Manifold regularized particle filter for articulated human motion tracking

Adam Gonczarek, Jakub M. Tomczak

Institute of Computer Science, Wrocław University of Technology,
wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
{adam.gonczarek,jakub.tomczak}@pwr.wroc.pl

Abstract. In this paper, a fully Bayesian approach to articulated human motion tracking from video sequences is presented. First, a filtering procedure with a low-dimensional manifold is derived. Next, we propose a general framework for approximating this filtering procedure based on the particle filter technique. The low-dimensional manifold can be treated as a regularizer which restricts the space of all possible distributions to the space of distributions concentrated around the manifold. We refer to our method as *Manifold Regularized Particle Filter*. The proposed approach is evaluated using real-life benchmark dataset *HumanEva*.

Keywords: articulated motion tracking, manifold regularization, particle filter, generative approach, Gaussian process latent variable model

1 Introduction

Articulated human motion tracking from video image sequences is one of the most challenging computer vision problems for the past two decades. The basic idea behind this issue is to recover a motion of the complete human body basing on the image evidence from a single or many cameras, and without using any additional devices, e.g., color or electromagnetic markers. Such system can be applied in many everyday life areas, see [8, 9]. Pointing out only a few, motion tracking may be used in control devices for Human-Computer Interaction, surveillance systems detecting unusual behaviors, dancing or martial arts training assistants, support systems for medical diagnosis.

During last years, a lot of effort has been put in solving the human motion tracking issue. However, excluding some minor cases, the problem is still open. There are several reasons worth mentioning that make the issue very difficult. First, there is a huge variety of different images corresponding to the same pose that may be obtained. This is caused by variability in human wear and appearance, changes in lighting conditions, camera noise, etc. Second, image lacks of depth information which makes impossible to obtain three-dimensional pose from two-dimensional images. Moreover, one has to handle different types of occlusions including self-occlusions and occlusions caused by external environment. Finally, efficient exploration of the space of all possible human poses

is troublesome because of high-dimensionality of the space and its non-trivial constraints.

To date, however, several conceptually different approaches have been proposed to address the human motion tracking problem. They can be roughly divided into two groups. In the first one, discriminative methods are used to model directly the probability distribution over poses conditioned on the image evidence, see [1, 3, 7]. On the contrary, in the second group a generative approach is used to model separately the prior distribution over poses and the likelihood of how well a given pose fits to the current image. Pure generative modeling assumes that one tries to model the true pose space as accurately as it is possible and uses Bayesian inference to estimate current pose, see [2, 6, 11, 13, 14]. Recent studies show that using more flexible models, i.e., part-based models and searching maximum a posteriori estimate (MAP) give very promising results, see [15]. However, these approaches are mainly applied to 2D pose estimation problems.

The contribution of the paper is threefold:

1. A general framework for human motion tracking using generative modeling and hidden low-dimensional manifold is proposed.
2. A particle filter regularized using low-dimensional manifold is introduced. Further, we refer to this particle filter as *Manifold Regularized Particle Filter* (MRPF).
3. A dynamics model using Gaussian process latent variable model (GPLVM) and low-dimensional manifold is presented.

Empirical results are evaluated on benchmark dataset HumanEva, see [11] for details.

The paper is organized as follows. In Section 2 the problem of the human motion tracking is outlined. First, the problem of pose estimation is given in section 2.1 and then the human motion tracking problem is stated in section 2.2. Next, the likelihood function is formulated in section 4. In Section 3 the particle filter with low-dimensional manifold is proposed. The model of dynamics with low-dimensional manifold is presented in section 5. At the end, the empirical study is carried out in section 6 and conclusions are drawn in section 7.

2 Human motion tracking

In this paper, we assume that a human body is represented by a set of articulately connected rigid parts. Each connection between two neighboring elements characterize a joint and can be described by up to three degrees of freedom, depending on movability of the joint. All connected parts form a kinematic tree with the root typically associated with the pelvis. A common representation of the state of the k^{th} joint uses Euler angles which describe relative rotation between neighboring parts in the kinematic tree. However, we prefer to use quaternions because they can be compared using the Euclidean metric. In case of angles of

rotation of the parts in the kinematic tree in the range between 0 and π we can take advantage of an approximation of quaternions (for details see [12]).

The set of quaternions for all K joints together with the global position and orientation of the kinematic tree in 3D constitutes the minimal set of variables that are used to describe the current state of the human body, which is denoted by $\mathbf{x}$. It is worth mentioning that $\mathbf{x}$ is usually around 40-50 dimensions, which is one of the fundamental reasons that makes the human motion tracking a difficult problem.

We assume that there are several synchronized cameras which provides video images of a human body from different perspectives. The cameras should be located so that to contribute as much information about the body as possible, i.e., they should register different parts of a scene. Let $\mathcal{I}$ denote a set of all available images from all cameras at current moment. Hence, we want to infer the human body configuration $\mathbf{x}$ basing on $\mathcal{I}$.

In next sections, first we formulate pose estimation problem and then state the human motion tracking problem.

2.1 Pose estimation problem statement

The goal of the pose estimation issue is to find a human pose estimate $\hat{\mathbf{x}}$ in every video frame, basing on all available images $\mathcal{I}$. Since this is a typical multivariate regression problem and the optimal solution (in the sense of decision making), that is, $\hat{\mathbf{x}} = \mathbb{E}[\mathbf{x}|\mathcal{I}]$.

The key issue is to model properly the true distribution $p(\mathbf{x}|\mathcal{I})$. It can be accomplished using discriminative models, i.e., explicit modeling of $p(\mathbf{x}|\mathcal{I})$, or generative models by applying Bayes' rule, $p(\mathbf{x}|\mathcal{I}) \propto p(\mathcal{I}|\mathbf{x})p(\mathbf{x})$, and then modeling $p(\mathcal{I}|\mathbf{x})$ and $p(\mathbf{x})$ separately. In this paper, we focus on the second approach.

2.2 Human motion tracking problem statement

Now let us extend the problem of the pose estimation to tracking the whole sequence of states $\mathbf{x}_{1:T} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, also called a *trajectory*. Additionally, we need to maintain the sequence of images $\mathcal{I}_{1:T} = \{\mathcal{I}_1, \ldots, \mathcal{I}_T\}$. The sequence of video frames corresponds to images achieved from all cameras in consecutive moments $1, \ldots, T$.

It is a fact that the high-dimensional pose space consists of human body configurations and most of them are unrealistic. Additionally, during specific motions (e.g. walking or running) all state variables exhibit strong correlations which depends on the current pose. These two remarks yields a corollary that the real trajectories of motion form a low-dimensional manifold.

We assume that any state $\mathbf{x}$ corresponds to a point on a low-dimensional manifold $\mathbf{z}$. Hence, the trajectories of human motion formulate a pattern which locally oscillates around the low-dimensional manifold. Further, we are interested in representing the joint probability distribution $p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}, \mathcal{I}_{1:T})$. The manner how it is factorized is graphically presented by the probabilistic graphical model in Figure 1. Notice that the current state $\mathbf{x}_t$ influences future state

and future point on the manifold $\mathbf{z}_{t+1}$ which in turn impacts $\mathbf{x}_{t+1}$. In the literature, several similar models have been proposed, e.g., in [16] a model assumes that the temporal dependence exists between low-dimensional variables only, in [13] a conditional Restricted Boltzmann Machine is used to represent information about the low-dimensional manifold which leads to undirected dependence between $\mathbf{x}$ and $\mathbf{z}$.

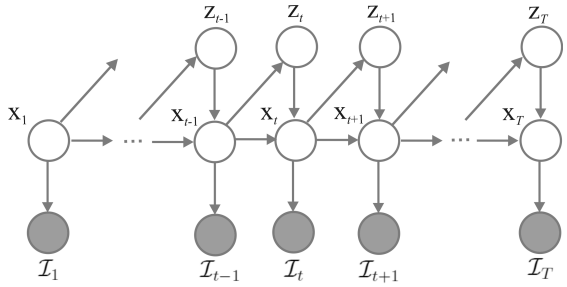

Fig. 1. Probabilistic graphical model with a low-dimensional manifold represented by variables $\mathbf{z}$.

We are interested in calculating the a posteriori probability distribution for $\mathbf{x}_t$ given images $\mathcal{I}_{1:t}$ by marginalizing $p(\mathbf{x}_{1:t}, \mathbf{z}_{1:t}|\mathcal{I}_{1:t})$ over all state variables $\mathbf{x}_{1:t-1}$ and hidden variables $\mathbf{z}_{1:t}$ which yields:

$$p(\mathbf{x}_t|\mathcal{I}_{1:t}) = \frac{p(\mathcal{I}_t|\mathbf{x}_t)}{p(\mathcal{I}_t|\mathcal{I}_{1:t-1})} \iint p(\mathbf{z}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t) p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}) \mathrm{d}\mathbf{x}_{t-1} \mathrm{d}\mathbf{z}_t, \tag{1}$$

where $p(\mathcal{I}_t|\mathcal{I}_{1:t-1})$ is the normalization constant.

We have obtained a filtering procedure which includes information about the low-dimensional manifold. Further, we need to determine the following aspects:

- an algorithm which allows to perform the filtering procedure (1);
- the likelihood function $p(\mathcal{I}_t|\mathbf{x}_t)$;
- models of dynamics: $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$ and $p(\mathbf{z}_t|\mathbf{x}_{t-1})$.

3 Manifold Regularized Particle Filter

In the context of the human motion tracking the filtering procedure is intractable and hence an approximation of (1) should be applied. Typically, a sampling method like a particle filter technique is applied. However, the main disadvantage of the particle filter is that it needs to generate a huge amount of particles in order to cover a high-dimensional state space. Otherwise, it fails to approximate true distribution. In order to cover the highly probable areas in the pose space only, an extension of the particle filter technique, called *Annealed Particle Filter*

(APF), has been proposed [2]. However, this method tends to be trapped in one or a few dominating extrema. Therefore, in the context of the human motion tracking, it is non-robust to noisy likelihood model and thus fails to track the proper trajectory.

In this paper, we propose a different approach which modifies the particle filter by introducing a regularization in a form of the low-dimensional manifold. This filtering procedure operates in the neighborhood of the low-dimensional space where the true poses are concentrated, and thus it guarantees that highly probable regions are covered and the particles are distributed around different local extrema.

First, let us remind that we want to calculate $p(\mathbf{x}_t|\mathcal{I}_{1:t})$. If we were able to sample from $p(\mathbf{x}_t|\mathcal{I}_{1:t-1})$, it would be possible to approximate (1), concentrated on points $\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(N)} \sim p(\mathbf{x}_t|\mathcal{I}_{1:t-1})$:

$$p(\mathbf{x}_t|\mathcal{I}_{1:t}) \approx \sum_{n=1}^{N} \pi(\mathbf{x}_t^{(n)})\delta(\mathbf{x}_t - \mathbf{x}_t^{(n)}), \tag{2}$$

where $\pi(\mathbf{x}_t^{(n)})$ is a normalized form of a single score calculated using the following expression $\tilde{\pi}(\mathbf{x}_t^{(n)}) = p(\mathcal{I}_t|\mathbf{x}_t^{(n)})$.

Usually, it is troublesome to generate $\mathbf{x}_t$ for given $\mathbf{x}_{t-1}$ and $\mathbf{z}_t$ using the distribution $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$ and thus we will introduce an auxiliary distribution $q(\mathbf{x}_t|\mathbf{x}_{t-1})$. Then, taking advantage of dependencies defined by the probabilistic graphical model in Figure 1, we get:

$$p(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1}) = \frac{1}{Z}\tilde{\omega}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t)Q(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1}), \tag{3}$$

where $\tilde{\omega}$ are weights:

$$\tilde{\omega}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t) = \frac{p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \tag{4}$$

and Q is an auxiliary joint distribution:

$$Q(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1}) = q(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{z}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}), \tag{5}$$

and Z is a normalization constant.

Eventually, we can approximate the a posteriori distribution (1) using the following formula:

$$p(\mathbf{x}_t|\mathcal{I}_{1:t}) \approx \sum_{n=1}^{N} \omega(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})\delta(\mathbf{x}_t - \mathbf{x}_t^{(n)}), \tag{6}$$

where the normalized weights are defined as follows:

$$\omega(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)}) = \frac{\tilde{\pi}(\mathbf{x}_t^{(n)})\tilde{\omega}(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})}{\sum_{j=1}^{N} \tilde{\pi}(\mathbf{x}_t^{(j)})\tilde{\omega}(\mathbf{x}_t^{(j)}, \overline{\mathbf{x}}_{t-1}^{(j)}, \mathbf{z}_t^{(j)})}. \tag{7}$$

Notice that we introduce the low-dimensional manifold in the manner that the particles are weighted by ω. The procedure of MRPF is presented in Algorithm 1.

Algorithm 1: Manifold Regularized Particle Filter

Input : initial state $\mathbf{x}_0$, sequence of images $\mathcal{I}_{1:T}$
Output: sequence of state estimates $\hat{\mathbf{x}}_{1:T}$

Duplicate the initial state $\mathbf{x}_0$ and formulate a set: $\overline{\mathcal{X}}_0 = \{\overline{\mathbf{x}}_0^{(1)}, \ldots, \overline{\mathbf{x}}_0^{(N)}\}$;
for $t = 1 : T$ **do**
  Generate a sample $\mathcal{Z}_t = \{\mathbf{z}_t^{(1)}, \ldots, \mathbf{z}_t^{(N)}\}$ using $\mathbf{z}_t^{(n)} \sim p(\mathbf{z}_t|\overline{\mathbf{x}}_{t-1}^{(n)})$;
  Generate a sample $\mathcal{X}_t = \{\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(N)}\}$ using $\mathbf{x}_t^{(n)} \sim q(\mathbf{x}_t|\overline{\mathbf{x}}_{t-1}^{(n)})$;
  Calculate $\tilde{\pi}(\mathbf{x}_t^{(n)})$ using the likelihood model $p(\mathcal{I}_t|\mathbf{x}_t)$;
  Calculate $\tilde{\omega}(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})$ using (4);
  Normalize the weights $\omega(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})$ using (7);
  Calculate the estimate of the state variables
  $\hat{\mathbf{x}}_t = \sum_{n=1}^{N} \omega(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)}) \mathbf{x}_t^{(n)}$;
  Generate a sample $\overline{\mathcal{X}}_t = \{\overline{\mathbf{x}}_t^{(1)}, \ldots, \overline{\mathbf{x}}_t^{(N)}\}$ using approximation (6);
end

4 Likelihood function

The likelihood function $p(\mathcal{I}_t|\mathbf{x}_t)$ aims at evaluating the given human body configuration $\mathbf{x}_t$ corresponds to the set of images $\mathcal{I}_t$. We compare images which contain a human body model projected onto camera views with binary silhouettes obtained from background subtraction procedure, by calculating the difference between them. This model is called *bidirectional silhouette likelihood* [11].

5 Dynamics model using low-dimensional manifold

The simplest model used for modeling the dynamics of the pose $\mathbf{x}$ is a Gaussian diffusion, i.e., new state is the old state disturbed by an independent Gaussian noise. However, this model turns to be too simplistic in the context of the human motion tracking. Therefore, we propose to model the dynamics of the pose using low-dimensional manifold and a nonlinear dependency. First, we need to learn the low-dimensional manifold. Second, a model for dynamics on the low-dimensional manifold has to be proposed, $p(\mathbf{z}_t|\mathbf{x}_{t-1})$. Third, the model of dynamics in the pose space with the low-dimensional manifold should be given, $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$.

5.1 Learning the low-dimensional manifold

For learning the low-dimensional manifold we apply the *Gaussian Process Latent Variable Model* (GPLVM) [4]. The GPLVM model constitutes a non-linear dependency between the pose and the low-dimensional manifold as follows: $\mathbf{x} = \mathbf{f}(\mathbf{z}) + \boldsymbol{\varepsilon}$, where i^{th} function is a realization of the Gaussian process [10], $f_i \sim \mathcal{GP}(f|0, k(\mathbf{z}, \mathbf{z}'))$, where k is a kernel function, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\varepsilon}|0, \sigma_z^2\mathbf{I})$, where σ_z^2 is variance, and $\mathbf{I}$ denotes the identity matrix. In this paper, we use the RBF kernel, $k(\mathbf{z}, \mathbf{z}') = \beta \exp\left(-\frac{\gamma_z}{2}\|\mathbf{z} - \mathbf{z}'\|^2\right) + \beta_0$.

We are interested in finding a matrix of low-dimensional variables corresponding to observed poses, i.e., a matrix $\mathbf{Z}$ for observed poses $\mathbf{X}$. Additionally, we want to determine the mapping between the manifold and the high-dimensional space by learning parameters β, β_0 and γ_z, and σ_z^2. The training corresponds to finding the parameters and points on the manifold that maximize the logarithm of the likelihood function in the following form:

$$\begin{aligned}\ln p(\mathbf{X}|\mathbf{Z}) &= \ln \prod_{i=1}^{D} \mathcal{N}(\mathbf{X}_{:,i}|0, \mathbf{K} + \sigma_z^2 \mathbf{I}_{T\times T}) \\ &= -\frac{DT}{2}\ln(2\pi) - \frac{D}{2}\ln|\overline{\mathbf{K}}| - \frac{1}{2}\mathrm{tr}(\mathbf{X}^{\mathrm{T}}\overline{\mathbf{K}}^{-1}\mathbf{X}),\end{aligned} \tag{8}$$

where $\mathbf{X}_{:,i}$ denotes i^{th} column of the matrix $\mathbf{X}$, $|\cdot|$ and $\mathrm{tr}(\cdot)$ are matrix determinant and trace, respectively, $\overline{\mathbf{K}} = \mathbf{K} + \sigma_z^2\mathbf{I}_{T\times T}$, and $\mathbf{K} = [k_{nm}]$ is the kernel matrix with elements $k_{nm} = k(\mathbf{z}_n, \mathbf{z}_m)$.

Let us notice that solutions of the maximization $\mathbf{z}_t$ and γ_z can be arbitrarily re-scaled, thus there are many equivalent solutions. In order to avoid this issue we introduce a regularizer $\frac{1}{2}\|\mathbf{Z}\|_F^2$, where $\|\cdot\|_F$ is the Frobenius norm, and the final objective function takes the form $L(\mathbf{Z}) = \ln p(\mathbf{X}|\mathbf{Z}) - \frac{1}{2}\|\mathbf{Z}\|_F^2$. The objective function can be optimized using standard numerical optimization algorithms, e.g., scaled conjugate gradient method. Additionally, the objective function is not concave and hence have multiple local maxima. Therefore, it is important to initialize the numerical algorithm properly, e.g., by using principal component analysis.

The kernel function used to determine the covariance function takes high values for points $\mathbf{z}_n$ and $\mathbf{z}_m$ that are close to each other, i.e., they are similar. Moreover, because the points on the manifold are similar, the original poses $\mathbf{x}_n$ and $\mathbf{x}_m$ are similar as well. However, the situation does not hold in the opposite direction. This issue is undesirable in the proposed filtering procedure 1 because the distribution $p(\mathbf{z}_t|\mathbf{x}_{t-1})$ is multi-modal and thus hard to determine. However, this effect can be decreased by introducing *back constraints* which leads to *Back-Constrained* GPLVM (BC-GPLVM) [5].

The idea behind BC-GPLVM is to define $\mathbf{z}$ as a smooth mapping of $\mathbf{x}$, $\mathbf{z} = \mathbf{g}(\mathbf{x})$. For example, this mapping can be given in the linear form, i.e., $g_i(\mathbf{x}) = \sum_{t=1}^{T} a_{ti}k_x(\mathbf{x}, \mathbf{x}_t) + b_i$, where g_i denotes i^{it} component of $\mathbf{z}$, a_{ti}, b_i are parameters, and $k_x(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\gamma_x}{2}\|\mathbf{x} - \mathbf{x}'\|^2\right)$ is the kernel function in the high-dimensional space of poses. We can incorporate the mapping into the objective function, i.e., $z_n^i = g_i(\mathbf{x}_n)$, and then optimize w.r.t. a_{ti} and b_i instead of z_n^i. The application of back constrains entails closeness of low-dimensional points $\mathbf{z}_t$ if high-dimensional points $\mathbf{x}_t$ are similar.

The big advantage of Gaussian processes is tractability of calculating the predictive distribution for new pose $\mathbf{x}_p$ and its low-dimensional representation $\mathbf{z}_p$. The corresponding kernel matrix is as follows:

$$\begin{bmatrix} \overline{\mathbf{K}} & \overline{\mathbf{k}} \\ \overline{\mathbf{k}}^{\mathrm{T}} & \overline{k}_z(\mathbf{z}_p, \mathbf{z}_p) \end{bmatrix}, \tag{9}$$

and finally the predictive distribution [10]:

$$p(\mathbf{x}_p|\mathbf{z}_p, \mathbf{X}, \mathbf{Z}) = \mathcal{N}(\mathbf{x}_p|\boldsymbol{\mu}_p, \sigma_p^2 \mathbf{I}_{D\times D}), \tag{10}$$

where:

$$\boldsymbol{\mu}_p = \mathbf{X}^{\mathrm{T}} \overline{\mathbf{K}}^{-1} \overline{\mathbf{k}}, \tag{11}$$

$$\sigma_p^2 = \overline{k}_z(\mathbf{z}_p, \mathbf{z}_p) - \overline{\mathbf{k}}^{\mathrm{T}} \overline{\mathbf{K}}^{-1} \overline{\mathbf{k}}. \tag{12}$$

5.2 Dynamics on the manifold

Idea of the model $p(\mathbf{z}_t|\mathbf{x}_{t-1})$ is to predict new position on the manifold basing on the previous pose. Therefore, we need a mapping which allows to transform a high-dimensional representation to a low-dimensional one. For this purpose we apply the back constraints. Adding Gaussian noise with the covariance matrix $\mathrm{diag}(\boldsymbol{\sigma}^2_{x\to z})$ to the back constraints, we obtain the following model of the dynamics on the manifold:

$$p(\mathbf{z}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{z}_t|\mathbf{g}(\mathbf{x}_{t-1}), \mathrm{diag}(\boldsymbol{\sigma}^2_{x\to z})). \tag{13}$$

5.3 Dynamics in the pose space with the low-dimensional manifold

The model $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$ determines the probability of the current pose basing on the previous pose and the current point on the low-dimensional manifold. A reasonable assumption is that the model factorizes into two components, namely, one concerning only previous pose, and second – the low-dimensional manifold. This factorization follows from the fact that these two quantities belong to two different spaces and thus are hard to compare quantitatively. Then, the model of dynamics takes the following form:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t) \propto p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_t|\mathbf{z}_t). \tag{14}$$

The first component is expressed as a normal distribution with the diagonal covariance matrix $\mathrm{diag}(\boldsymbol{\sigma}^2_{x\to x})$:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathrm{diag}(\boldsymbol{\sigma}^2_{x\to x})). \tag{15}$$

The second component is constructed using the mean of the predictive distribution (11) and is disturbed by a Gaussian noise with the diagonal covariance matrix $\mathrm{diag}(\boldsymbol{\sigma}^2_{z\to x})$ which leads to the following model:

$$p(\mathbf{x}_t|\mathbf{z}_t) = \mathcal{N}(\mathbf{x}_t|\mathbf{X}^{\mathrm{T}} \overline{\mathbf{K}}^{-1} \overline{\mathbf{k}}, \mathrm{diag}(\boldsymbol{\sigma}^2_{z\to x})). \tag{16}$$

It is important to highlight that the training of the parameters $\mathrm{diag}(\boldsymbol{\sigma}^2_{z\to x})$ has to be performed using a separate validation set which contains data. Otherwise, using the same training set as for determining $\mathbf{Z}$ leads to underestimation of the parameters.

5.4 Dynamics models and the filtering procedure

At the end, let us consider the application of the particle filter proposed earlier (see Algorithm 1) in the context of the outlined models of dynamics. First, we need to propose the auxiliary distribution $q(\mathbf{x}_t|\mathbf{x}_{t-1})$. In our case it is given in the form (15), i.e., $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathrm{diag}(\boldsymbol{\sigma}^2_{x\to x}))$. Then, the weights $\tilde{\omega}$ are given in the form (16), i.e., $\tilde{\omega}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t) = \mathcal{N}(\mathbf{x}_t|\mathbf{X}^{\mathrm{T}}\overline{\mathbf{K}}^{-1}\overline{\mathbf{k}}, \mathrm{diag}(\boldsymbol{\sigma}^2_{z\to x}))$.

6 Empirical study

The aim of the experiment is to compare the proposed approach using MRPF with methods using the ordinary particle filter (PF) and the annealed particle filter (APF). In both methods Gaussian diffusion as dynamics model was applied. The performance evaluation is measured using real-life benchmark dataset *HumanEva* [11]. The motion sequence is synchronized with measurements from the *MOCAP* system and thus it is possible to evaluate the difference between the true values of pose configuration with the estimated ones using the following equation ($\mathbf{w}(\cdot) \in \mathcal{W}$ denotes M points on a body for given state variables): $\mathrm{err}(\hat{\mathbf{x}}_{1:T}) = \frac{1}{TM}\sum_{t=1}^{T}\sum_{\mathbf{w}\in\mathcal{W}} \|\mathbf{w}(\mathbf{x}_t) - \mathbf{w}(\hat{\mathbf{x}}_t)\|$. The obtained value of the error $\mathrm{err}(\hat{\mathbf{x}}_{1:T})$ is expressed in millimeters.

In the experiment we used two motion types, namely, *walking* and *jogging*, performed by three different persons, i.e., S1, S2, S3, which results in six various sequences. In each sequence we used 350 and 300 frames from different training trials for training and validation sets, respectively. Only the sequence S1-Jog contained 200 and 200 frames in training and validation sets, respectively. For testing we utilized first 200 frames from validation trial.

In the empirical study we used the following number of particles: (i) MRPF with 500 particles, (ii) PF with 500 particles, and (iii) APF with 5 annealing layers with 100 particles each. The low-dimensional manifold had 2 dimensions. All parameters (except $\gamma_{\mathbf{x}} = 10^{-4}$) were set according to the optimization process. The methods were initiated 5 times.

Results and discussion The averaged results obtained within the experiment are gathered in Table 1. The results show that the proposed approach with MRPF gave the best results except the sequence S1-Jog for which PF was slightly better. It is probably caused by the low-quality of this sequence which resulted in shorter training and validation sets. Because of this fact the manifold was not fully discovered.

The worst performance was obtained by the APF. The explanation for such result can be given as follows. First, the likelihood model used in the experiment is highly noised by the low quality of the silhouettes achieved in the background subtraction process. The noise in the likelihood model leads to displacement of extrema and thus wrong tracking. Second, the number of particles can be insufficient.

In the summary, the manifold regularized particle filter seems to correctly follow the trajectory on the low-dimensional manifold. In other words, the prior knowledge about the low-dimensional manifold of the human pose configuration in motion is properly introduced. Additionally, MRPF obtained not only lower average error but also lower standard deviation in comparison to PF and APF. This result allows to presume that MRPF is more stable. However, in order to resolve this issue conclusively more experiments are needed.

Table 1. The motion tracking errors $\mathrm{err}(\hat{\mathbf{x}}_{1:T})$ (in millimeters) for all methods are expressed as an average and a standard deviation (in brackets). The best results are in bold.

Sequence	APF	SIR	MRPF
S1-Walk	107(31)	82(18)	**69**(**7**)
S1-Jog	111(17)	**81**(**4**)	82(8)
S2-Walk	106(16)	95(7)	**86**(**12**)
S2-Jog	121(9)	106(13)	**94**(**8**)
S3-Walk	114(27)	88(13)	**79**(**10**)
S3-Jog	111(27)	117(29)	**70**(**8**)

7 Conclusions

In this paper, a fully Bayesian approach to the articulated human motion tracking was proposed. The modification of the particle filter technique is based on introducing low-dimensional manifold as a regularizer which incorporates prior knowledge about the specificity of human motion. The application of the low-dimensional manifold allows to restrict the space of possible pose configurations. The idea is based on the application of GPLVM with back constraints. At the end of the paper, the experiment was carried out using the real-life benchmark dataset *HumanEva*. The proposed approach was compared with two particle filters, namely, PF and APF, and the obtained results showed that it outperformed both of them.

References

1. Agarwal A., Triggs B.: Recovering 3D human pose from monocular images, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(1):44-58 (2006)
2. Deutscher J., Reid I., Articulated body motion capture by stochastic search, International Journal of Computer Vision, 61(2):185-20 (2005)
3. Kanaujia A., Sminchisescu C., Metaxas D.: Semi-supervised hierarchical models for 3D human pose reconstruction, in CVPR '07 Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (2007)

4. Lawrence N.D.: Probabilistic non-linear principal component analysis with gaussian process latent variable models, Journal of Machine Learning Research, 6:1783-1816 (2005)
5. Lawrence N.D., Quiñonero-Candela J.: Local distance preservation in the GP-LVM through back constraints, in ICML '06 Proceedings of the 23rd international conference on Machine learning, pp. 513-520 (2006)
6. Li R., Tian T., Sclaroff S., Yang M.: 3D human motion tracking with a coordinated mixture of factor analyzers, International Journal of Computer Vision, 87:170-190 (2010)
7. Memisevic R., Sigal L., Fleet D.J.: Shared kernel information embedding for discriminative inference, IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(4):778-790 (2012)
8. Moeslund T.B., Hilton A., Krüger V.: A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding, 104:90-126 (2006)
9. Poppe R.: Vision-based human motion analysis: An overview, Computer Vision and Image Understanding, 108:4-18 (2007)
10. Rasmussen C.E., Williams C.K.I.: Gaussian Processes for Machine Learning, The MIT Press, Cambridge (2006)
11. Sigal L., Balan A.O., 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)
12. Sigal L., Bhatia S., Roth S., Black M.J., Isard M.: Tracking loose-limbed people, in CVPR '04 Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition (2004)
13. Taylor G.W., Sigal L., Fleet D.J., Hinton G.E.: Dynamical binary latent variable models for 3D human pose tracking, in CVPR '10 Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (2010)
14. Tian T., Li R., Sclaroff S., Tracking human body pose on a learned smooth space, Technical Report 2005-029, Boston University Computer Science Department (2005)
15. Yang Y., Ramanan D.: Articulated pose estimation with flexible mixtures-of-parts, in CVPR '11 Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (2011)
16. Wang J., Fleet D.J., Hertzmann A.: Gaussian process dynamical models for human motion, IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(2):283-298 (2008)