Human Motion Tracking with Monocular Video by Introducing a Graph Structure into Gaussian Process Dynamical Models

Jianfeng Xu, Koichi Takagi, and Shigeyuki Sakazawa

KDDI R&D Laboratories Inc.
{ji-xu,ko-takagi,sakazawa}@kddilabs.jp

Abstract. This paper presents a novel approach to tracking articulated human motion with monocular video. In a conventional tracking system based on particle filters, it is very challenging to track a complex human pose with many degrees of freedom. A typical solution to this problem is to track the pose in a low dimensional latent space by manifold learning techniques, e.g., the Gaussian process dynamical model (GPDM model). In this paper, we extend the GPDM model into a graph structure (called *GPDM graph*) to better express the diverse dynamics of human motion, where multiple latent spaces are constructed and dynamically connected to each other appropriately by an unsupervised learning method. Basically, the proposed model has both intra-transitions (in each latent space) and inter-transitions (among latent spaces). Moreover, the probability of inter-transition is dynamic, depending on the current latent state. Using the proposed GPDM graph model, we can track human motion with monocular video, where the average tracking errors are improved from the state-of-the-art methods in our experiments.

Keywords: motion tracking, monocular video, manifold learning, Gaussian process dynamical model, motion graph.

1 Introduction

In the computer vision community, much effort has been put into inferring the human pose or 3D articulated human body parts from videos [6,11]. Basically, there are two kinds of approaches: on one side, *discriminative approaches* employ a parametric model mapping directly from image observation to the pose space [1,15]. Although recent techniques are developed with promising performance [15], it is generally quite difficult to learn of such mapping because the mapping itself is generally ambiguous, e.g. two different poses may have almost the same observation. On the other side, the inverse problem of generating image observations by a given pose is well defined, leading the *generative approaches* to optimize the pose (or pose distribution). As a typical technique of generative approaches, particle filters are widely adopted to track human motion from videos [7,13,16] and are also employed in this paper.

Y.-S. Ho (Ed.): PSIVT 2011, Part I, LNCS 7087, pp. 370–383, 2011.

In most papers [1,7,13,14,15,16], the human pose is represented as articulated human body parts in a tree structure with many degrees of freedom [6,11]. Therefore, the human pose is very difficult to track directly in the high dimensional pose space due to the curse of dimensionality with such techniques as the particle filters [7]. Fortunately, recent studies demonstrate that human motion can essentially be described in a much lower dimensional space (called *latent space*) [9,16,17], given that targeted motion has regular dynamics. In this paper, the Gaussian process dynamical models (GPDM) proposed by Wang et al. [17] is employed because of the good performance as reported by Quirion et al. [12] for many applications in tracking human motion [7,16]. However, it is unsatisfactory for a single GPDM to express complicated motion that has several motion patterns [7].

Our basic idea is to separate complicated motion into simple segments, where a GPDM model (i.e., latent space) is learned for each segment. Naturally, those latent spaces should be transited with a probability. Moreover, the transition probability among latent spaces (called *inter-transition*) should depend on the current state of the current latent space. For example, the probability of inter-transitions is much higher at the landing state than that at the flight state from the jumping space to walking space. Generally speaking, it is very challenging to learn such a complicated latent model in a reasonable way. For this purpose, we combine the techniques of the motion graph [3,8,10] and GPDM to construct our novel model *GPDM graph*. As far as we know, it is the first latent dynamics model with a graph structure. In addition, our approach is a completely unsupervised learning method by the data-driven scheme.

Although monocular approaches are much more challenging than multi-view approaches due to incomplete information, such as the occlusion problem [6,11], a single camera is more ubiquitous and cheaper, thus making it suitable for non-professional users. Moreover, a single camera solution can open up a new possibility to capture motion from video archives such as past Olympic games. In both cases, currently, we do not require real time processing, targeting to the applications for entertainment, coaching, etc.

The rest of this paper is organized as follows. Section 2 presents a brief survey on related work. Section 3 describes the proposed algorithm in detail. Section 4 discusses our experimental results on the HumanEva dataset [13]. The conclusions and future work are addressed in section 5.

2 Related Work

A plethora of literature is reported on video-based human motion tracking. See the comprehensive reviews in previous surveys [6,11]. In this section, we focus on dimension reduction and particle filter techniques for human motion tracking, which are the categories of our core techniques.

Although principle component analysis (PCA) is widely used for dimension reduction in human motion [2], linear mapping has poor ability to reduce the dimensions because human motion is highly non-linear. As a non-linear approach,

the Gaussian process latent variable model (GPLVM) can learn the latent space and the mapping function [9]. GPLVM is an efficient tool for modeling distribution in a high dimensional space with a compact low dimensional representation. Wang et al. extend GPLVM to GPDM [17], which models the dynamics in the learned latent space. GPDM and its variants, including BGPDM [16], are widely employed in tracking human motion because it simultaneously models the latent space, the dynamics in the latent space, and the mapping from latent space to the pose space. GPDM is an unsupervised method and only needs a minimum of learning data [17]. However, Chen et al. [7] have reported that GPDM cannot model complicated motion. They introduce a switching GPDM model that is successfully used in human motion tracking [7]. In their model, the transition probability of switching states is static. Moreover, labels of switching states in the learning data are usually required, which means that it is a supervised learning method. The essential difference between our GPDM graph and the switching GPDM is whether to learn *dynamic* switching probability with an *unsupervised* method, which is very challenging but important in real applications.

On the other hand, particle filters and variants are successfully applied to track objects in video because of the compatibility of non-linear and non-Gaussian elements [4]. However, the workable dimensionality for particle filters is small as pointed out by Chen et al. [7]. With the above dimension reduction methods, it is possible to track human motion using particle filters in a low dimensional latent space. In this paper, we employ a particle filter technique similar to Sigal et al. [13]. Our experimental results show that performance is further improved from the state-of-the-art methods [7,14,17]. See the details in Section 4.

3 Proposed Method

Our system includes learning the GPDM graph and inference with GPDM graph. To learn the GPDM graph, training motion data are divided into several short segments, and a GPDM model is simultaneously learned for each segment. At the same time, the candidates for inter-transitions among GPDM models are detected using the short-term principle component analysis, originally proposed by Xu et al. [19]. With the learned GPDM graph, which includes the mapping function from latent space to pose space, the human pose is inferred with the low dimensional latent space by particle filters. In this stage, inter-transitions are dynamically determined by the similarity of human poses. In Section 3.1, we will first describe the concept of the GPDM graph in detail.

3.1 Concept of GPDM Graph

The basic hypothesis is that a complicated motion consists of a sequence of elemental motions, and each elemental motion, originally in many degrees of freedom, is essentially controlled by low dimensional latent space as shown in Eq. (2) [7,17]. At the same time, the first-order Markov dynamics is assumed

for simplicity in latent spaces as shown in Eq. (1). Furthermore, we connect the latent spaces with a dynamic probability as shown in Fig. 1 (called *inter-transitions*).

$$\mathbf{z}_t^k = f(\mathbf{z}_{t-1}^k; \mathbf{A}) + \mathbf{n}_{z,t} \tag{1}$$

$$\mathbf{x}_t = g(\mathbf{z}_t^k; \mathbf{B}) + \mathbf{n}_{x,t} \tag{2}$$

where $\mathbf{z}_t^k \in \mathbb{R}^d$ denotes the d-dimensional coordinates at time-t in the k-th latent space, $\mathbf{x}_t \in \mathbb{R}^D$ denotes the D-dimensional coordinates at time-t in pose space ($D >> d$), f and g are non-linear mappings parameterized by $\mathbf{A}$ and $\mathbf{B}$, and $\mathbf{n}_{z,t}$ or $\mathbf{n}_{x,t}$ denotes zero-mean, isotropic, white Gaussian noise processes. Note that our model has multiple latent spaces but a single pose space while the original GPDM has a single latent space and a single pose space. Therefore, our model is more general and suitable for complex motions.

One of the unique characteristics in our model is that the inter-transition probability depends on the current state of the current latent space, which infers that probability changes dynamically. The example in Fig. 1 explains the reasonableness of our model, where two kinds of elemental motions exist including "walking" and "jumping". As shown in Fig. 1(b), it is natural that the transition probability at the landing state is much higher than at the flight state when transiting from "jumping" to "walking". Similarly, the transition probability must be dynamic when transiting from "walking" to "jumping" as shown in Fig. 1(c). Surely, besides the inter-transitions, we have intra-transitions in each latent space as the original GPDM did [17]. Note that our model is designed for not only the above scenario that clearly has two motions but also the complex motion with multiple short phases that can transit in-between such as the gesture motion in Table 2.

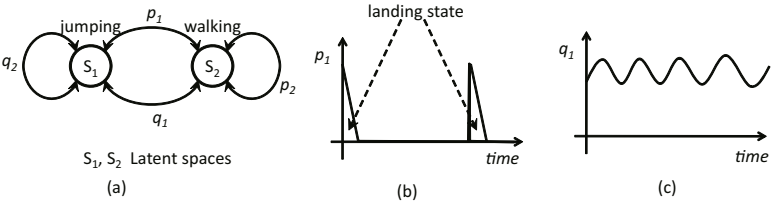


Fig. 1. Concept of the proposed GPDM graph model (a): multiple latent spaces are connected in a probability depending on the current state of the current latent space. Naturally, the probability of inter-transition is much higher at the landing state than that at the flight state from the jumping space to walking space in (b). Similarly, the transition probability is dynamic when transiting from walking to jumping in (c).

Specifically, when using GPDM to learn latent space, the above model can be further represented as Eqs. (3) and (4) through Gaussian process regression, where the dynamics of the latent space is the former, and the mapping from latent space to pose space is the latter. Note that both are probability functions, which are desirable for particle filters. For more details, please refer to [17].

$$p(\mathbf{Z}^k \mid \bar{\alpha}^k) = \frac{p(\mathbf{z}_1^k)}{\sqrt{(2\pi)^{(N-1)d} \mid \mathbf{K}_{Zk} \mid^d}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{K}_{Zk}^{-1}\mathbf{Z}_{\mathbf{2:N}}{}^k\mathbf{Z}_{2:N}^{k\,T})\right) \quad (3)$$

$$p(\mathbf{X} \mid \mathbf{Z}^k, \bar{\beta}^k, \mathbf{W}^k) = \frac{\mid \mathbf{W}^{\mathbf{k}} \mid^N}{\sqrt{(2\pi)^{ND} \mid \mathbf{K}_{Xk} \mid^D}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{K}_{Xk}^{-1}\mathbf{X}\mathbf{W}^{\mathbf{k}2}\mathbf{X}^T)\right) \quad (4)$$

where $\mathbf{Z}^k \equiv \mathbf{Z}_{1:N}^k \equiv \{\mathbf{z}_1^k, \mathbf{z}_2^k, ..., \mathbf{z}_N^k\}$ denotes all the coordinates in the k-th latent space, $\mathbf{X} \equiv \{\mathbf{x}_t : t = 1, ..., N\}$ denotes all coordinates in the pose space, $\bar{\alpha}^k$ denotes kernel hyperparameter vector for dynamics in latent space, which is used in calculating the kernel function $(\mathbf{K}_{Zk})_{ij} \equiv k_{Zk}(\mathbf{z}_i^k, \mathbf{z}_j^k)$ in Eq. (5), $\bar{\beta}^k$ and $\mathbf{W}^k \equiv diag(w_1^k, ..., w_D^k)$ are hyperparameters for the mapping function, where the kernel function $(\mathbf{K}_{Xk})_{ij} \equiv k_{Xk}(\mathbf{x}_i, \mathbf{x}_j)$ is calculated by Eq. (6). In a word, a GPDM model is represented as $\{\mathbf{Z}^k, \bar{\alpha}^k, \bar{\beta}^k, \mathbf{W}^k\}$, which is learned in a segment of motion data.

$$k_{Zk}(\mathbf{z}_i^k, \mathbf{z}_j^k) = \exp\left(-\frac{\beta_1^k}{2} \parallel \mathbf{z}_i^k - \mathbf{z}_j^k \parallel^2\right) + (\beta_2^k)^{-1}\delta_{\mathbf{z}_i^k, \mathbf{z}_j^k} \quad (5)$$

$$k_{Xk}(\mathbf{x}_i, \mathbf{x}_j) = \alpha_1^k \exp\left(-\frac{\alpha_2^k}{2} \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2\right) + \alpha_3^k \mathbf{x}_i^T \mathbf{x}_j + (\alpha_4^k)^{-1}\delta_{\mathbf{x}_i, \mathbf{x}_j} \quad (6)$$

The probability of an inter-transition is intuitively calculated according to the distance between two poses that are transited as Eq. (7), where the principle in the so-called motion graph technique is adopted [3,8,10]. Basically, the more similar the poses are, the higher the transition probability is.

$$-\log p(\mathbf{z}_t^k \rightarrow \mathbf{z}_{t'}^{k'}) \propto dist(\mathbf{x}_t, \mathbf{x}_{t'}) \quad (7)$$

where $\mathbf{z}_t^k$ denotes the departure coordinates in the k-th latent space, where the mean of the mapping function is $\mathbf{x}_t$ in the pose space, and $\mathbf{z}_{t'}^{k'}$ denotes the destination coordinates in the k'-th latent space, where the mean of the mapping function is $\mathbf{x}_{t'}$ in the pose space. The function *dist* is a distance function between two poses. See an implementation by Wang et al. [18], where the weighted difference of joint orientations is calculated as Eq. (8).

$$dist(\mathbf{x}_t, \mathbf{x}_{t'}) = \sum_{n=1}^{m} w_k \parallel \log(q_{t',n}^{-1} q_{t,n}) \parallel^2) \quad (8)$$

where m denotes the number of joints in the human pose, and $q_{t,n}$ denotes the orientation of joint n in the t-th frame, expressed as quaternion.

3.2 Learning of GPDM Graph

Given human motion, the proposed GPDM graph will be learned with an unsupervised method.

Inter-transition Candidate Detector: It is necessary to detect the possible inter-transitions in the training motion data, e.g. the time instants for hitting

the ground in walking motion, where a short-term principal component analysis (short-term PCA) method [19] is employed. The basic idea in short-term PCA is piece-wise linear approximation for non-linear human motion because motion data are almost linear in the short term due to strong temporal coherence. Short-term PCA is executed in a sliding window in the joint position space. And the peaks and valleys of the coordinates in the first principal component are regarded as candidates for inter-transitions $\{\mathbf{b}_i : i = 1, ..., I\}$. The detected candidates for inter-transitions are stored as potential time instants to transit to other motions. See the detailed procedure in [19].

Construction of GPDM Graph: We simultaneously segment training motion data and learn a sequence of GPDM models. The basic idea is to use the trial and error approach iteratively with a sliding window as shown in Table 1. The motion in a window is called a *motion clip*, which is empirically set as 60 frames or 0.5 seconds in our implementation. We merge the motion clips when the reconstruction error, calculated as Eq. (9), is smaller than the threshold as shown in Table 1. Here, the threshold is set as 1.0. Otherwise, it is divided into two segments at the boundary of additional motion clip as shown in Fig. 5(b). In concept, a segment for a motion pattern is desired. In practice, the real concern in the inference is the reconstruction error.

$$error(t) = dist(\mathbf{x}_t, \hat{\mathbf{x}}_t) \tag{9}$$

$$\hat{\mathbf{x}}_t = g(\hat{\mathbf{z}}_t^{\mathbf{k}}) \tag{10}$$

where $\hat{\mathbf{x}}_t$ is the t-th reconstructed pose from the t-th coordinates $\hat{\mathbf{z}}_t^k$ of a so called *mean prediction sequence* in the current latent space, generated from $\mathbf{z}_1^k$ by simulating the dynamical process one frame at a time [17].

Now, our GPDM graph is composed of the GPDM models $\{\mathbf{Z}^k, \bar{\alpha}^k, \bar{\beta}^k, \mathbf{W}^k : k = 1, 2, ..., K\}$ and all the candidates for inter-transitions $\{\mathbf{b}_i : i = 1, ..., I\}$, which will be used in the next section. An example is shown in Fig. 2, where a walking motion in Section 4 is used.

Table 1. Procedure for learning a sequence of GPDM models

```
while training data are not finished
    do add a motion clip
        merge the current clip
        learn the GPDM for merged motion
        if error(t) < TH for any t
                then continue
        else learn the GPDM without the added clip
                output the learned GPDM
                reset the start point as the head of current clip
                break
```

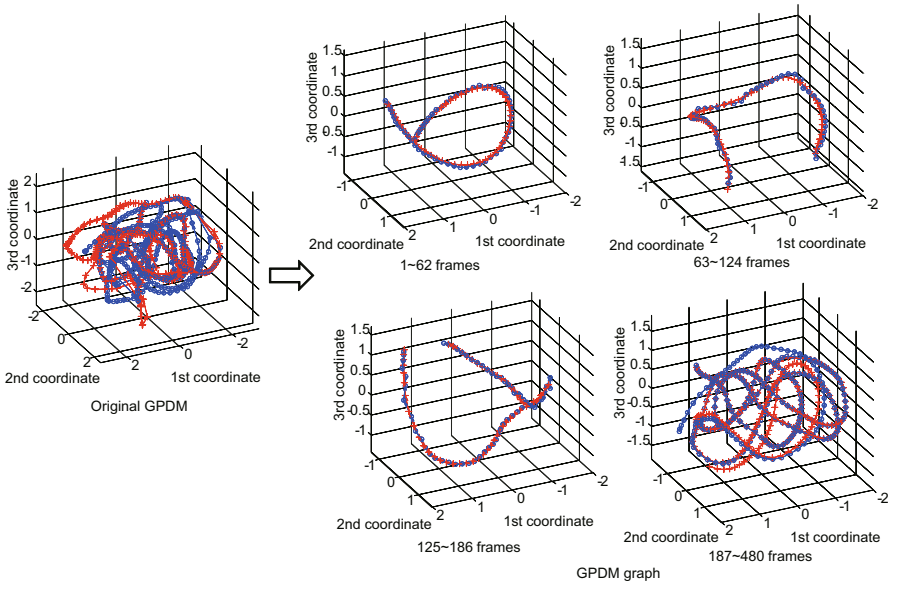


Fig. 2. An example of the proposed GPDM graph model for a walking motion in Table 2, where the circles denote the learned coordinates in latent space and the crosses denote the predicted coordinates in latent space.

3.3 Inference with GPDM Graph

As mentioned before, particle filters are used to infer the human pose from the input video, where the main difference from conventional particle filters [13] is that the particles are generated in the latent spaces instead of the pose space, reducing the space dimension greatly. Later, the particles in the latent space are called *latent particles* $\mathbf{z}_t^k(1:P^k)$, which denotes the P^k coordinates in the k-th latent space for time t. The corresponding particles in the pose space are called *pose particles* $\mathbf{x}_t(1:P)$, which denotes the $P(=\sum P^k)$ coordinates in the pose space for time t and is calculated by the mean of GP regression in Eq. (6) as Wang et al. [17] reported.

Similar to conventional particle filters [13], the initialization is specially processed. In detail, the ground truth of the first frame is used to generate particles. First, we search the human poses in the training motion data to find pose candidates, which are required to be similar to the first frame (i.e. satisfied by Eq. (11)). The corresponding coordinates in the learned latent space are the seeds for latent particles $\mathbf{z}_1^k(1:P^c)$ with P^c particles. P^c is determined by Eqs. (12)-(14) given P particles in total. With the seeds and particle number, the latent particles $\mathbf{z}_1^k(1:P^c)$ for the first frame are generated by a Gaussian distribution. Those latent particles are further mapped to the pose particles $\mathbf{x}_1(1:P)$ ($P=\sum P^c$ is the total particle number). The importance weights $w_t(1:P)$ are equally set as $1/P$.

$$dist(\mathbf{x}_1^{gt}, \mathbf{x}_t^*) < \overline{dist} \text{ and } \frac{d(dist(\mathbf{x}_1^{gt}, \mathbf{x}_t^*))}{dt} < 0 \tag{11}$$

$$-\log q(c) = dist(\mathbf{x}_1^{gt}, \mathbf{x}_t^*(c)) / \sum_i dist(\mathbf{x}_1^{gt}, \mathbf{x}_t^*(i)) \tag{12}$$

$$p(c) = q(c) / \sum_i q(i) \tag{13}$$

$$P^c = p(c) * P \tag{14}$$

where $\mathbf{x}_1^{gt}$ denotes the ground truth of the first frame, $\mathbf{x}_t^*(c)$ denotes a pose candidate, and $\overline{dist}$ denotes the average distance for all the pose candidates.

Then, the human pose is inferred by the following steps iteratively. Note that this scheme can easily be extended to variants of the particle filters, such as the annealed particle filter [13].

1. **Likelihood calculation:** With the pose particles $\mathbf{x}_t(1:P)$ and video frame $\mathbf{y}_t$, the importance weights $\hat{w}_t(1:P)$ are updated by the same likelihood functions as [13], which includes the edge and silhouette features in the video frame.
2. **Resampling:** According to the updated importance weights, resample the latent particles $\hat{\mathbf{z}}_t^k(1:P^k)$, which is similar to [13].
3. **Prediction by inter-transition:** This step is unique for our GPDM graph model. The above latent particles are checked whether they should be transited to other latent spaces. By this step, the particles are adaptively distributed among the latent spaces. Since all possible inter-transitions are learned in section 3.2, the distances are calculated between $\{\mathbf{b}_i : i = 1, ..., I\}$ and each pose particle $\mathbf{x}_t(p)$, which is mapped from a latent particle $\hat{\mathbf{z}}_t^k(p)$. If the distance with $\mathbf{b}_i$ and $\mathbf{x}_t(p)$ is smaller than the threshold, the latent particle $\hat{\mathbf{z}}_t^k(p)$ will be transited to the k'-th latent space corresponding to the human pose $\mathbf{b}_i$. The transited particle number is determined by the distances and the original particle number, which is similar to Eq. (14).
4. **Prediction by intra-transition:** Although this step exists in conventional particle filters, much more advanced dynamics is available in the latent space using GPDM models [7,16]. The purpose of this step is to generate latent particles at the next time instant $\mathbf{z}_{t+1}^k(1:P^k)$, which is calculated by the learned dynamics in Eq. (5).
5. **Mapping to pose particles:** With the above latent particles $\mathbf{z}_{t+1}^k(1:P^k)$, the pose particles $\mathbf{x}_{t+1}(1:P)$ are obtained by the mapping function in Eq. (6). Now go to Step (1) for tracking human pose in the next frame.

4 Experimental Results

Experimental Conditions: In Section 4, we evaluate our algorithm in both the learning and inference stages using the HumanEva dataset [13], where the training and test data from S1 subject are used as shown in Table 2.

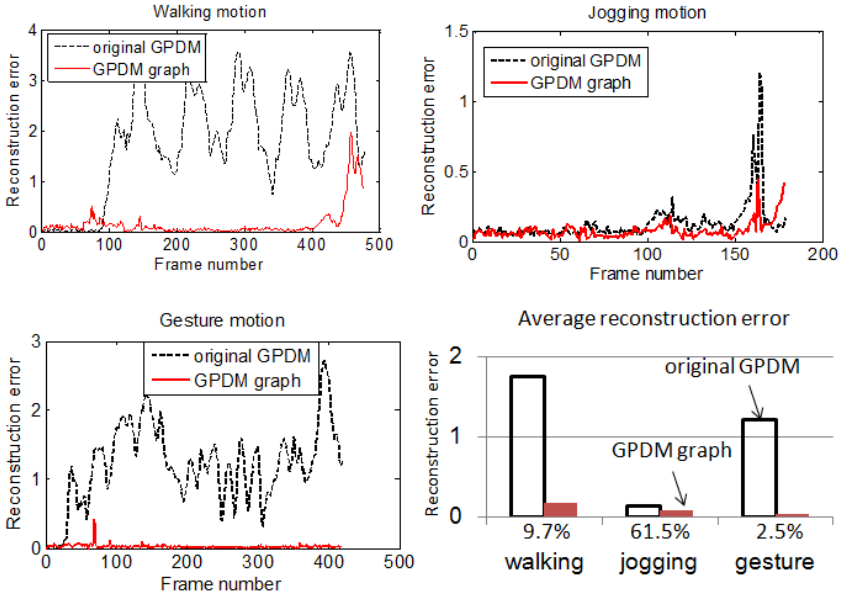


Fig. 3. Comparison of reconstruction error

Evaluation of GPDM Graph Learning: We compare our GPDM graph with the original GPDM model [17] for the three motions in Table 2. The reconstruction errors are shown in Fig. 3, where the average errors are reduced to 9.7%, 61.5%, and 2.5% in the three motions, respectively. As expected, the model precision is much improved. Basically, the more complex the motion is, e.g. gesture motion, the more benefit the proposed method provides.

Figure 4 shows the inter-transition candidates for training data. The frame distance, which means the probability of inter-transition in our method, changes a lot in Fig. 4, requiring that the transition probability should dynamically depend on the current state. Similar results were reported in motion graph technique [3,8,10]. At the same time, the inter-transition candidates should locate the similar poses with short distances in those cyclic motions. The experiments show our inter-transition candidate detector works well, which detects the local extreme values by short-term PCA [19] as shown by the crosses in Fig. 4.

Table 2. Experimental data used in the learning and inference stages

motion	description	training data	test data (C1 camera)
walking	cyclic motion	1∼480	481∼600
jogging	cyclic motion	1∼180	531∼650
gesture	multiple patterns	1∼420	421∼570

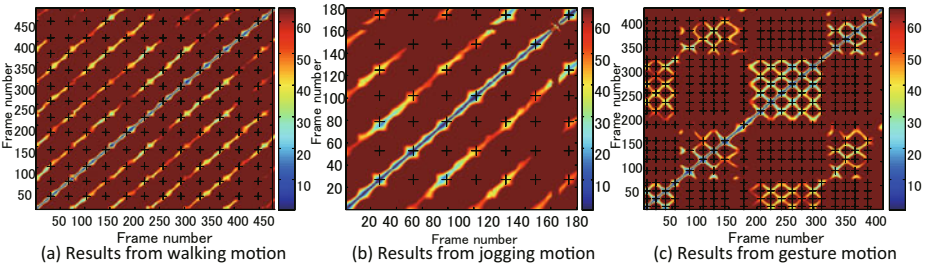


Fig. 4. Frame distances and detected candidates for inter-transitions from a walking motion (a), a jogging motion (b), and a gesture motion (c). Blue color denotes the low distance and deep red color denotes the high distance. Crosses denote the detected candidates for inter-transitions.

An interesting observation from Figs. 2 and 3 is that there are multiple patterns in a semantically simple walking motion. This is due to the following fact that the signals in two cycles are rather different. Figure 5 (a) shows the learned latent space from the first two cycles (frame #1∼#150) of the walking motion in Table 2. It is clear that the predicted latent coordinates (crosses, generated by the GPDM model) are almost the same in two cycles while the learned latent coordinates (circles, learned directly from the training data) are quite different, which infers that the learned GPDM model cannot confidently generate correct latent coordinates and leads to the reconstruction errors become rather large in the second cycle as shown in Fig. 5(b). By segmenting into two models, the reconstruction ability is greatly improved as shown in Fig. 2.

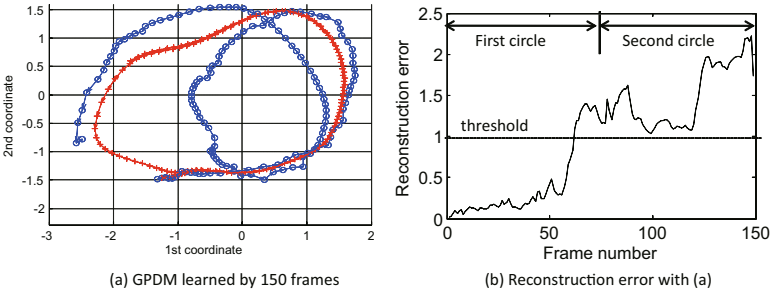


Fig. 5. Learned latent space from frame #1∼#150 of the walking motion (about two cycles), where the circles denote the learned coordinates in latent space and the crosses denote the predicted coordinates in latent space. A single GPDM model may fail to model a semantically simple motion.

Evaluation of Pose Inference: We compare the GPDM graph model with the original GPDM model [17] and the switching GPDM model [7] where the probability of inter-transitions is constant (i.e. independent of the latent state

Table 3. Average errors of tracking human motion by different methods

motion	original GPDM	switching GPDM	GPDM graph
walking	44.16 mm	56.09 mm	40.88 mm
jogging	57.21 mm	57.55 mm	53.26 mm
gesture	17.80 mm	16.33 mm	13.23 mm

in GPDMs). In all the methods, the total particle number is set as 1000. For evaluation, the tracking error is calculated by the inferred pose and the ground truth as described by Sigal et al. [13].

Figure 6 shows the tracking errors of the above three methods respectively, whose average error is listed in Table 3. In the above experiments, the proposed GPDM method achieves the best performance by combining the merits of original GPDM and switching GPDM[1]. As Fig. 6 shows, the GPDM graph method basically has the errors similar to the lower ones of the original GPDM and the switching GPDM. When the motion is in a single pattern, the particles in particle filter are preferred to stay in a GPDM model. On the other hand, when the motion transits to a new pattern, the particles are preferred to transit to another GPDM model. Our experimental results infer that neither the original GPDM nor the switching GPDM deals with the situations well. In this meaning, by the adaptive probability of inter-transitions, the efficiency of using particles in the particle filter is improved in the proposed GPDM graph model, leading to better performance. Figure 7 shows the particles are transited among different GPDM models by GPDM graph and switching GPDM respectively. As the dashed line in Fig. 7 (a) shows, the particles are transited properly with the motion patterns in GPDM graph while they are equally transited in switching GPDM as Fig. 7 (b) shows. Basically, in the proposed GPDM graph, the particles can automatically follow the changes of motion patterns by the adaptive transition probability among different GPDM models, which is the essential advantage of our method.

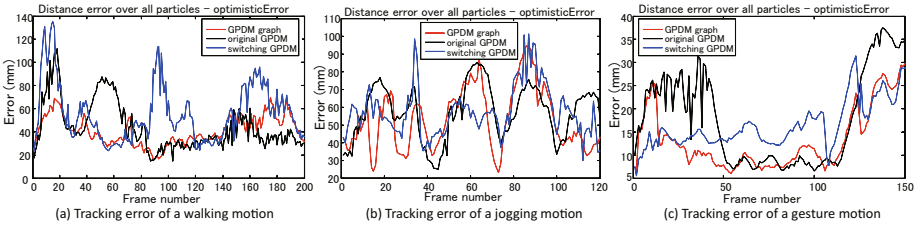
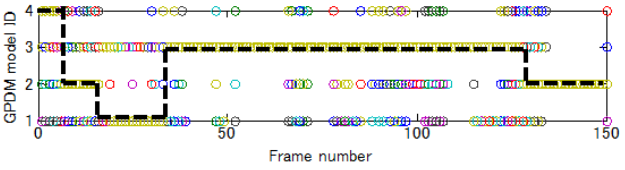
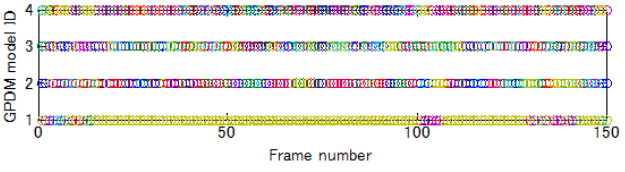


Fig. 6. Tracking errors by the proposed GPDM graph (red dotted curves), the original GPDM (black solid curves), and the switching GPDM (blue dashed curves) in a walking video (a), a jogging video (b), and gesture video (c) of the S1 subject from the C1 camera

[1] As a latest result on walking motion of S1 subject in HumanEva dataset, Taylor et al. reported an average error of 47.29 mm by a sixth-order model of Implicit Mixture of Conditional Restricted Boltzmann Machines in a similar condition [14].



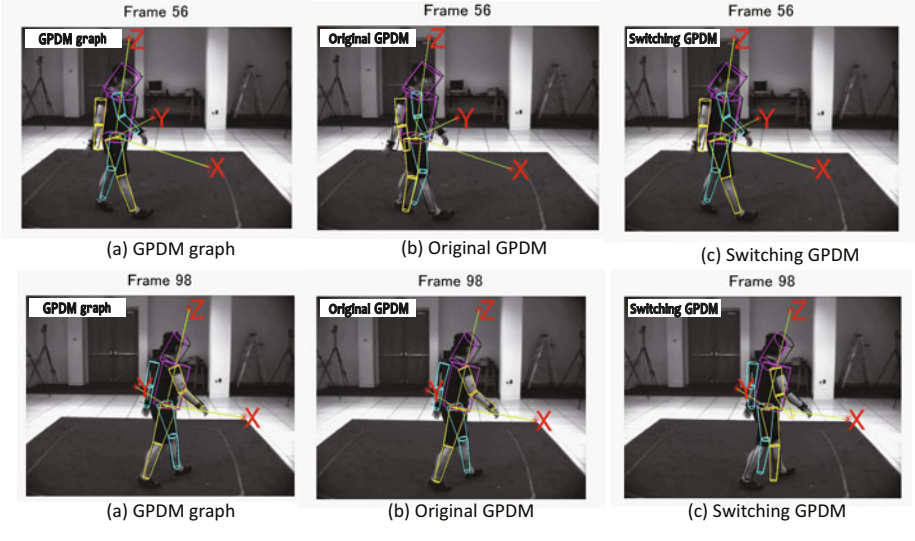
(a) Particle transition in GPDM graph



(b) Particle transition in switching GPDM

Fig. 7. Particle transitions among different GPDM models by GPDM graph (a) and switching GPDM (b) from the gesture motion. The dashed line shows the transition trace of most particles in GPDM graph. The particles are transited properly with the motion patterns in GPDM graph (a) while they are equally transited in switching GPDM (b). The color of points denotes the particle ID.

Finally, we show two samples in Fig. 8 where the proposed method tracks the pose correctly while other methods may fail to track the legs.



(a) GPDM graph (b) Original GPDM (c) Switching GPDM

(a) GPDM graph (b) Original GPDM (c) Switching GPDM

Fig. 8. Tracking result of frame #56 and #98 by GPDM graph (a), original GPDM (b), and switching GPDM (c) in the test video of walking motion. The colored cylinders show the tracking results and the black cylinders denote the ground truth.

5 Conclusions and Future Work

In this paper, our main contribution is to propose a novel model for tracking human motion from a monocular video, where the novelties are as follows.

- It is the first latent dynamics model with graph structure. With inter-transitions in the graph, the long-term correlation is possible to be used. We simultaneously segment the training motion and learn the GPDM models by the trial and error approach.
- Our data-driven approach is a completely unsupervised learning method. For this purpose, we employ the short-term PCA method to search the candidates for inter-transitions. In the inference stage, the connections (inter-transitions) are dynamically determined by the similarity of human poses, which is inspired by the motion graph technique [3,8,10].

In the future, we plan to improve the likelihood function in the tracking stage using more advanced features, such as robust local and global appearance features [5,16].

Acknowledgment. The HumanEva dataset and baseline codes are provided by Brown University. Part of GPDM codes are downloaded from Toronto University.

References

1. Agarwal, A., Triggs, B.: Recovering 3d human pose from monocular images. IEEE Trans. on PAMI 28(1), 44–58 (2006)
2. Arikan, O.: Compression of motion capture databases. ACM Trans. on Graphics 25(3), 890–897 (2006)
3. Arikan, O., Forsyth, D.A.: Interactive motion generation from examples. ACM Trans. on Graphics 21(3), 483–490 (2002)
4. Arulampalam, M., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. IEEE Trans. on Signal Processing 50(2), 174–188 (2002)
5. Balan, A., Black, M.J.: An adaptive appearancemodel approach formodel-based articulated object tracking. In: IEEE CVPR, vol. 1, pp. 758–765 (2006)
6. Moeslund, T.B., Hilton, A., Kruger, V.: A survey of advances in vision based human motion capture and analysis. CVIU 104(2), 90–126 (2006)
7. Chen, J., Kim, M., Wang, Y., Ji, Q.: Switching gaussian process dynamic models for simultaneous composite motion tracking and recognition. In: IEEE CVPR, pp. 2655–2662 (2009)
8. Kovar, L., Gleicherl, M., Pighin, F.: Motion graphs. ACM Trans. on Graphics 21(3), 473–482 (2002)
9. Lawrence, N.: Gaussian process latent variable models for visualization. In: Proc. Adv. Neural Inf. Process. pp. 329–336 (2003)
10. Lee, J., Chai, J., Reitsma, P.S.A., Hodgins, J.K., Pollard, N.S.: Interactive control of avatars animated with human motion data. ACM Trans. on Graphics 21(3), 491–500 (2002)

11. Poppe, R.: Vision-based human motion analysis: an overview. CVIU 108(1/2), 4–18 (2007)
12. Quirion, S., Duchesne, C., Laurendeau, D., Marchand, M.: Comparing gplvm approaches for dimensionality reduction in character animation. Journal of WSCG 16(1-3), 41–48 (2008)
13. Sigal, L., Balan, A., Black, M.J.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal of Computer Vision 87(1), 4–27 (2010)
14. Taylor, G.W., Sigal, L., Fleet, D.J., Hinton, G.E.: Dynamical binary latent variable models for 3d human pose tracking. In: IEEE CVPR, pp. 631–638 (2010)
15. Urtasun, R., Darrell, T.: Sparse probabilistic regression for activity independent human pose inference. In: IEEE CVPR, pp. 1–8 (2008)
16. Urtasun, R., Fleet, D., Fua, P.: 3d people tracking with gaussian process dynamical models. In: IEEE CVPR, pp. 238–245 (2006)
17. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models for human motion. IEEE Trans. on PAMI 30(2), 283–298 (2008)
18. Wang, J., Bodenheimer, B.: Synthesis and evaluation of linear motion transitions. ACM Trans. on Graphics 27(1), 1:1–1:15 (2008)
19. Xu, J., Takagi, K., Yoneyama, A.: Beat induction from motion capture data using short-term principal component analysis. The Journal of The Institute of Image Information and Television Engineers 64(4), 577–583 (2010)