modes) and *diffusion* (the estimation within a mode). Initially, M modes are uniformed distributed with large variances along the circle, and will be updated during the inference in an annealing fashion. The Metropolis-Hasting algorithm is used here where the acceptance ratio of the candidate gait type $\boldsymbol{\nu}^*$ or ν^* is $\alpha = \min \left\{ 1, \frac{\mathcal{P}(\boldsymbol{\nu}^*) \mathcal{Q}(\boldsymbol{\nu}^{(i)} | \boldsymbol{\nu}^*)}{\mathcal{P}(\boldsymbol{\nu}^{(i)}) \mathcal{Q}(\boldsymbol{\nu}^* | \boldsymbol{\nu}^{(i)})} \right\}$, where $\mathcal{P}(\boldsymbol{\nu}^*)$ is the posterior probability $p(\boldsymbol{\nu}_t^* | \mathbf{y}_t, \mathbf{x}_{t-1})$ determined by $p(\mathbf{y}_t | \mathbf{x}_t)$ and previous estimation, and $\mathcal{Q}(\cdot | \cdot)$ is the proposal distribution of the dynamics of $\boldsymbol{\nu}^*$. The pseudocode of the gait type inference is listed below.

- Initialize the sample $\nu_t^{(0)} = \nu_{t-1}$ and its underlying mode $m_t^{(0)} = m_{t-1}$, as well as the diffusion variance σ_d^2 .
- Compute the segmental prior P_{seg} .
- FOR $i = 1, \dots, (B + MN)$ (N is the number of samples, B is the burn-in period and M is the thinning interval.)
 1. Randomly sample $\gamma \sim U[0, 1]$.
 - IF $P_{seg} \geq \gamma$, **Jump**. Randomly select a mode m^* and sample ν^* using $\mathcal{Q}(\nu^* | \nu_t^{(i)}) = \mathcal{N}(\nu^*; \mu_{m^*}, \delta_{m^*}^2)$. Compute α .
 - ELSE **Diffusion**. Sample ν^* from $\mathcal{Q}(\nu^* | \nu_t^{(i)}) = \mathcal{N}(\nu^*; \nu_t^{(i)}, \sigma_d^2)$. Compute α .
 2. Randomly sample $\eta \sim U[0, 1]$.
 - IF $\alpha \geq \eta$ them accept ν^* as $\nu_t^{(i+1)} = \nu^*$.
 - * IF ν^* is generated by *diffusion*, $\sigma_d \leftarrow \sigma_d^{\frac{1}{D}}$ (l is an annealing constant and D is the number of diffusion samples).
 - ELSE, reject ν^* and let $\nu_t^{(i+1)} = \nu_t^{(i)}$.
- Return the new sample set $\{\nu_t^{(n)}\}_{n=1}^N$. ν_t is the mean of all samples and the estimated mode is obtained by the maximum likelihood estimation, i.e., $m_t = \arg \max_r \mathcal{N}(\nu_t | \mu_r, \delta_r^2)$.
- Update mode m_t in terms of the mean and variance: $\mu_{m_t} \leftarrow \frac{1}{2}(\mu_{m_t} + \nu_t)$, $\delta_{m_t} \leftarrow \delta_{m_t}^{\frac{1}{c}}$, where c is an annealing constant.

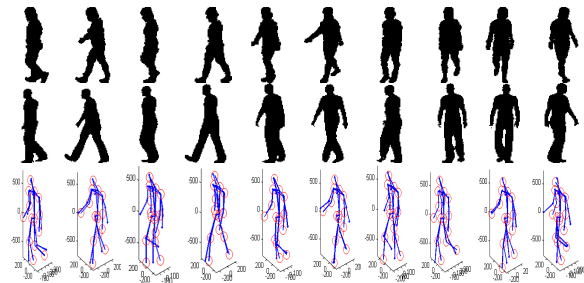


Figure 3. Experimental results (from top to bottom: input silhouettes, synthesized silhouettes, the estimated/ground-truth gait kinematics (the stick man with/without circle)).

3. Experimental Results and Discussions

We used CMU Mocap data [2] and HumanEva-I data [11] as the training and test data respectively. Specifically, we used 20 gait motions to learn the kinematic gait generative model. We also created 1,200 gait animations by using MotionBuilder that is composed of 20 gait motions, 12 views, and 5 body shapes, and each gait animation has 30 poses. Motion retargetting provided by MotionBuilder was used to ensure the compatibility between the gait motion data and 3D character models. HumanEva provides video sequences and corresponding gait motion data. Three video sequences were used, each of which has one person walking along a circle. We used a background substraction method to obtain gait silhouettes that are followed by manual clean-up and normalization. Since the training and test data have different marker configurations that complicate gait matching, we used a skeleton mapping scheme that can make gait kinematics from different marker systems more compatible and comparable [8].

Fig. 3 illustrates experimental results on gait synthesis and estimation, and Table 1 shows numerical results on three subjects where our algorithm has four implementations. Firstly we examine the capability of the KGGM for gait synthesis by computing the lower error bound (LEB) of gait approximation. Alg-1 is the implementation with a basic particle filter without dynamic gait tracking; Alg-2 is the implementation where the gait type is fixed to be the one estimated from Alg-1. Alg-3 is the implementation where the particle filter is embedded with the segmental jump-diffusion MCMC for dynamic gait tracking and estimation, and Alg-4 is the same as Alg-3 except that the M modes of the gait type variable are learnt from Alg-3. The overall error is computed by averaging 14 joints' errors (30 DOF) over the whole sequence. The estimation results are obviously improved progressively from Alg-1 to Alg-4. By comparing with [14], our results are very promising.

Subjects	LEB	Alg-1	Alg-2	Alg-3	Alg-4	[14]
S1	32.20	77.45	69.14	66.18	61.86	140.35
S2	46.25	91.42	80.81	78.76	73.34	149.37
S3	44.87	98.72	87.50	82.35	78.39	156.30

Table 1. Gait estimation errors (mm) on HumanEva.

4. Conclusion

We have proposed dual generative models to estimate gait kinematics from video sequences captured by an uncalibrated camera. The two generative models provide a general gait representation in the kinematic and visual spaces respectively. A manifold enforcement technique is proposed to fill the gap between two generative models, so that we can infer unknown gait kinematics from visual observations. A new particle filtering algorithm is proposed to support dynamic gait tracking and estimation and achieves state-of-the-art results.

References

- [1] <http://www.idleworm.com/how/anm/02w/walk1.shtml>.
- [2] CMU Human Motion Capture Database. Available at <http://mocap.cs.cmu.edu>.
- [3] A. Agarwal and B. Triggs. 3D human pose from silhouettes by relevance vector regression. In *Proc. of IEEE CVPR*, 2004.
- [4] A. Agarwal and B. Triggs. Recovering 3D human pose from monocular images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28, 2006.
- [5] F. Guo and G. Qian. 3D human motion tracking using manifold learning. In *Proc. of IEEE ICIP*, 2007.
- [6] T. Jaeggli, E. Koller-Meier, and L. V. Gool. Multi-activity tracking in LLE body pose space. In *Proc. of IEEE ICCV 2nd Workshop on Human Motion*, 2007.
- [7] C.-S. Lee and A. Elgammal. Modeling view and posture manifolds for tracking. In *Proc. of IEEE ICCV*, 2007.
- [8] J.-S. Monzani, P. Baerlocher, R. Boulic, and D. Thalmann. Using an intermediate skeleton and inverse kinematics for motion retargeting. *Computer Graphics Forum*, 19:11–19, 2000.
- [9] S. M. Oh, J. M. Rehg, T. Balch, and F. Dellaert. Learning and inferring motion patterns using parametric segmental switching linear dynamic systems. *International Journal of Computer Vision*, 77(1-3), May 2008.
- [10] R. Poppe. Evaluating example-based pose estimation: experiments on the HumanEva set. In *Proc. of CVPR Workshop on EHUM*, 2007.
- [11] L. Sigal and M. Black. HumanEva: Cynchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, 2006.
- [12] C. Tian, X. Gao, and G. Fan. Multi-view face recognition by nonlinear tensor decomposition. In *Proc. of ICPR*, 2008.
- [13] R. Urtansun. *Motion Model for Robust 3D Human Body Tracking*. PhD thesis, EPFL, 2006.
- [14] X. Xu and B. Li. Learning motion correlation for tracking articulated human body with a Rao-Blackwellised particle filter. In *Proc. of IEEE ICCV*, 2007.