UNIVERSITAT POLITÈCNICA DE CATALUNYA

Feature-Based Annealing Particle Filter for Robust Motion Capture

by

Adolfo López Méndez

Image and Video Processing Group
Teoria de Senyal i Comunicacions

January 2009

Para Nuria,
a mis padres Adolfo y Lourdes
y a mi familia

"To be natural is such a very difficult pose to keep up"

Oscar Wilde

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Abstract

This thesis presents a new annealing method for particle filtering aiming at body pose estimation. Particle filters are Monte Carlo methods commonly employed in non-linear and non-Gaussian Bayesian problems, such as the estimation of human dynamics. However, they are inefficient in high-dimensional state spaces. Annealed particle filter copes with such spaces by introducing a layered stochastic search. Our algorithm aims at generalizing and enhancing the classical annealed particle filter. Different image features are exploited in a sequential importance sampling scheme to build better proposal distributions from likelihood. This technique, termed Feature-Based Annealing, is inferred from the required function properties in the annealing process and the properties of the weighting functions obtained with common image features in the field of body tracking. Comparative results between the proposed strategy and common annealed particle filter are shown to assess the robustness of the algorithm.

Acknowledgements

I would like to thank professor Josep R. Casas for his useful comments. I would like to thank also Serafeim Perdikis, Martin Lojka, Ananthakrishnan Gopal, Usman Saeed, Albert Ali Salah, Athanasios Vogiannou, Hamdi Dibeklioglu, Dimitrios Tzovaras and all who helped us to record the data that has been partially used in this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Inroduction

Automated inference of human pose from images is a challenging and often ill-posed problem. It basically involves the estimation of the configuration of a three-dimensional underlying structure of the human body, traditionally represented by an articulated model. This configuration is a high-dimensional hidden variable experimenting non-linear transformations as a result of the human dynamics and the mapping to images. Solutions to this challenge may offer many diverse applications. Not in vain, pose estimation and tracking has arisen as a highly active area within the fields of computer vision and image processing.

To overcome such a challenging problem, the search space is constrained by the prior knowledge about the human body. A virtual skeleton is often used to model the underlying structure of the human body and, indeed, the constraints on human pose space. Analysis-by-synthesis approaches have become very relevant due to the efficient use of this information. In such approaches, human body models are used to produce hypothesis that are matched with images to find pose correspondences. Within these approaches, stochastic sampling, and more concretely particle filters (PF) [1] [2] have become predominant methods due to their ability to precisely model non-linear and non-Gaussian processes. However, particle filters are inefficient for high-dimensional state spaces, thus requiring further improvements. To this end, partitioned sampling [3], hierarchichal sampling [4], covariance scaled sampling [5] or annealed particle filter (APF) [6] have been proposed. Among those proposals, APF has become on of the most relevant by treating the bayesian estimation of the human pose in an optimization context. In addition, hierarchical sampling or scaled covariance ideas can be applied to the APF scheme.

In this thesis, we investigate the role of image features in the annealing algorithm proposed for particle filters [6]. Besides, we exploit annealing as a method to generate

better proposal distributions in the context of Monte Carlo Importance Sampling [7] [8]. We address annealing to sample from likelihood instead of simply sampling from the transition prior. To do so, the properties of image features involved in the likelihood approximation are very important, since they provide different functions to sample from.

As a result of the research work on these areas, we propose a new annealing method in the context of image-based pose estimation and tracking named Feature-Based Annealing. Different image features are used to build weighting functions that are appropriately weighted through different annealing layers. We experimentally show the increased robustness of this technique under challenging conditions for pose estimation by comparing it with classical APF.

Chapter 2

Related Work

In this chapter, we review some fundamental concepts involving pose estimation from images. The chapter is broken down into two main blocks. The first one presents the modelling framework. Since we aim at model-based analysis-by-synthesis estimation, this framework is of high relevancy. The second block is devoted to particle filtering. First, the generic tracking problem is formulated as a non-linear non-Gaussian Bayesian state estimation. Then, particle filtering is introduced as a precise Bayesian estimator for the underlying statistics of non-linear and non-Gaussian dynamics.

2.1 Kinematic Chain Framework

The design of a model-based motion capture system implies implementing explicitly the prior knowledge about the body pose configurations. The body modelling framework must accomplish several requirements:

- **Provide a simple and compact representation of the possible body poses.** The relationship between body part locations and the body parameters must be simple and unique, while keeping a relatively low number of body parameters.
- **Capability of incorporating motion constraints**. Constraints can be set at the tracking level, but it is desirable that the body model incorporates the majority of them, thus leading to a more efficient tracking scheme.

The kinematic chain framework satisfies both requirements. Every 3D part location can be easily determined by the product of the twists affecting the motion of that point and most of the problem constraints are incorporated by means of hard kinematic restrictions.

In the following, we present twists and exponential map formula for kinematic chain framework and we introduce its application in an articulated human body model.

2.1.1 Twists and the Exponential Map Formulation

A kinematic chain [9] is an assembly of joints with several degrees of freedom (DOF), connecting rigid segments. Let SE(3) be the Euclidean group of rigid body motions and SO(3) the group of 3x3 proper rotation matrices. Both groups are Lie groups thus presenting an associated Lie algebra [10].

Let us consider the rotation of one segment with respect a single DOF of a joint to which the segment is connected. To this end, we model this DOF as a rotation axis and we consider two elements on it: a unit vector $\omega \in \Re^3$ and a point $\mathbf{q} \in \Re^3$. Assuming unitary velocity of rotation, the velocity $\dot{\mathbf{p}}$ of a point $\mathbf{p}$ on a rigid object about the rotation axis is determined by:

$$\dot{\mathbf{p}} = \omega \times (\mathbf{p} - \mathbf{q}) \tag{2.1}$$

If we re-write the above expression in homogeneous coordinates we obtain the SE(3) Lie algebra $\hat{\xi}$:

$$\begin{aligned} \begin{bmatrix} \dot{\mathbf{p}} \\ 0 \end{bmatrix} &= \begin{bmatrix} \hat{\omega} & -\omega \times \mathbf{q} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p} \\ 1 \end{bmatrix} = \hat{\xi} \begin{bmatrix} \mathbf{p} \\ 1 \end{bmatrix} \\ \underline{\dot{\mathbf{p}}} &= \hat{\xi}\underline{\mathbf{p}} \end{aligned} \tag{2.2}$$

where $\underline{\mathbf{p}} = \begin{bmatrix} \mathbf{p} & 1 \end{bmatrix}^T$ is the point $\mathbf{p}$ in homogeneous coordinates and $\hat{\omega}$ is the Lie algebra of SO(3):

$$\hat{\omega} = \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix} \tag{2.3}$$

Hence, Lie algebra of SE(3) is representing a twist. The solution for the differential equation in homogeneous coordinates 2.2 leads to the connection between the Lie algebra and the Lie group:

$$\underline{\mathbf{p}}(\theta) = e^{\hat{\xi}\theta}\underline{\mathbf{p}}(0) \tag{2.4}$$

Any element $\hat{\xi}$ of the Lie algebra can be mapped to its corresponding rigid body motion by means of an exponential. Hence, the above solution shows the transformation from an initial location of the point to the current location after rotating θ radians, by means of the exponential map associated to the twist $\hat{\xi}$.

For an open kinematic chain with n axes of rotation, this formulation provides an interesting property. Let $\underline{\theta} = [\theta_1 \dots \theta_n]$:

$$g_P(\underline{\theta}) = e^{\hat{\xi}_1\theta_1} .... e^{\hat{\xi}_n\theta_n} g_P(\mathbf{0}) \tag{2.5}$$

where $g_P(\theta)$ is the transformation from the rotations θ to the 3D locations of the chain points. This property allows us to compute the 3D location of every point in the chain by means of a product of the exponential maps associated to previous joints in the chain and a reference configuration $g_P(\mathbf{0})$. Moreover, this product is independent of the order in which it is computed.

An interesting property of Lie groups is that the exponential of its matricial elements is the matrix exponential. Therefore, we can compute the elements of an exponential mapping matrix by means of Taylor expansions, i.e, $\exp(\hat{\xi}) = \mathbf{I} + \hat{\xi} + \frac{\hat{\xi}^2}{2!} + \dots$. As a consequence we can develop the terms of the mapping as follows:

- Let $\|\hat{\omega}\| = 1$ where $\|\cdot\|$ is the Euclidean norm. Then, for any $\theta \in \Re$

$$\mathbf{R} = e^{\hat{\omega}\theta} = \mathbf{I} + \sin\theta\hat{\omega} + (1 - \cos\theta)\hat{\omega}^2 \tag{2.6}$$

- Let $\|\hat{\omega}\| = 1$ and $\mathbf{v} = -\omega \times \mathbf{q}$. Then, for any $\theta \in \Re$

$$\exp\left(\begin{bmatrix}\hat{\omega} & \mathbf{v}\\ 0 & 0\end{bmatrix}\theta\right) = \begin{bmatrix}\mathbf{R} & \mathbf{t}\\ 0 & 1\end{bmatrix} \tag{2.7}$$

  where

$$\mathbf{t} = \left(\theta\mathbf{I} + (1 - \cos\theta)\hat{\omega} + (\theta - \sin\theta)\hat{\omega}^2\right)\mathbf{v} \tag{2.8}$$

Note that if the rotation axes ω are aligned with the world coordinate system, the rotation matrices are very easy to find, because they will correspond to an x, y or z

rotation matrix. Since this choice verifies $\hat{\omega}^2 = \mathbf{I} - \omega\omega^T$ and $\hat{\omega}\hat{\omega}^2 = -\hat{\omega}$, we obtain the following result:

$$\mathbf{M} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} e^{\hat{\omega}\theta} & -(1-\cos\theta)\hat{\omega}^2\mathbf{q} - \sin\theta\hat{\omega}\mathbf{q} \\ 0 & 1 \end{bmatrix} \tag{2.9}$$

where $\mathbf{M} \in \mathrm{SE}(3)$ is used to denote the exponential map.

2.1.2 Articulated Body Model

The articulated body model consists in a set a of open kinematic chains, starting from the torso and ending in the terminal limbs. All these chains are built from a reference point, usually located in the torso, that will determine the position and orientation of the whole body. For given limb sizes, the positions of joints are denoted with respect to the body reference point.

Let us call the mapping involving the body translation and orientation $\mathbf{M_0}$ and $\mathcal{C}^p$ the set of joints belonging to a chain that affects a point in the body. As shown in 2.4, the motion of this point depends on the exponential maps of the preceding joints. Since the reference point is considered as the basic joint with an exponential map and it is always included in $\mathcal{C}^p$, this formulation allows to encode the pose in such a way that is invariant to rotation and translation.

$$\underline{\mathbf{p}}(\mathbf{x}) = \left(\prod_{i \in \mathcal{C}^p} \mathbf{M_i} \right) \underline{\mathbf{p}}(0) = \left(\prod_{i \in \mathcal{C}^p \setminus \mathcal{O}} \mathbf{M_i} \right) (\mathbf{M_0})\, \underline{\mathbf{p}}(0) \tag{2.10}$$

Equation 2.10 shows how the mapping of a given pose can be factorized in order to separate the mapping of the whole body translation and orientation. Hence, the vector parameter $\mathbf{x}$ can be split into pose-determinant angles and body translation and orientation.

2.2 Particle Filters

Pose estimation and tracking implies modeling the dynamics of the underlying structure of the human body. State-space models with non-linear and non-Gaussian transitions are a widely used approach. However, these models lead to analitically intractable statistics that require sub-optimal estimation algorithms. To this end, Monte Carlo methods

approximate probability density functions by means of a set of weighted samples. Particle Filters are a particular application of Monte Carlo methods over a sequence of noisy measurements. In the following, we will introduce the principles of importance sampling, sequential importance sampling (Particle Filters) and Sampling Importance Resampling Particle Filters.

2.2.1 Non-linear non-Gaussian Bayesian Tracking

The tracking problem [1] can be defined by means of a state-space model. Consider a sequence of states over time $\mathbf{x}_t$ constituting the dynamic process. From this sequence, one can recursively find the new state as follows:

$$\mathbf{x}_t = f_t(\mathbf{x}_{t-1}, \mathbf{v}_{t-1}) \tag{2.11}$$

where $\mathbf{v}_{t-1}$ is the process noise and $f_t : \Re^{n_x} \times \Re^{n_v} \rightarrow \Re^{n_x}$ is a possibly non-linear state transition function; n_x, n_v are the dimensions of the state vector and the noise of the dynamic process respectively. As a common Bayesian problem, $\mathbf{x}_t$ is a hidden variable producing observations or measurements:

$$\mathbf{z}_t = h_t(\mathbf{x}_t, \mathbf{n}_t) \tag{2.12}$$

where $\mathbf{n}_t$ is the observation noise and $h_t : \Re^{n_x} \times \Re^{n_n} \rightarrow \Re^{z_x}$ is a possibly non-linear state transition function; n_z, n_n are the dimensions of the observation vector and the noise of the observation process respectively.

These two equations constitute the Gauss-Markov state-space model that has been adopted to a great extent in tracking problems. The objective of such model is to recursively estimate the degree of belief that a state $\mathbf{x}_t$ is being produced given a collection of observations $\mathbf{z}_{1:t}$ up to time t, i.e, to infer the posterior pdf $p(\mathbf{x}_t|\mathbf{z}_{1:t})$. Note that an initialization $p(\mathbf{x}_0|\mathbf{z}_0) \equiv p(\mathbf{x}_0)$ must be available to recursively estimate the posterior. Then $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ may be obtained by prediction and update. Given the previous posterior $p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})$, prediction is performed by means of the Chapman-Kolmogorov equation:

$$p(\mathbf{x}_t|\mathbf{z}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1}) d\mathbf{x}_{t-1} \tag{2.13}$$

where $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is the state transition prior of a Markov process, which is a commonly adopted assumption for tracking. This probabilistic model is built according to the knowledge of the dynamical process given in equation 2.11.

The update step is the Bayesian computation of the posterior:

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) = \frac{p(\mathbf{z}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{z}_{1:t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})} \tag{2.14}$$

where the normalizing constant is obtained by the total probability theorem:

$$p(\mathbf{z}_t|\mathbf{z}_{1:t-1}) = \int p(\mathbf{z}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{z}_{1:t-1})d\mathbf{x}_t \tag{2.15}$$

Prediction and update are the formulation to find the optimal recursive Bayes solution for the tracking problem. However, the statistics of a non-linear non-Gaussian process are analitically intractable and therefore we cannot compute the integrals involved in the optimal solution.

2.2.2 Sequential Importance Sampling (SIS)

A way to deal with complex distributions is by drawing Monte Carlo samples of them [7]. Hence, a criterion describing the "goodness" of a sample set is needed:

Definition *A random variable x drawn from a distribution q is said to be* ***properly weighted*** *by a weighting function $w(x)$ with respect to the distribution p if for any integrable function h:*

$$E_q\{h(x)w(x)\} = E_p\{h(x)\}$$

A set of random draws and weights (x^i, w^i), $i = 1...N_s$ is ***properly weighted*** *with respect to p if*

$$\lim_{N_s\to\infty} \frac{\sum_{i=1}^{N_s} h(x^i)w^i}{\sum_{i=1}^{N_s} w^i} = E_p\{h(x)\}$$

for any integrable function h.

According to this definition, a pdf is approximated by discrete distribution of samples with probability proportional to the weights. Following this spirit, Particle Filters (PF) [1] are designed as recursive Bayesian estimators to approximate the posterior density $p(\mathbf{x}_t|\mathbf{z}_t)$ by means of a set of N_s weighted samples or particles. Given a Bayesian recursive estimation problem with Markov state transitions:

$$p(\mathbf{x}_{0:t}|\mathbf{z}_{1:t}) = \frac{p(\mathbf{z}_{1:t}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})} p(\mathbf{x}_{0:t-1}|\mathbf{z}_{1:t-1}) \tag{2.16}$$

we want to draw a set of properly weighted samples from the posterior. This set can be formulated as follows:

$$p(\mathbf{x}_{0:t}|\mathbf{z}_{1:t}) \approx \sum_{i}^{N_s} w_t^i \delta(\mathbf{x}_t - \mathbf{x}_t^i) \tag{2.17}$$

where w_t^i is the weight associated to the i-th particle. This discrete approximation of the posterior requires the evaluation of weights. This is done by means of the importance sampling principle [11], with a probability density function (pdf) $q(\mathbf{x}_{0:t}|\mathbf{z}_{1:t})$ from which we generate samples that can be evaluated with the posterior (up to proportionality). Applying the importance sampling principle in Eq. 2.16:

$$\begin{aligned} w_t^i &\propto \frac{p(\mathbf{x}_{0:t}^i|\mathbf{z}_{1:t})}{q(\mathbf{x}_{0:t}^i|\mathbf{z}_{1:t})} \\ w_t^i &\propto \frac{p(\mathbf{z}_{1:t}|\mathbf{x}_t^i)p(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i)}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})q(\mathbf{x}_{0:t}^i|\mathbf{z}_{1:t})} p(\mathbf{x}_{0:t-1}^i|\mathbf{z}_{1:t-1}) \end{aligned} \tag{2.18}$$

and choosing this importance distribution in a way that factors appropriately, we have:

$$\begin{aligned} w_t^i &\propto \frac{p(\mathbf{z}_{1:t}|\mathbf{x}_t^i)p(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i)p(\mathbf{x}_{0:t-1}^i|\mathbf{z}_{1:t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})q(\mathbf{x}_t^i|\mathbf{x}_{0:t-1}^i, \mathbf{z}_t)q(\mathbf{x}_{0:t-1}^i|\mathbf{z}_{1:t-1})} \\ w_t^i &\propto w_{t-1}^i \frac{p(\mathbf{z}_{1:t}|\mathbf{x}_t^i)p(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i)}{q(\mathbf{x}_t^i|\mathbf{x}_{0:t-1}^i, \mathbf{z}_t)p(\mathbf{z}_t|\mathbf{z}_{1:t-1})} \end{aligned} \tag{2.19}$$

Moreover, if we apply the Markov assumption, the expression is simplified regarding the fact that observations and current state only depend on the previous time instant. Therefore, the Particle Filter is a sequential propagation of the importance weights.

A major problem affects the PF. After several iterations the majority of the particles have negligible weights and, as a consequence, the estimation efficiency decays. An effective measure for the particle degeneracy is the survival rate [7] given by:

$$\alpha = \frac{1}{N_s \sum_{i=1}^{N_s} (w_t^i)^2} \tag{2.20}$$

In order to avoid this effect, there are two main strategies that can be combined. The first is the choice of the importance distribution. This is crucial since the samples drawn from $q()$ must be properly weighted with respect to the posterior. It has been shown in [11] that $q(\mathbf{x}_t|\mathbf{x}^i_{t-1}, \mathbf{z}_t) = p(\mathbf{x}_t|\mathbf{x}^i_{t-1}, \mathbf{z}_t)$ is optimal in terms of variance of the true weights, i.e., the weights obtained by sampling directly the posterior. However, this optimal importance distribution is not always a possible choice.

The second technique consists in resampling the particle set. After likelihood evaluation a new particle set must be drawn from the posterior estimation, hence particles with higher weights are reproduced with higher probability. Once the new set has been drawn all the weights are set to $\frac{1}{N_s}$, leading to a uniformly weighted sample set concentrated around the higher probability zones of the estimated posterior. Resampling is usually applied when the survival rate of the sample set is below a threshold.

The PF estimate should be computed before resampling, because resampling introduces additional random variation. As in Bayesian estimation, there are several options to produce an estimation by means of a significant point of the state-space. Posterior mean, maximum a posteriori or median are valid strategies. Since it is optimal in terms of mean squared error and provides a reduction of the noise introduced when sampling the distributions, the most common option is to use the Monte Carlo approximation of the posterior mean:

$$\hat{\mathbf{x}}_\mathbf{t} = \sum_{i=1}^{N_s} w^i_t \mathbf{x}^i_t \tag{2.21}$$

2.2.3 Sampling Importance Resampling (SIR)

The SIR Particle Filter proposed by Gordon et. al [12] is a method commonly used in computer vision problems. It is characterized by applying resampling at every iteration and by defining the importance distribution as the prior or prediction density $p(\mathbf{x}^i_t|\mathbf{x}^i_{t-1})$. By substituting this importance density in 2.19:

$$\begin{aligned} w^i_t &\propto w^i_{t-1} \frac{p(\mathbf{z}_{1:t}|\mathbf{x}^i_t)p(\mathbf{x}^i_t|\mathbf{x}^i_{t-1})}{p(\mathbf{x}^i_t|\mathbf{x}^i_{0:t-1})p(\mathbf{z}_t|\mathbf{z}^i_{1:t-1})} \\ w^i_t &\propto p(\mathbf{z}_{1:t}|\mathbf{x}^i_t) \end{aligned} \tag{2.22}$$

Hence, the computation of weights only depends on the likelihood. Consequently, the design of the particle filter is basically a problem of finding an appropiate likelihood

function. A particular advantage of the SIR PF is that we can easily sample $\mathbf{x}_t^i \sim p(\mathbf{x}_t|\mathbf{x}_{t-1}^i)$ by first generating a noise sample $\mathbf{v}_t^i$ (according to the noise distribution of our system) and then setting $\mathbf{x}_t^i = \mathbf{f}_t(\mathbf{x}_{t-1}^i, \mathbf{v}_t^i)$. The major drawback is that we generate samples without any insight about the observations, thus making the algorithm sensitive to outliers and possibly inefficient.

Chapter 3

State of the Art

Up to now, we have presented a background framework to focus our research in articulated model-based particle filter-based body trackers. Consequently, the goal of the following state of the art review is to show existing techniques that can be circumscribed within these topics.

3.1 Facing High Dimensional Limitations of the SIR PF

The large number of degrees of freedom that can be found in an articulated body model makes the body tracking problem a high-dimensional state-space problem. SIR Particle Filters are a good approach for tracking in low dimensional spaces, but they become inefficient in high-dimensional problems, because they require a number of particles that grows exponentially with the number of dimensions.

In high-dimensional bayesian problems, likelihood functions are often multi-modal and sharpened. Let us call principal mode to the mode that will be defined around the global maxima, and secondary modes to the rest (which can be linked to the local maxima). If we see the likelihood as a mixture of unimodal pdfs, then a valid estimation would be produced by selecting the unimodal pdf that best matches the principal mode (although this may yield to a degenerate posterior).

The prior density term of the Bayesian estimation is a function that is, in general, broader than every mode of the likelihood. Hence, the samples drawn from the prior will not constitute a properly weighted set, i. e., the variance of the weights will be high because only a few particles will hit the principal mode of the likelihood. Furthermore, the likelihood function may present secondary modes in the typical set of a common prior density (typically a Gaussian density). In this case, the importance weights are

unlikely to point at the global maximum of the likelihood, in which close vicinity we expect to find a good estimation of the pose. Hence, the Monte-Carlo approximation of the posterior mean estimate is likely to be highly biased (see Fig. 3.1).

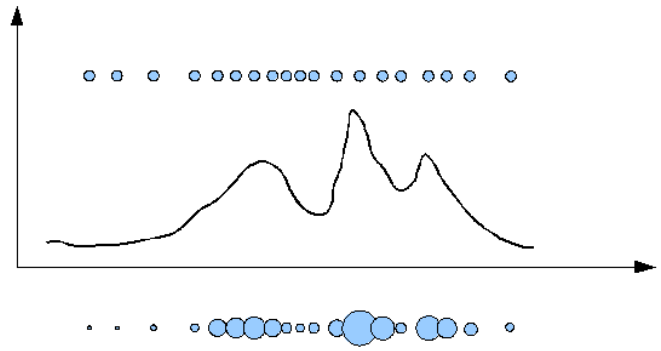


FIGURE 3.1: Particle Filter and Multi-Modal Likelihood. Since the prior is broader than the modes of the likelihood, the Monte-Carlo approximation of the posterior mean estimate will be highly biased. In addition, the sample set presents a high variance in its weights

3.1.1 Partitioned Sampling

Partitioned Sampling [3] is one of the first succesful strategies to cope with the high dimensional limitations of the SIR PF. This technique basically aims at splitting the vector parameter to sample the state-space efficiently. It is based on the assumption that $p(\mathbf{x}_t|\mathbf{x}_{t-1}) = p(x_{t1}, ..., x_{tK}|x_{(t-1)1}, ..., x_{(t-1)K})$ where $\mathbf{x}_t \in \Re^K$ can be factorized in a way that samples are drawn as follows:

- $x^i_{t1} \sim p(x_{t1}|x_{(t-1)1}, x_{(t-1)1}, ..., x_{(t-1)K})$.
- $x^i_{t2} \sim p(x_{t2}|x_{(t-1)2}, x_{(t)1}, ..., x_{(t-1)K})$.
- ...
- $x^i_{tk} \sim p(x_{tk}|x_{t1}, ..., x_{t(k-1)}, x_{(t-1)(k+1)}, ..., x_{(t-1)K})$.
- ...
- $x^i_{tK} \sim p(x_{tK}|x_{t1}, ..., x_{t(K-1)}, x_{(t-1)K})$.

Note that the Markov assumption has been taken into account.

Under such factorization, the state vector can arbitrarily be split if we have statistical models of the isolated dynamics of each one of its elements. Furthermore, we can apply the SIR to every dimension whenever appropiate weighting functions can be found. As a consequence, the required number of particles is reduced because of the linear dependency between the dimensions of the space and the particle cardinality. Unfortunately, finding the required weighting functions for the separate dimensions is not an easy task and often requires high-level features.

3.1.2 Hierarchical Sampling

An articulated body model yields to a state vector comprising a set of angles as pose variables. It seems evident that partitioning can be applied to this set of angles. However, partitioned sampling does not propose a partitioning related to the body model structure. In [4], the underlying hierarchy of the human body is used to define state vector partitions (see Fig. 3.2).

This algorithm is a stochastic version of that of [13], in which the state-space is decomposed to increase the efficiency of the local search of different limbs.

FIGURE 3.2: Examples of the Mitchelson's hierarchical sampling for 3D body tracking. First, torso is sampled independently and weighted to find an estimate for the whole body location and orientation. Then, arms and legs are sampled to refine the pose estimate.

However, as discussed in [6], self-occlusions and self-overlaps found in multiple 2D views make independent limb location very difficult, unless labelling cues or very reliable local color information is available.

3.1.3 Annealing Particle Filter

Deutscher et. al [6] proposed a variation of the SIR framework by introducing the variant of the simulated annealing [14] in the Particle Filtering concept. Annealing PF deals with multiple peaked maxima functions by evaluating the particles in several smoothed versions of the likelihood approximation (see Fig. 3.3). After the weights are computed via these smoothed versions of the likelihood approximation, particles are resampled and propagated with Gaussian noise with zero mean and a covariance that decreases at every step. By doing so, particles are likely to be drawn in the vicinity of the likelihood's global maxima.

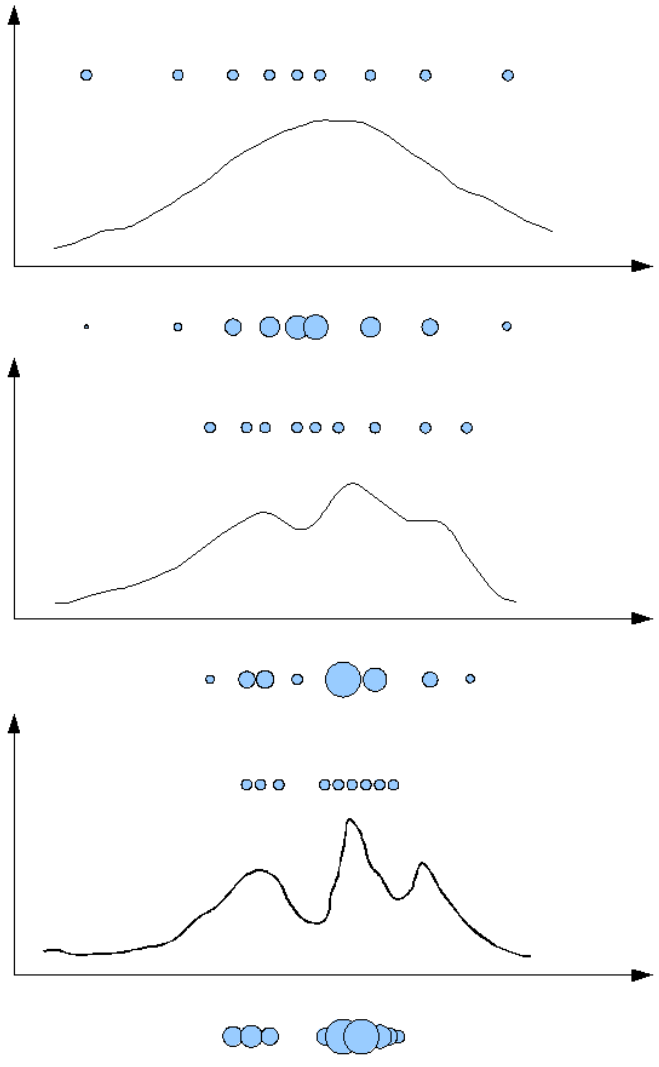


FIGURE 3.3: Annealing Particle Filter with 3 layers. The bias of the Monte-Carlo estimation and the variance of the weights are reduced. The probability of hitting the typical set of the principal mode is increased with every layer, thus leading to a more properly weighted set with respect to the posterior

The steps comprising weighting with a smoothed function, resampling and propagation are called annealing runs. In the last annealing run, the estimation is given by means of the Monte Carlo approximation of the posterior mean.

The most usual way to smooth the weighting function is by means of an annealing rate, an exponent $\beta < 1$. In the first layer, β is minimum and progressively increases with each layer, sharpening the likelihood approximation. In [6] a method for tuning β with the survival rate after each annealing run is proposed.

The sharpness of the likelihood function is due to the high dimensional space in which it is defined. Using well-defined hierarchical models [15] is another possible strategy in order to have annealing layers. The ordered exploration of spaces of increasing dimensionality helps in avoiding the sampling procedure to get misdirected by local maxima. From this particular definition of annealing, the APF can be seen as a generalization of the aforementioned partitioned schemes.

Since Annealed Particle Filter addresses the dimensionality limitations of the SIR by means of a layered stochastic search and avoids strong assumptions on motion and data availability, it is more suitable than other techniques [16] such as Local Search in Decomposed State [13], Maximum Likelihood-based trackers [17], Relevance Vector Regression [18] or other Bayesian approaches [19]. Hence, APF is the basis of most of the best-performing body trackers found in the literature and the starting point of our work.

FIGURE 3.4: Example of Deutscher's Annealing Particle Filter

3.2 Likelihood Evaluation

As stated in section 2.2.3, a crucial design step for the SIR-based trackers is the evaluation of a likelihood function. In computer vision problems, probability density functions

usually are not directly accessible, thus an observation model is required to approximate the likelihood function. It is necessary to determine which image features are more correlated with the true body configuration. Therefore, finding the appropiate likelihood approximation involves both image and body model. A skeletal body model requires a flesh model in order to have an entity that can be compared with image features.

In general, the best way to combine several feature weightings is by means of a product of Gibbs-like functions:

$$\omega = exp\left(-\sum_{c=1}^{C}\sum_{n=1}^{N}\omega_n\right) \tag{3.1}$$

where c corresponds to the view index, n to feature index and ω denotes the weighting function for the n-th feature.

In [13], the flesh model consists in a set of tappered superquadrics that are projected onto several images and matched with foreground segmentation and Chamfer distances to the extracted edges. Similarly, Deutscher et al. [6] matched a flesh model with foreground and edges, but they avoided the computation of the Chamfer distance using a smoothing of the detected edges. They flesh out the articulated model by means of conic sections with elliptical cross-sections surronding virtual skeleton segments (see Fig. 3.4). Raskin et al. [20] use a similar model and they add the body part histogram as an additional feature. They also estimate the visibility of each limb in order to weight every view.

Other authors use Visual Hull approaches [21] to work with voxel data. In that case, a common flesh model is a set of ellipsoids surrounding the skeletal segments [19]. [22] presented these ellipsoids as three-dimensional Gaussian mixtures. In this approach, a cross-entropy measure between target and data mixtures is used for the likelihood approximation.

3.3 Motion Priors

Articulated models incorporate hard kinematic constraints, but further restrictions can be considered in order to reduce the state space to a more tractable subspace. A usual way to achieve such reduction is by means of specific motion priors [23], thus exploiting the bayesian prediction component. Instead of sampling from a given distribution, this approach samples from a motion history database. To do so, we replace the prior pdf $p(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i)$ by $p(\mathbf{x}_t^i|\hat{\mathbf{x}}_{t-1})$ where $\hat{\mathbf{x}}_{t-1}$ is the estimated pose in the previous time instant. Let $\mathbf{s}_{t-1}^i$ be a pose sample in the stored motion sequence $\mathcal{S} = (\mathbf{s}_0^i \dots \mathbf{s}_T^i)$ in the database. Hence, the formulation of the learnt motion prior is as follows:

$$p(\mathbf{x}_t^i|\hat{\mathbf{x}}_{t-1}) = p(\mathbf{x}_t^i|\mathbf{s}_{t-1}^i)p(\mathbf{s}_{t-1}^i|\hat{\mathbf{x}}_{t-1}) \tag{3.2}$$

Therefore, sampling in this new prediction scheme implies matching the previous estimation with the stored poses and setting $\mathbf{x}_t^i = \mathbf{s}_t^i$. Note that, in the SIR framework, the prior pdf is also the importance distribution.

Since motion transition from one pose to another is assumed to be a Markov process, a classical approach is to learn the motion priors as Markov sequences by means of HMM [24]. Caillette et al. [22] use training data to automatically divide complex motions into elementary dynamics. They basically cluster the feature space into Gaussian clusters and use this probabilistic model to build a Variable Length Markov Model (VLMM) [25].

Gaussian Process Latent Variable Models(GPLVM) [26] and Gaussian Process Dynamical Models(GPDM) [27] are non-linear mappings to a latent space that provides a compact and efficient representation of data. The latter have been succesfully applied in articulated model-based body tracking by several authors. Raskin et al. [20] introduce the Gaussian Process Dynamical Models in the APF scheme to improve tracking, specially in low frame rate sequences. [28] combines this non-linear embedding with the VLMM to improve the results obtained in [22].

In spite of providing chances for a tracking improvement, these approaches impose a reduction of the solution generality. They imply a training procedure that yields to the problem of constructing a motion prior database with the minimum loss of generality or, at least, the exhaustivity required to avoid terminal failures caused by motions outside the database.

3.4 Summary

In the preceding sections of this chapter, we have presented a state of the art review focused on model-based multiview 3D pose estimation and tracking.

We have made emphasis on stochastic sampling techniques because their ability to cope with non-linear and non-Gaussian processes, such as human dynamics. Moreover, literature shows that, when a single point is provided as an estimation (commonly the Monte Carlo posterior mean estimate), APF strategies outperform other approaches by dealing with the Bayesian estimation problem in an optimization context. Hence, this algorithm will be the basis of our work.

Regarding the likelihood evaluation, it is not necessary that the flesh models are highly realistic. However, a necessary condition is that the flesh model should provide a support that can be easily matched with image features. Hence, volume primitive, and more concretely conical sections, are enough to design a good body tracking.

Finally, our goal is to enhance the tracking without loss of generality. To this end, state transition priors should be as general as possible. This basically implies simple and uninformative pdfs such as Gaussians. In any case, in our solution we will try to keep an open door for motion models to enhance priors for specific scenarios, where high accuracy and/or efficiency is required.

Chapter 4

Study on Image Features for Likelihood Approximation

In our approach we prefer not to rely on a 3D reconstruction that could be difficult to build and, indeed inaccurate. Therefore, we opt for a projection of the flesh model onto the images. Our proposal is to avoid the computational cost of projecting the whole set of sampling points of a 3D flesh model by projecting a reduced set of points per body part. The flesh model will be a set of cylinders around all the skeleton segments except the head, which will be modeled by a sphere (see Fig. 6.2(a)).

In the following, we show how common image features are processed and used to define weighting functions in order to approximate the likelihood. In addition, we make a preliminary analysis of the properties of these functions. This study is the basis for the major contribution of the present Ms Thesis.

4.1 Model Projection

Likelihood evaluation is the bottleneck of SIR Particle Filters and derived algorithms. Computational capacity is usually spent in feature extraction, flesh model building/projection and particle evaluation. While feature extraction is bounded in terms of computational cost (and it is herein commented because when dealing with voxel data the projection is somehow included in the feature extraction), flesh model building/projection and particle evaluation clearly depend on the number of particles. Even though existing rendering software provide efficient tools to project volume primitives, we make our own flesh model projection by building the flesh templates directly on the images.

We proceed in such manner in order to reduce the number of projected points, thus reducing the computational cost of the projection step.

Our reduced set of projected points will be defined by the vertices of the trapezoidal section resulting from the intersection of a plane, approximately parallel to the image plane, with the cylindric shape modelling the limb (or spherical shape in the case of the head).

To define an intersecting plane for a given cylinder, we compute the vectors going from the camera center towards each one of the ending points of the limb. Then the cross product of these vectors with the one defined by the principal axis of the limb itself is computed to determine two normal vectors that lie on the intersecting plane and along which we will find the key points to project (see Fig. 4.1). The head template is handled with a similar procedure using as limb vector the one going from the base of the neck to the head center.

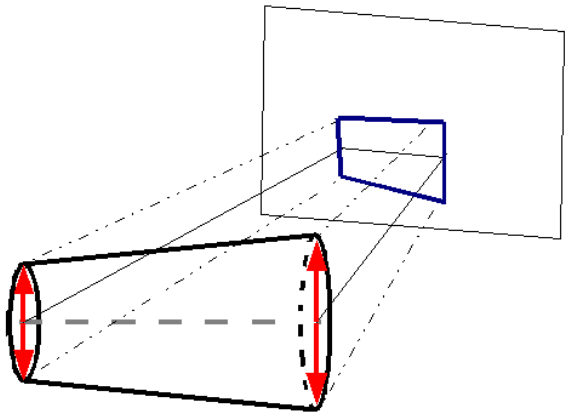


FIGURE 4.1: Illustration of the model projection procedure for a given cylinder. The normal vectors are depicted in red and the final model projection is shown in dark blue.

The norm of the cross product, as well as the area of the projected trapezoid, can be used as a quality measure in order to determine whether the limb is properly aligned with the view (this does not apply for the head). If this quality measure is above a certain threshold, we can change the trapezoidal projected shape by a circle or an ellipse or simply correct the projected shape. For instance, in scenarios where the subject is expected to notably change his or her orientation, we use a box as torso volumetric primitive. However, we project the vertices of a rectangle defined by the shoulder positions and the torso dimensions. The approximated area of the projected torso template is used to determine wether the projection is suitable for that view or

not. If it is found that is under a certain threshold, the projected template is replaced by a rectangle with the minimum required area to match with image features.

This procedure saves the computational effort of projecting the whole super-quadrics or a set of sufficient sampling points. However, the projection model is inaccurate in comparison to an exhaustive projection method. Therefore, the pose estimator should also face the model inaccuracies.

4.2 Feature Extraction

Regarding the image features, we propose modifications on a likelihood approximation like the one proposed in [6] while keeping common features that are easy to extract, like foreground silhouettes, edges and detected skin (see Fig. 4.2).

- We extract foreground silhouettes by means of a background learning technique based on Stauffer and Grimson's method [29]. A single multivariate Gaussian $\mathcal{N}(\mu_t, \mathbf{\Sigma}_t)$ with diagonal covariance in the RGB space is used to model every pixel value $\mathbf{I_t}$. For simplicty, $\mathbf{\Sigma}_t = \sigma^2\mathbf{I}$. The algorithm learns the background model for every pixel using a set of background images and then, for the rest of the sequence, evaluates the likelihood of a pixel color value to belong to the background. With every pixel that matches the background the pixel model is updated, adaptively learning smooth illumination changes:

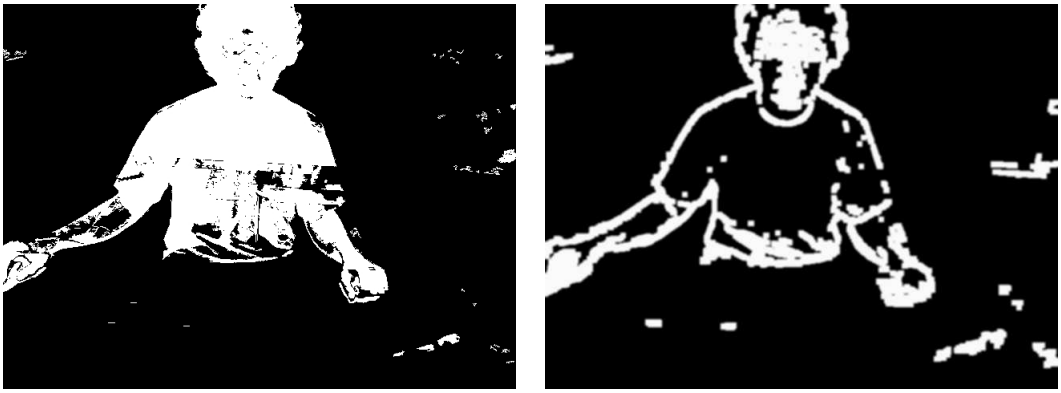
$$\mu_t = (1-\rho)\mu_{t-1} + \rho\mathbf{I_t} \tag{4.1}$$

$$\sigma_t^2 = (1-\rho)\sigma_{t-1}^2 + \rho(\mathbf{I_t} - \mu_{t-1})^T(\mathbf{I_t} - \mu_{t-1}) \tag{4.2}$$

  A shadow removal algorithm [30], based on the color and brightness distortion, is used to enhance the segmentation.

- Edge detection is performed in the luminance channel by means of the Canny edge detector [31]. Depending on the proximity of the views, the result is dilated with a square structuring element. This is done in order to increase the probability of hitting the edge. Finally, the edge map is smoothed with a Gaussian mask. In order to avoid background spurious edges, we previously mask the edge detection provided by Canny's algorithm with a dilation of the foreground mask.

- A simple skin detection method based on evaluating the likelihood ratio between skin and non-skin hypothesis is performed. The likelihood functions are estimated by 8-bin RGB color histograms of several skin and non-skin samples.



(a) Foreground Mask (b) Edges Mask

FIGURE 4.2: Extracted Image features

4.3 Feature Weightings

The final likelihood approximation will be a combination of several measures constructed with the aforementioned features. The following weightings are the exponents of the final likelihood approximation and regarding the way they are presented, they must be understood as penalties.

N sampling points of the projected flesh model are matched with the extracted foreground corresponding to a view. The weight is computed as follows:

$$\omega^{fgl} = \frac{1}{N}\sum_{n=1}^{N}(1 - I_n^{fg}) \tag{4.3}$$

Since pixel intensities in the foreground masks (I_t^{fg}) have 0 or 1 as possible values, the weighting function is obtained by a normalized sum of the background pixels falling inside the projected flesh model. In the case of the head, we add skin detection information:

$$\omega^{fgh} = \frac{1}{N}\sum_{n=1}^{N}(1 - I_n^{fg} I_n^{s}) \tag{4.4}$$

where I_n^s is the pixel intensity in the skin map. Therefore, the final foreground weight ω^{fg} is the averaged sum of all the limbs ω^{fgl} and head weights ω^{fgh}.

The proposed weighting function for edges is a sum of squared differences between the contour pixels I_n^e and the edges of the flesh model aligned with the axis of the limb:

$$\omega^e = \frac{1}{N}\sum_{n=1}^{N}(1 - I_n^e)^2 \tag{4.5}$$

where N stands for the sampling points along the occluding edges of the projected model.

4.3.1 Foreground Divergence Measure

The proposed foreground matching measure shows how well the model fits the observation, but does not evaluate how well the observations are being explained by the model. Suppose the likelihood $p(\mathbf{z}_t|\mathbf{x}_t)$ is available and that a given pose generates a pdf. A measure that can be used to assess the similarity of the likelihood and the generated pdf is the Kullback-Leibler divergence [32]. At this point, it is important to remark that the KL divergence will provide different results depending on the factor order (except if both pdfs are identical).

We can establish an analogy with our likelihood approximation. Suppose that the foreground mask is a discrete two-dimensional function. If we normalize all the foreground values by the total number of foreground pixels we construct a function with uniform pdf appearance. Now, we project the whole flesh model. With the appropiate normalization, this projection defines another uniform pdf-like function in a two-dimensional discrete domain. Let us consider that both the foreground silhouette and the projected model have a similar number of pixels N_{fg}. Let us also make an approximation of the KL divergence between the projected flesh model and the foreground silhouette:

$$D(model||silhouette) = \sum_{i=1}^{N} model(i) \log\left(\frac{model(i)}{silhouette(i)}\right)$$

Consider the following limit cases:

- Outside the model projection ($model(i) = 0$) the divergence is not considered and therefore its value is 0. This stands for all the possible values of the foreground mask $\{0, \frac{1}{N_{fg}}\}$, since the limit of a function of the class $x \log x$ is 0 when x tends to zero (when both values are 0 the convention $0 \log \frac{0}{0} = 0$ is followed).
- If $model(i) = \frac{1}{N_{fg}} = silhouette(i)$ then the divergence is zero.
- If $model(i) = \frac{1}{N_{fg}}$ and $silhouette(i) = 0$ then the divergence tends to infinity.

Note that, if we force the divergence to be 1 in the last case and the model and silhouette to be binary, the meaning of this approximation of the KL divergence is the same as the

foreground matching measure in equation 4.3. The question is why the approximation of the KL divergence is not computed the other way around, $D(silhouette||model)$, as it is computed in information theory problems, where the first pdf represents the observations. Hence, we propose to include an additional divergence measure between the projection of the flesh model and the foreground masks to see how well a particle explains the observations.

$$\omega^d = \frac{1}{N_{fg}} \sum_{n=1}^{N_{fg}} (I_n^{fg}(1 - B_n)) \tag{4.6}$$

This divergence basically consists in measuring the occupancy of the foreground silhouette (comprising N_{fg} foreground pixels) by the B_n pixels of the projection of a given particle. Note that with binary silhouettes we only take into account regions over the foreground pixels where there is no projection of the model.

We avoid this computation with edge information because this feature is more sensitive to spurious data.

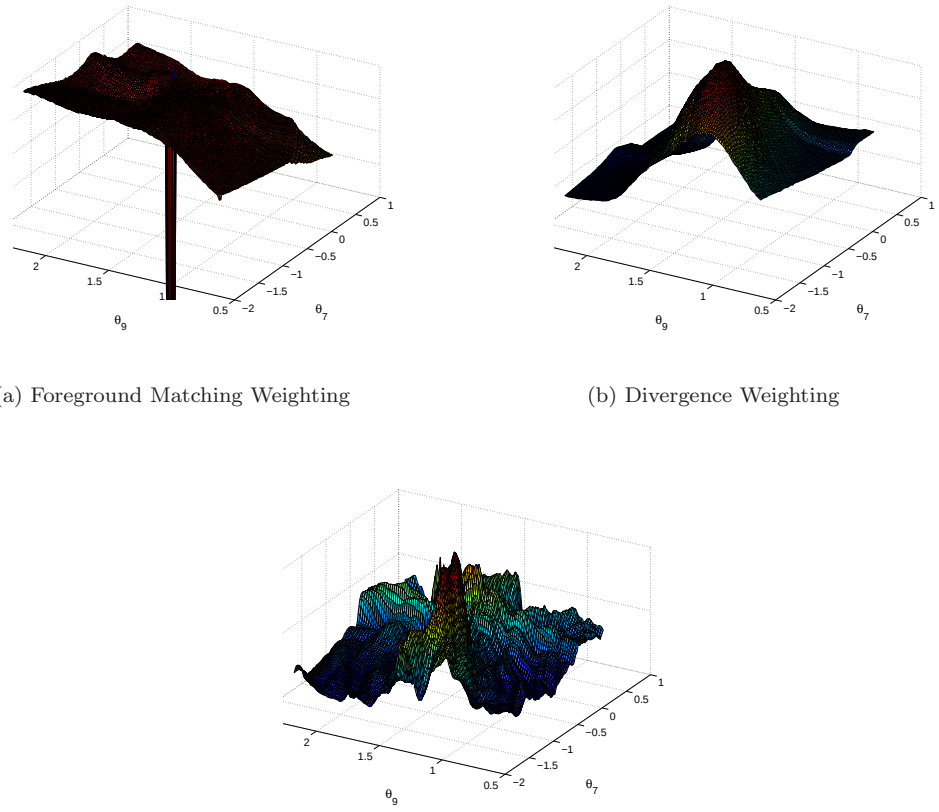
4.4 Weighting Functions Properties

In order to properly combine the weighting functions in the final likelihood approximation, one would like to have insights about their properties. A priori, we know that edge information tends to produce more sharpened and multi-modal likelihood approximations. Both foreground matching and foreground divergence measures may be smoother since they are defined over data that is smoother than edge information. In spite of that, both measures are, in general, multi-modal in the whole pose-space, mainly due to ambiguities of the projected poses.

In order to check the assumptions made on the different functions, we have evaluated several likelihood approximations based on one of the presented image features (taking into account several views). We uniformly sample over two parameters of a given pose space in order to visualize the resulting functions.

The foreground matching measure produces a smooth and flat function (in almost every point) in which many different poses take considerable degrees of likelihood. However, foreground matching has the property of being discriminative with several wrong states. These properties can be observed in Fig. 4.3(a), where the weighting function is shown with actual data as a function of two pose angles. The foreground divergence measure is a smooth function that presents, in general, a broad global maximum (see Fig. 4.3(b)).

Edge matching is the most determinant measure in the sense that high values can only be reached when a particle is very close to the true pose. Nevertheless, spurious edges can also produce high values of the likelihood approximation (see Fig. 4.3(c)).



(a) Foreground Matching Weighting

(b) Divergence Weighting

(c) Edges Matching Weighting

FIGURE 4.3: Plots of the different likelihood approximations resulting from separate image feature weightings, represented as functions of two angles of the left arm. The rest of parameters are set to values close to the true pose. Two views have been considered for all the likelihood approximations depicted.

Chapter 5

Feature-Based Annealing

Foreground measures produce very broad and generally flat functions. On the other hand, matching with edges tends to produce peaked functions with several sharpened local maxima. The combination of all these measures is, in general, a peaked function with several local maxima.

When dealing with such likelihood approximation, we want our sampling scheme to converge to the global maxima, which we assume to be very close to the posterior mean. In the following, we are going to gather some definitions and assumptions taken in similar approaches in statistical or pose estimation problems and in section 5.2 we are going to introduce the main contribution of this Ms thesis.

5.1 Pose estimation and tracking problem

In chapter 2 we have presented the tracking problem as a recursive estimation problem. In this context, we take advantage of prior knowledge provided by the body model and some physical constraints to use a Bayesian framework. The use of Monte Carlo methods is therefore justified because of the non-linear and non-Gaussian underlying statistics of the human dynamics.

Pose estimation, as well as other estimation problems, produces a single point output rather than a whole posterior pdf. Obviously, this point is taken from the posterior estimation either as its mean, maximum or median. The true pose is a single point in the pose space emitted at a precise time instant. We can make an analogy with frequency estimation problems by formulating the pose as $\delta(\mathbf{x}_t - \hat{\theta}_t)$. Then, the pose is a hidden input of a non-linear system whose outputs are our observations, just as in a Gauss-Markov State-Space model (see Fig. 5.1). This model is implicitly assumed in

pose estimation problems and is used here to illustrate the further assumptions about the underlying statistics of the elements in the model.

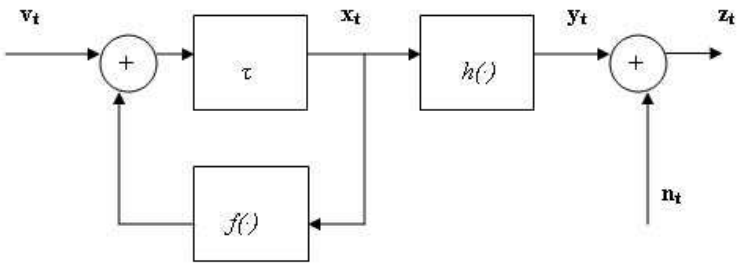


FIGURE 5.1: Example of a Gauss-Markov State-Space Model. In pose estimation problems the true state can be seen as a tone in the pose space $\delta(\mathbf{x}_t - \hat{\theta}_t)$ and the observations can be seen as the channel response at this frequency.

Although we cannot state many things about the non-linear dynamics of the body model, it seems logical to assume that, for a given time t, the uncertainty associated to the physical phenomena of "producing" a pose is low, at least much lower than the observation noise. If we observe a moving person, most of the uncertainty to determine the pose comes from the observation (occlusions, clothing, limitations of human sight when motion is fast, etc.). Therefore, we expect that the underlying pdf of such process is somehow a narrow function around the true pose $\hat{\theta}_t$.

Regarding the observations, projection of poses onto images is a highly noisy and non-linear mapping that, altogether with the high dimensionality of the state-space, make the likelihood functions multi-modal and peaked. Similar conclusions are drawn in [33], where the multi-modality of the problem comes mainly from the implicit many-to-one transformation from the state to observations. To estimate the true pose, annealing-based approaches try to concentrate the particle population around the global maxima of the likelihood. Some authors [34] consider this concentration as a drawback of annealing methods, since they imply loss of information. Particles are not representing the whole posterior thus becoming a degenerate pdf estimation. However, experimental results show that annealing outperforms other sequential importance sampling approaches when the estimation is given in terms of expectation of the posterior, and no learnt motion prior is used [34]. Furthermore, when learnt motion priors are used, transition prior becomes narrower than a simple and often uninformative gaussian, thus the posterior estimate is likely to be "less multimodal". Based on these empirical results, an optimization context seems suitable for the pose estimation problem and, as a consequence, the following condition is assumed:

- Likelihood's global maxima is close to the true pose

$$\exists \epsilon > 0 \not| \left| \arg\max_{\mathbf{x_t}} p(\mathbf{z}_t|\mathbf{x}_t) - \hat{\theta}_t \right| < \epsilon, \forall t, \mathbf{x} \in \mathcal{X}$$

In fact, since likelihood is not available for the image problem, this assumption is applied to the final approximation. This condition must be understood in a Sampling Resampling Importance context, i.e., $\mathbf{x}$ is confined to the discrete support of the samples belonging to the prediction based on the previous posterior estimate.

We are assuming that the likelihood function has a principal mode which is a function that, together with the adequate prior $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, produces a properly weighted set with respect to the posterior $p(\mathbf{x}_t|\mathbf{z}_t)$. Unfortunately, common prior pdf's used in body tracking problems are broad, thus yielding a multimodal posterior estimate and a subsequently biased posterior mean estimate. Therefore, an algorithm dealing with such functions should avoid the secondary modes by means of an improved importance sampling strategy.

We have already seen that annealing is a technique aiming at the convergence to this principal mode or global maxima. Annealing can be seen also as a technique to obtain proper weighting functions with a simple importance distribution. This viewpoint is very similar to the Annealed Importance Sampling [8]. In the first layer annealing provides a broad function centered or close to the principal mode which is used to generate a new set of samples. With every layer, the new generated set is expected to be closer to the proper weighting with respect to the principal mode of the likelihood function and, consequently, to the posterior.

A more formal probabilistic approach for this interpretation can be derived extending the sequential importance sampling concept to the annealing layers [8]. By definition, every layer can output an estimation of the posterior pdf and, as a consequence, it makes sense to use an importance distribution and a prediction pdf per layer. If we call $l = 0, ..., L-1$ the layer index (where 0 is the smoothest layer and $L-1$ the more sharpened) at time t and $p(\mathbf{z}_{t,l}^{\beta(l)}|\mathbf{x}_{t,l}^i)$ to the corresponding smoothed version of the likelihood function:

$$w_{t,L-1}^i \propto \prod_{l=1}^{L-1} \left(\frac{p(\mathbf{z}_{t,l}^{\beta(l)}|\mathbf{x}_{t,l}^i) p(\mathbf{x}_{t,l}^i|\mathbf{x}_{t,l-1}^i)}{q(\mathbf{x}_{t,l}^i|\mathbf{z}_{t,l}^{\beta(l)})} \right) \frac{p(\mathbf{z}_{t,0}^{\beta(0)}|\mathbf{x}_{t,0}^i) p(\mathbf{x}_{t,0}^i|\mathbf{x}_{t-1,L-1}^i)}{q(\mathbf{x}_{t,0}^i|\mathbf{x}_{t-1,L-1}^i, \mathbf{z}_{t-1,L-1})} w_{t-1,L-1}^i \quad (5.1)$$

Note that if L=1, the above expressions are equivalent to the basic particle filtering scheme (if $q(\mathbf{x}_{t,0}^i|\mathbf{x}_{t-1,L-1}^i, \mathbf{z}_{t-1,L-1}) = p(\mathbf{x}_{t,0}^i|\mathbf{x}_{t-1,L-1}^i)$ and $w_{t-1,L-1}^i = \frac{1}{N_s}$ then SIR PF is applied). Setting $q(\mathbf{x}_{t,l}^i|\mathbf{z}_{t,l}^{\beta(l)}) = p(\mathbf{z}_{t,l-1}^{\beta(l-1)}|\mathbf{x}_{t,l-1}^i) p(\mathbf{x}_{t,l}^i|\mathbf{x}_{t,l-1}^i)$ and $p(\mathbf{z}_{t,l}^{\beta(l)}|\mathbf{x}_{t,l}^i) =$

$p(\mathbf{z}_t|\mathbf{x}_{t,l}^i)^{\beta(l)}$is equivalent to the APF propagation strategy between layers. Moreover, it yields to the SIR concept for annealing layers, where the importance weights of every layer only depend on the smoothed likelihood. Hence, one can see annealing as a layered likelihood proposal distribution, since it aims at finding better estimates just sampling from the adequate versions of the likelihood. In the classical annealing, however, adequate means smoothed by an exponent.

An alternative idea to exploit annealing to sample from likelihood is to find, at least, two groups of image features: one that produces coarse but locally almost convex functions around the global maxima and a second group providing highly peaked and determinant likelihood approximations, despite its multi-modality. The word 'locally' must be understood as within the support of the prediction based on the previous estimate. From our viewpoint, it may be very difficult to construct the coarse weighting from a simple smoothing of the peaked and multi-modal weighting, and that is what common annealing strategies aim at. Therefore, we need to find several sets of weighting functions or combinations of weighting functions that fulfill the described requirements.

5.2 Annealing enhacements towards tracking robustness

A particular characteristic of foreground and edge information is that they complement themselves for body tracking problems. Both the literature and our experiments with the image features evidence that edge information is very determinant. At the same time, this information can easily misdirect the estimation due to the presence of spurious edges, like, for instance, wrinkles caused by lose clothing. On the other hand, mainly due to self-overlaps and self-occlusions, foreground information does not produce determinant information to estimate the pose, but it shows robustness against background and clothing. This is a high-level interpretation of the shape of the weighting functions produced by these features and also a justification for being widely used in the literature. Besides, the properties of the weighting functions resulting from these features approximately match the requirements of the two groups described in the previous section.

In the context of human body tracking, we have mentioned two procedures to anneal the likelihood approximation: by means of an exponent $\beta < 1$ or by exploiting the underlying hierarchical structure of the human body. Nevertheless, neither exponent-based annealing nor hierarchical body structure seem to be conceived focusing in the problem of highly noisy and, at the same time, determinant features.

In image processing problems, likelihood approximations are often defined with several image features extracted from the same source, the raw observation (see Fig. 5.2).

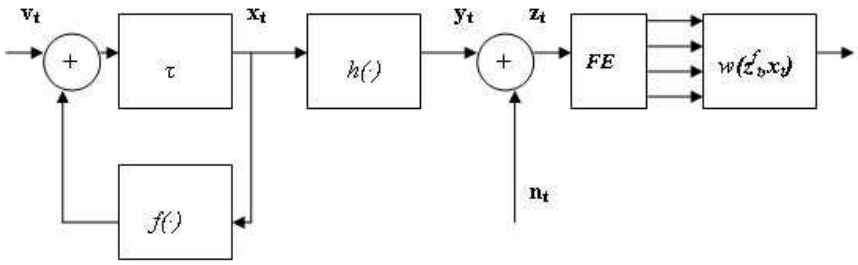


FIGURE 5.2: Particularization of a Gauss-Markov State-Space Model and Likelihood Approximation for Image Processing problems (FE stands for Feature Extraction)

Our goal is to take advantage from the diversity of image features to generate better layered proposal distributions in a new annealing strategy.

Regarding the features we have analyzed so far, we would like the particles to be sampled so that they fit the foreground silhouette and, from this sample set, to locally guide the estimation by means of the edge information. This scheme has two basic disadvantages:

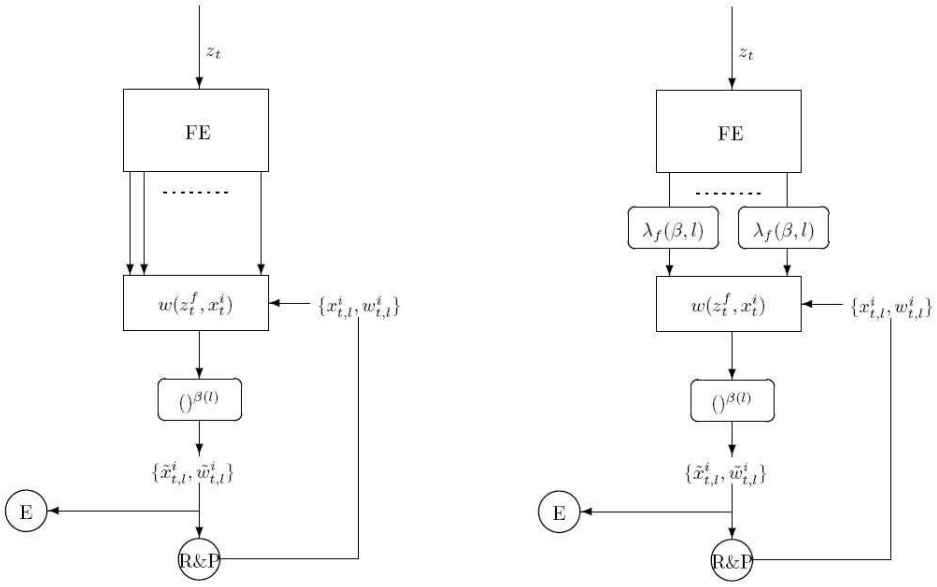
- A set sampled from a function of the foreground information will not be very reliable in case of self-overlaps. Nevertheless, we can use smoothed edge information to emphasize regions with occluding contours
- We cannot assure the local quasi-unimodality and convexity of such proposal distribution. We still can use overall smoothing to force a quasi-unimodal distribution.

One can formalize the aforementioned ideas by introducing $p(\mathbf{z}_{t,l}^{\beta(l)}|\mathbf{x}_{t,l}^{i}) = \prod_{f=1}^{F} p(\mathbf{z}_{f,t}|\mathbf{x}_{t,l}^{i})^{\beta(l)\lambda^{f}(\beta(l),l)}$ in equation 5.1 (where f denotes the feature measure index and $\lambda^{f}(\beta(l), l)$ is a parameter controlling the proportion of each feature measure). Proceeding in such manner, we are annealing the likelihood approximation by means of its features. Besides, it allows more flexibility in the importance pdf definition for each layer. In this new scheme, different measures derived from different image features can be combined with different importances per layer, regarding that it should be two groups of measure functions: one coarse and smooth and one highly peaked and probably multi-modal. Since it is difficult to make this classification with real data it does not make sense to find shades in between.

Considering the features analyzed in chapter 4, the resulting likelihood approximation for the presented scheme can be formulated as follows:

$$\omega = exp\left(-\sum_{c=1}^{C}(\lambda_c^{fg}(\beta(l),l)\omega^{fg} + \lambda_c^{e}(\beta(l),l)\omega^{e} + \lambda_c^{d}(\beta(l),l)\omega^{d})\right) \tag{5.2}$$

where λ_c is a weighting coefficient depending on the annealing rate β, the feature importance and the camera view, fg denotes foreground matching, d foreground divergence and e edges. We call it Feature-Based Annealing, and it constitutes a generalization of the classical exponent-based annealing involving image features. We show a simple example with actual data in figure 5.4. Foreground divergence measure is used as a predominant function in the first layers, thus penalizing some secondary modes produced mainly by edges. Therefore, the probability of the particle set to hit the probability mass volume of the sharpened likelihood increases.



(a) Classical Annealed PF

(b) Feature-Based Annealing PF

FIGURE 5.3: Annealing Schemes for Particle Filtering. Input Images (observations z_t) are processed in a Feature Extraction (FE) module. The output is a set of image features that are used to define a separate weighting functions. Note that in Feature-Based annealing a coefficient is applied to every single measure function in order to ponder each image feature. The result is smoothed and an estimation (E) can be provided after each weight computation. Resampling and Propagation (R&P) are applied also after every computation of weights

.

Obviously, this concept can be extended to more features or even the same feature extracted with several thresholds or successively filtered. However, an important issue

must be addressed: how do we determine if a feature produces a function which is suitable for sampling from?

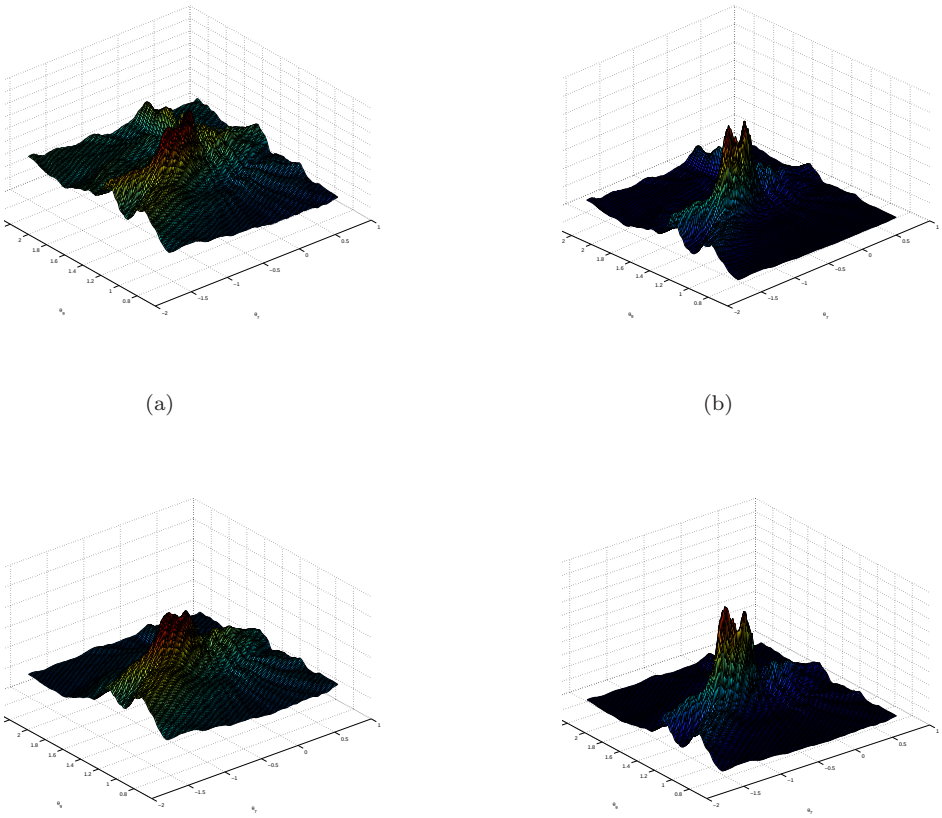
5.2.1 Practical Issues

In order to tune and to extend the concept of feature-based annealing, the role of every feature and its corresponding weighting function becomes crucial. In order to determine whether a weighting function is suitable to sample from in the first layers or is very determinant to locally estimate the pose, we have used two basic tools:

- Common knowledge and intuition about the appearance of the weighting functions.
- Evaluation of the weighting functions with actual data.

According to these sources of information and the features analyzed in chapter 4, we can give some insights of the parameters in equation 5.2: λ_c should be directly proportional to β in edge measure and inversely proportional in the foreground measures. Therefore, edge measures would be strongly smoothed in the first layers and progressively sharpened while foreground functions would behave the other way around. This is closely linked to the proposed classification of features. Based on actual data, we have considered foreground measures as locally quasi-unimodal and quasi-convex functions (belonging to the first group of features) and edge measurements as strongly peaked and multimodal (belonging to the second group).

However, for high-dimensional state spaces (such as pose space) intuition and visualization of actual data are weak tools for tuning the algorithm. As well as in the classical APF, is not easy to find a method to properly tune the parameters. We propose a simple test to determine the quasi-unimodality and quasi-convexity conditions needed for some functions to lead the first layers of the algorithm. This test is parameter dependent; it depends on the values of λ_c (and, consequently, of $\beta(l)$). The algorithm is run with several layers for separated image features and different $\beta(l)$ (evolving with a known rate, for instance a geometric progression). In each annealing run, the resampled particle set is modelled as a mixture of Gaussians with diagonal covariance in the pose space (the reference coordinate is marginalized). To this end, an Expectation-Maximization algorithm with automatic selection of the number of Gaussians is applied. In our case, the Figueiredo and Jain algorithm [35] has been used. The goal is to determine the number of gaussians that maximize the likelihood of the given particle set. This is similar to approximately determine the number of modes of the weighting functions, since particle sets are drawn from smoothed versions of these functions.



(a) (b)

(c) (d)

FIGURE 5.4: Common Annealing using Foreground and Edge Matching (top row) vs Feature-Based Annealing with Foreground divergence as predominant feature in the first layers (bottom row). While the last layers are very similar, the first layers of the Feature-Based annealing are better in terms of presence of secondary modes thus leading to a more properly weighted set

When testing edge measures with two views, we typically obtain from 2 to 10 Gaussians in different layers. Surprisingly, a similar number of modes is produced by foreground matching function when 2 views are used. The reason is that many-to-one mapping and foreground misses strongly affect this function. Contrarily, foreground divergence measure presents only one mode for most of the tests run with two views. Hence, the test provides additional reasons to use the foreground divergence to build a weighting function for the 3D body tracker.

Chapter 6

Implementation Details

The present chapter is a prelude of the experimental results used to show the achievements of this thesis. Section 6.1 describes the experimental setup used to evaluate our contribution. Special emphasis is made on the conditions that make it realistic and therefore challenging. Section 6.2 presents the body model used in for the previously established experimental conditions and 6.3 describes the simplistic initialization procedure employed in the tests. Finally, section 6.4 is devoted to the implementation issues concerning the particle filtering and the annealing schemes discussed in this thesis.

6.1 Experimental Setup

We have tested the different body tracking approaches in an office desktop environment. The goal was to characterize the pose of several people while performing several common actions at a workplace. More concretely they did the following actions:

- Mouse dragging
- Picking the phone and talking on it
- Typing on the keyboard
- Picking a pen, writing a sentence on a paper and leaving the pen.
- Picking a cup and drinking
- Reading

Performing all these actions took approximately two minutes. Since the relevant motion is basically due to the arms, we have focused on upper body tracking.

Our setup was built under the premise of being portable, low cost and easy to configure. The hardware consisted of:

- 1 Laptop 2GHz Intel Core Duo with 2GB RAM
- 2 Logitech QuickCam Pro 9000 webcams connected to the same laptop.

  Both webcams were calibrated by means of the open source library ARToolkit [36]

  The selected views were one lateral and one frontal, with a little downtilt (see Fig. 6.1).

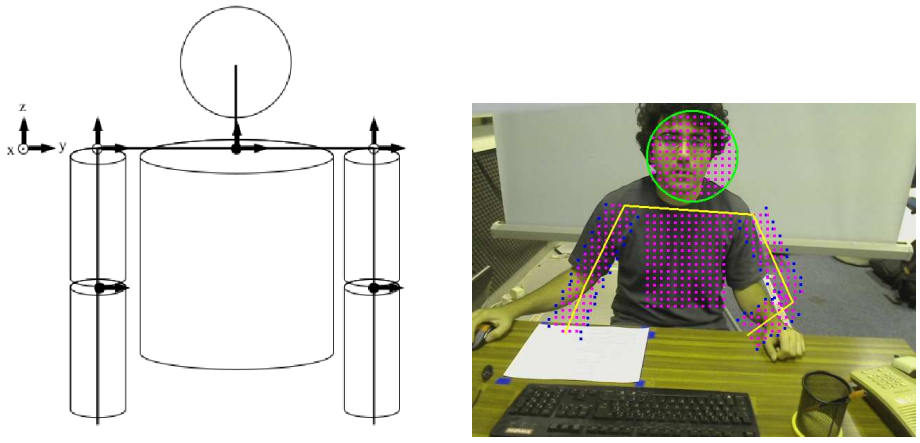
The total frame rate achieved with this configuration was 9.5 fps. Both views are relatively close to the subject, thus the apparent size of some limbs in the image can change notably depending on their 3D position. Besides, the spurious features introduced by clothing wrinkles and moving wires are not neglectible.

Moving objects, loose clothes and hardware setup simplicity make the scenario realistic and challenging. In addition to the aforementioned challenges, background is uncontrolled in the sense that is cluttered, illumination changes are not controlled and shadows appear. Furthermore, low frame rate makes some pose changes to become apparently abrupt and in some cases edges become blurred due to the apparent celerity of the motion.

FIGURE 6.1: Available views for the experimental setup

6.2 Upper Body Model

A simplistic articulated upper body model will fulfill the requirements of the described scenario. The model joints are the base of the neck, shoulders and elbows with a total of



(a) (b)

FIGURE 6.2: Articulated upper body model and its projection for a given particle

Angle	Joint	Rotation Axis	Range
θ_1	Base of the Neck	**y**	$\left[-\frac{\pi}{4}, \frac{\pi}{4}\right]$
θ_2	Right Shoulder	**x**	$\left[-\frac{\pi}{4}, \pi\right]$
θ_3	Left Shoulder	**x**	$\left[-\frac{\pi}{4}, \pi\right]$
θ_4	Right Shoulder	**y**	$\left[-\frac{\pi}{4}, \pi\right]$
θ_5	Left Shoulder	**y**	$\left[-\frac{\pi}{4}, \pi\right]$
θ_6	Right Shoulder	**z**	$\left[-\frac{\pi}{4}, \frac{\pi}{2}\right]$
θ_7	Left Shoulder	**z**	$\left[-\frac{\pi}{4}, \frac{\pi}{2}\right]$
θ_8	Right Elbow	**y**	$[0, \pi]$
θ_9	Left Elbow	**y**	$[0, \pi]$

TABLE 6.1: Articulated Body Model Joints

nine degrees of freedom (see table 6.1). x and z rotation axis are defined with different signs in order to allow the same angle range for left and right joints.

In order to set the model in a world position, a three-dimensional coordinate system built with the base of the neck as origin and a body orientation are defined. The world reference point for our model is set to be the base of the neck (see Fig. 6.2(a)). Therefore, the body model defines a thirteen-dimensional state vector:

$$\mathbf{x_t} = \{x_0, y_0, z_0, \theta_0, ..., \theta_9\} \tag{6.1}$$

Angle θ_0 is the orientation of the whole body model while all the other angles are designed following hard kinematic constraints.

6.3 Algorithm Initialization

The initialization is mainly a manual procedure. A default pose is sent to the algorithm in the first frame to generate the initial samples. This pose is approximately similar to an expected initial pose, although there might be differences that the tracker should reduce. An automatic local search of the head, based on skin and foreground detection, is performed in order to produce a better whole body location initial value and a better particle set initialization.

Limb sizes are written in a configuration file and are not modified during the tracker run, although they are slightly adapted for different subjects (mainly due to their height and inter-shoulder distance).

6.4 Particle Filtering

6.4.1 General PF Implementation Issues

So far we have explained the general annealing framework intended to increase the robustness of a generic particle filter-based human body tracker. Our purpose is to preserve a high degree of generality thus keeping as possible all the motions that are physically possible. We have shown the kinematic constraints added to our model in order to produce feasible poses. We expect that the subspace spanned by the possible poses of our model is representative of the physically possible subspace of human poses. However, hard kinematic constraints do not take into account the inter-penetration of limbs. We use an additional prior with the underlying idea of [5]. Some key points of the model are evaluated in terms of Euclidean distance to obtain the probability of being inside another limb, the head or the torso. Since it may be difficult to draw samples on the fly from this new prior, we incorporate the inter-penetration factor in the SIR formulation by forcing a dependence with an intermediate variable:

$$w_t^i \propto \frac{p(\mathbf{z}_t|\mathbf{x}_t^i)p(\mathbf{x}_t^i|\mathbf{x'}_t^i)p(\mathbf{x'}_t^i|x_{t-1}^i)}{p(\mathbf{x'}_t^i|x_{t-1}^i)}w_{t-1}^i \tag{6.2}$$

$$p(\mathbf{x}_t^i|\mathbf{x'}_t^i) \propto exp(-f(\mathbf{x}_t^i)|f(\mathbf{x}_t^i)|^p)\delta(\mathbf{x}_t^i - \mathbf{x'}_t^i) \tag{6.3}$$

where $f(\mathbf{x}_t) < 0$ denotes that a point is inside another limb and p controls the probability decay rate.

The computation of the annealing rate is performed somehow similarly to that in [6]. We use the survival rate to tune this parameter but we do not use an exact gradient descent algorithm, but simply a single correction based on the current survival rate:

$$\beta(l) = \beta(l-1) - \epsilon(\alpha_{target} - \alpha(l-1)) \tag{6.4}$$

where α_{target} is the desired survival rate per layer, $\alpha(l-1)$ is the survival rate computed after the weighting in the last layer and ϵ is a learning factor; we typically set it to $\frac{1}{1+l}$. We also force that $\beta(l) \geq \beta(l-1)$.

6.4.2 Annealing schemes for PF

We have implemented several versions of the annealing schemes for particle filtering presented:

1. **APF**: The APF proposed by Deutscher in [6].
2. **M-APF**: The aforementioned algorithm was modified to allow edges to be more determinant in the computation of weights. There are two reasons that justify this. The first one is the evident importance of edges when locating limbs. The second one is the fact that a perfect match between model and natural edges is very unlikely to happen in comparison to a foreground match.
3. **FBAPF**: A Feature-Based Annealing Particle Filter with the following likelihood approximation:

$$\omega^{\beta(l)} = exp\left(-3\beta(l)\sum_{c=1}^{C}[\beta(l)(5\omega^{e}) + \frac{1}{\beta(l)}(2\omega^{fg} + (e^{1.5\omega^{d}} - 1))]\right) \tag{6.5}$$

   The factor 3 multiplying the whole exponent has been also applied in the first and second annealing schemes. Edge information is scaled according to the reasons described in the above scheme (in fact, the scale is the same for both edges and foreground matching measures). Based on experimental tests, an exponential is used to produce a narrower proposal from foreground divergence measure.

4. **FBAPF2**: A Feature-Based Annealing Particle Filter with the following likelihood approximation:

$$\omega^{\beta(l)} = exp\left(-3\beta(l)\sum_{c=1}^{C}[\beta(l)(2\omega^{fg} + 5\omega^{e}) + \frac{1}{\beta(l)}(e^{1.5\omega^{d}} - 1)]\right) \tag{6.6}$$

5. **FBAPF3**: A Feature-Based Annealing Particle Filter with the following likelihood approximation:

$$\omega^{\beta(l)} = exp\left(-3\beta(l)\sum_{c=1}^{C}[\beta(l)(5\omega^{e}) + \frac{1}{\beta(l)^{2}}(2\omega^{fg} + (e^{1.5\omega^{d}} - 1))]\right) \tag{6.7}$$

Note that all the proposed feature-based annealing schemes are very simple and they basically aim at exploiting the quasi-convex nature of some well-defined foreground measures in order to have better proposal distributions in the first layers.

Chapter 7

Results and Discussion

In this chapter we show experimental results to assess the increased robustness of our proposal under the challenging conditions described in chapter 6.

7.1 Evaluation Procedure and Metrics

In order to evaluate the precision of the proposed body tracking systems, we use 3D error measures [37]. Consider a set of M virtual markers placed in specific body locations. Then we can write the body model as a vector comprising the positions of all the markers as $X = \{x_1, \ldots, x_m, \ldots, x_M\}$, where $x_m \in \Re^3$. In our particular case, these virtual markers are specific upper-body locations: head centroid, shoulders, elbows and wrists. These 3D virtual markers are obtained at approximately every second by means of manual annotation. Let us call $\hat{X}$ the vector of estimated locations corresponding to the ones found in X. $\hat{X}$ is easily obtained from the estimate $\hat{\mathbf{x}}_t$ by means of the product of exponential maps formulation. Then, the pose estimation error is computed as follows:

$$D(X, \hat{X}) = \frac{\sum_{m=1}^{M} \|x_m - \hat{x}_m\|}{M} \tag{7.1}$$

where $D(X, \hat{X})$ is the distance expressed in mm. Note that the existence of the pairs ground truth-estimate is assumed and that there is a one-to-one correspondance between both. In benchmarking campaings could happen that different models with different complexity need to be evaluated, thus requiring a selection of the common body locations. In our case this is not needed, since we have control over the whole modelling,

Method	μ_1 (mm)	σ_1 (mm)	μ_2 (mm)	σ_2 (mm)	μ_3 (mm)	σ_3 (mm)
APF	143	69.2	187	72.5	203	99.6
M-APF	242	86.9	131	97.63	169	70.2
FBAPF	**81**	**20.2**	98	39.3	**107**	**38.7**
FBAPF2	87	27.5	**94**	**34.1**	113	40.0
FBAPF3	120	29.6	175	49.6	145	59.66

TABLE 7.1: Tracking results for three sequences of three different subjects. 3 layers and 200 particles per layer have been used for all the schemes and subjects

estimation and annotation process. Furthermore, no matching criterion is required to make pairs $x_m, \hat{x}_{m'}$, since annotation and estimation are given in the same exact order.

When dealing with a whole sequence of T frames, the overall pose error can be expressed as the mean and the standard deviation:

$$\mu_{seq} = \frac{1}{T}\sum_{t=1}^{T} D(X_t, \hat{X}_t) \tag{7.2}$$

$$\sigma_{seq} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\Big(D(X_t, \hat{X}_t) - \mu_{seq}\Big)^2} \tag{7.3}$$

7.2 Experimental Results

Each one of the annealing schemes[1] described in section 6.4.2 is run 5 times over 3 sequences from 3 different subjects performing the list of actions detailed in section 6.1. For every time instant, the mean pose error over the 5 trials is computed. The final averaged sequence of pose errors is used to obtain the error metrics detailed in the previous section. The result of these metrics is shown in table 7.2.

We can clearly see that, although outperforming classical annealing in mean in most of the frames, FBAPF3 is not correctly parametrized, regarding the results obtained by the other feature-based annealing schemes.

Part of the performance gain is due to the additional foreground divergence measure. We include this function in Deutscher's APF scheme and we compare the mean results obtained with FBAPF (table 7.2).

[1] APF is Deutscher's annealing concept for particle filtering, M-APF is a modified version of APF; FBAPF, FBAPF2 and FBAPF3 are three simple parametrizations of the feature-based annealing concept

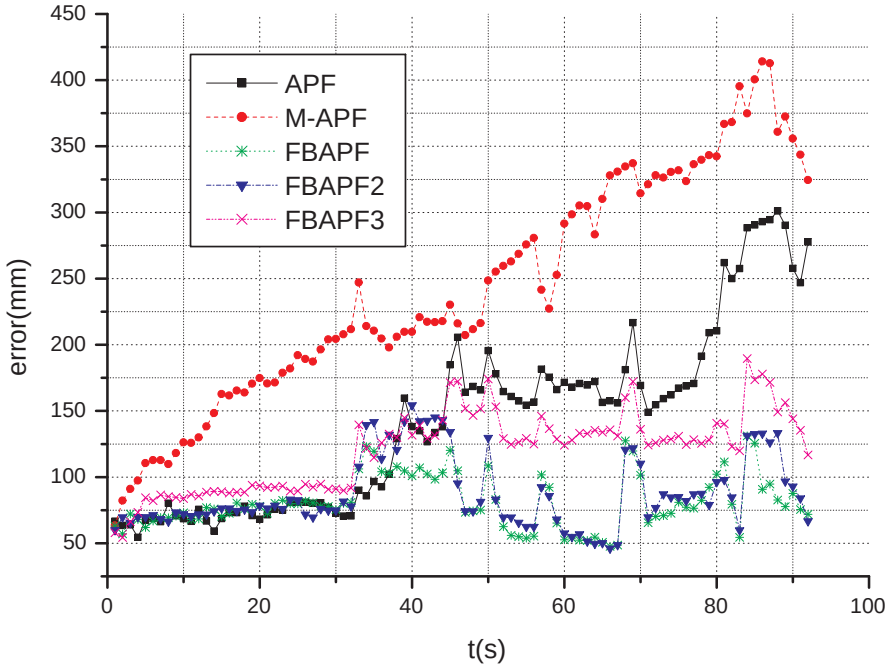


FIGURE 7.1: Comparative results for subject 1 using 3 layers and 200 particles per layer

Method	μ_1 (mm)	σ_1 (mm)	μ_2 (mm)	σ_2 (mm)	μ_3 (mm)	σ_3 (mm)
FBAPF	81	20.2	98	39.3	107	38.7
APFdivergence	107	54.3	112	50.2	116	41.8

TABLE 7.2: Comparative results between the classical annealing and feature-based using the same features. 3 layers and 200 particles per layer have been used in the three sequences

7.3 Discussion

Experimental results show that simple feature-based annealing strategies outperform classical annealing methods in particle filtering with similar parameters. The increased robustness is clearly perceived in moderately long sequences, where annealing particle filter becomes instable with a low number of views and low frame-rate. Recall that, as mentioned in section 6.4.2, the evaluated feature-based annealing schemes are not sophisticated because they only aim at showing the potential performance gain of simple parametrizations.

Considering our test conditions, emphasizing edge measures in the weighting function used by classical annealing is risky, as figure 7.1 shows, but can provide better accuracy in terms of mean distance to ground-truth if critical failures do not take place (see figures 7.2 and 7.3). Special attention must be paid to the second sequence, where a bad

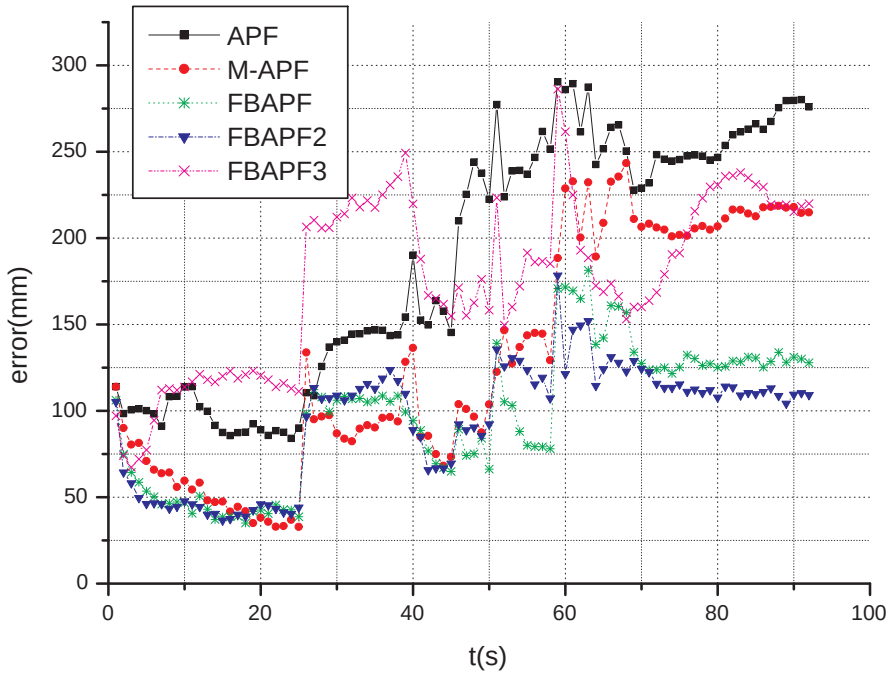


FIGURE 7.2: Comparative results for subject 2 using 3 layers and 200 particles per layer

initialization is given and the modified version of APF is able reduce the error, while common APF cannot.

The stability gain produced by sampling from improved likelihood proposal distributions is reflected in the standard deviation values computed in the experiments. Although the values are high in general (regarding the mean error for the sequences), the use of feature-based annealing implies deviation reductions ranging from 7 to 63% with respect classical annealing and the same weighting functions (edges, foreground matching and foreground divergence) in this difficult scenario.

Experiments evidence that the foreground divergence measure improves the performance of the Deutscher's APF in this difficult scenario. Besides, the properties of the resulting weighting function, make it a good candidate to sample from in the initial stages of the annealing process. Consequently, this measure is a crucial element in our achievements. However, it presents important drawbacks. First, and most evident, it is model dependent in the sense that in order to explain a silhouette, the model should produce a reasonable human shape. Size missadjustments of model body parts may affect the reliability of the foreground divergence. Nevertheless, our simple initialization process and flesh model shows that it is not a critical issue when views are controlled. In the case of a highly moving subject (typically walking or running along strange paths),

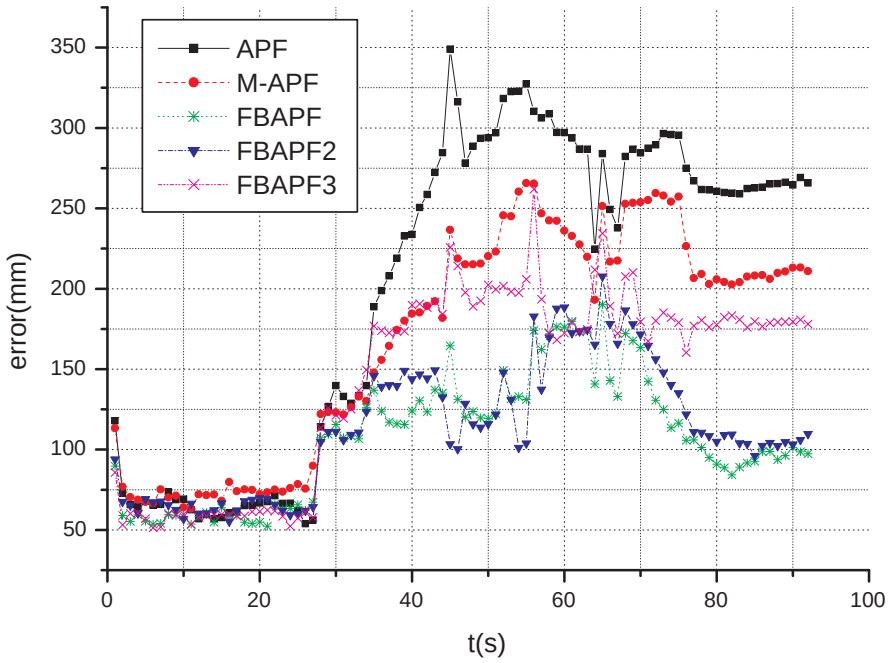


FIGURE 7.3: Comparative results for subject 3 using 3 layers and 200 particles per layer

two views might not provide such determinant silhouettes, hence more views and an self-occlusion/self-overlap test may be needed. A more critical factor may appear in multi-person tracking or in scenarios with apparently big and occluding objects detected as foreground objects. In those cases, the maximum occupancy of the observations by the model should be redefined. If the observations of interest are multiple foreground silhouettes, then the divergence should be measured in a bounding volume enclosing the target and occlusions should be taken into account. Another possibility is to use an explicit representation such as a volumetric reconstruction to avoid occlusions. Finally, the computational cost of matching the generated model silhouette and the observation silhouette must be taken into account. This could be partially overcome using hardware graphics algorithms.

With 200 particles and 3 layers, none of the schemes was able to consistently track some specific motions (picking the phone, for instance; see figure 7.4). Low frame rate and spatial ambiguity are the main limitations to obtain consistent tracks. However, feature-based annealing is able to keep acceptable results after track loss. In some runs, due to the random sampling, these difficult motions were approximately tracked (see figures 7.5 and 7.6). Since these samples are likely to occupy the tails of the gaussian state transition priors, we expect that a higher number of particles may increase the probability of drawing favorable random samples.

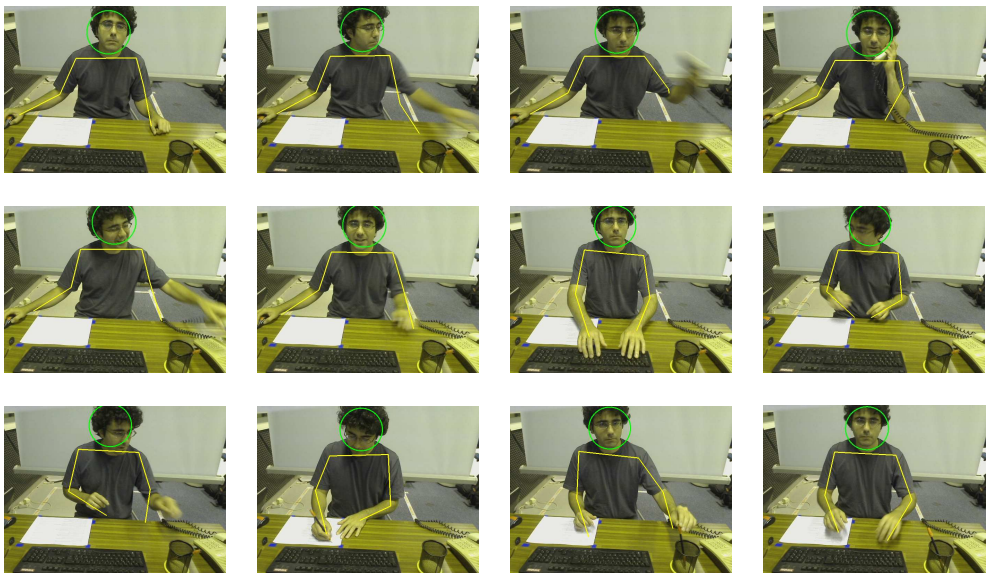


FIGURE 7.4: Tracking examples of sequence 1. Feature-based annealing with 200 particles and 3 layers has been used. Picking the phone produces a big tracking error. However, the tracker is able to recover the pose after several errors. Note the blurring effect when arms are moving.

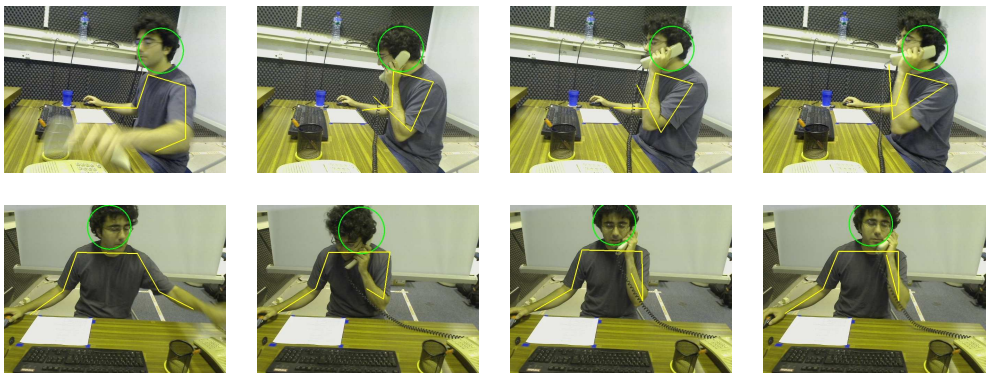


FIGURE 7.5: Tracking examples of sequence 1 for action "picking the phone" for a run in which the feature-based annealing particle filter (200 particles and 3 layers) approximately estimates the poses involved in the action. Top row: Lateral view. Bottom row: Frontal View. Pose ambiguities for these two views can be observed. The imprecise estimation of the left arm in the two last frames is mainly due to the strong edge introduced by the wire. In this case, two strong modes appear in the likelihood, thus the posterior mean estimate falls between them.

FIGURE 7.6: Tracking examples of a single run in sequence 3 with feature-based annealing particle filter (200 particles and 3 layers). First picture of bottom row shows a tracking loss of the right arm. The next picture illustrates the limitations of the model to estimate pronounced torso inclinations. In addition, right arm cannot be viewed from lateral camera, thus yielding an imprecise right arm pose estimate.

Chapter 8

Conclusions and Future Work

We have presented a generalization of the annealing schemes for particle filtering in the context of body pose estimation and tracking. An image processing signal model for the pose space has been presented to justify the use of a technique treating the estimation of the dynamical system from an optimization viewpoint.

Feature-Based Annealing aims at providing better proposal distributions from likelihood, which it is expected to be more correlated with the posterior. The technique is based on two main assumptions. First, the true underlying statistics of a pose at a precise time instant are somehow a narrow function masked mainly by observation noise. Then, it makes sense to use a single statistic figure, the posterior mean, in order to give an estimate of the pose. Second, two main groups of weighting functions can be constructed from multiple images: one group of quasi-convex and quasi-unimodal functions and a second group of highly peaked and probably highly multi-modal weighting functions. The conditions of both groups must hold in the high probability zones of the resulting pdf after the prediction step or, equivalently for Monte Carlo methods, in the support of the available sample set after the prediction step. In order to classify them accordingly, we have proposed simple tests to have insights about the properties of weighting functions obtained by different features .

We have introduced foreground divergence measure consisting in measuring the occupancy of the foreground silhouette by a generated model silhouette. This model silhouette is determined by the projection of the flesh model in a given pose. The need for this complementary foreground measure is justified because of the properties of the resulting weighting function. This complementary foreground measure reduces the pose ambiguity, thus increasing the robustness of the classical annealing scheme for particle filtering. Besides, it is locally quasi-convex and quasi-unimodal, thus providing a good function to sample from in the first layers of the feature-based annealing.

We have tested our proposal with a simple body configuration and a simplified projection method in a challenging scenario. A quantitaive comparison between our proposal and common annealed particle filter has been presented. The feature-based annealing strategy shows an increased robustness under challenging experimental conditions.

Like in the simple annealed particle filter we have tried to preserve the tracker generality by only adding hard kinematic constraints to our model and a prior function for inter-penetration of limbs. Under a challenging scneario, our approach is not able to consistenly track some fast motions. This could be attributed to a limitation of the hardware and means of the setup that has been used to test the algorithm, the state-space model and the common propagation model of the Sampling Importance Resampling framework from which annealing particle filter is derived.

Future research involves further validation of feature-based annealing with full body models and several recording conditions, and the extension of this study to other image features, including spatio-temporal features.

Another important issue concerns the statistics of the pose. We have assumed that these are somehow a narrow function centered around the true pose at a given time instant, thus justifying the need of an estimator treating the likelihood in an optimization context. In spite of this, the final posterior estimate appears to be multimodal, mainly due to observation ambiguities. Hence, the common posterior mean seems to be a weak statistical figure. Further research can be done in order to study the modes of the posterior to provide better pose estimates. In addition, the variety of weighting functions used in pose estimation can be addressed to exploit the correlation between several likelihood approximations in order to infer the mode that better approximates the true pose.

Bibliography

[1] MS Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-GaussianBayesian tracking. *Signal Processing, IEEE Transactions on*, 50(2):174–188, 2002.

[2] M. Isard and A. Blake. CONDENSATION-Conditional density propagation for visual tracking. *Int. Journal of Computer Vision*, 29(1):5–28, 1998.

[3] J. MacCormick and M. Isard. Partitioned Sampling, Articulated Objects, and Interface-Quality Hand Tracking. *Lecture Notes in Computer Science*, pages 3–19, 2000.

[4] J. Mitchelson and A. Hilton. Simultaneous pose estimation of multiple people using multiple-view cues with hierarchical sampling. *In Proc. of BMVC, September*, 2003.

[5] C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3d body tracking. *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, 1:I–447–I–454 vol.1, 2001. ISSN 1063-6919.

[6] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, 2:126–133 vol.2, 2000.

[7] J.S. Liu and R. Chen. Sequential Monte Carlo methods for dynamical systems. *Journal of the American Statistical Association*, 93(5):1032–1044, 1998.

[8] Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11: 125–139, 1998.

[9] F.C. Park. Computational aspects of the product-of-exponentials formula for robot kinematics. *Automatic Control, IEEE Transactions on*, 39(3):643–647, 1994. ISSN 0018-9286.

[10] B. Hall. Lie Groups, Lie Algebras, and Representations: An Elementary Introduction (Graduate Texts in Mathematics vol 222), 2003.

[11] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3): 197–208, July 2000.

[12] N.J. Gordon, D.J. Salmond, and A.F.M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2):107–113, Apr 1993. ISSN 0956-375X.

[13] D. M. Gavrila and L. S. Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *CVPR '96: Proceedings of the 1996 Conference on Computer Vision and Pattern Recognition (CVPR '96)*, page 73, Washington, DC, USA, 1996. IEEE Computer Society. ISBN 0-8186-7258-7.

[14] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[15] C. Canton-Ferrer, JR Casas, and M. Pardas. Exploiting Structural Hierarchy in Articulated Objects Towards Robust Motion Capture. *Lecture Notes in Computer Science*, pages 82–91, 2008.

[16] Fabrice Caillette. *Real-Time Markerless 3-D Human Body Tracking*. PhD thesis, University of Manchester, 2005.

[17] C. Bregler and J. Malik. Tracking People with Twists and Exponential Maps. *In Proc. CVPR (1998)*, 1998.

[18] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevance vector regression. *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 2:II–882–II–888 Vol.2, June-2 July 2004. ISSN 1063-6919.

[19] I. Mikic. *Human Body Model Acquisition and tracking using multi-camera voxel Data*. PhD thesis, University of California, San Diego, 2003.

[20] L. Raskin, E. Rivlin, and M. Rudzsky. Using Gaussian Process Annealing Particle Filter for 3D Human Tracking-Volume 2008, Article ID 592081, 13 pages. *EURASIP Journal on Advances in Signal Processing*, 2008.

[21] A. Laurentini. The visual hull concept for silhouette-based image understanding. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(2):150–162, 1994.

[22] F. Caillette, A. Galata, and T. Howard. Real-Time 3-D Human Body Tracking using Variable Length Markov Models. *British Machine Vision Conference*, 1:469–478, 2005.

[23] Hedvig Sidenbladh, Michael J. Black, and Leonid Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV'02) - volume 1*, number 2350 in Lecture Notes in Computer Science, pages 784–800, Copenhagen, Denmark, May 2002. ISBN 3-540-43745-2.

[24] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989. ISSN 0018-9219.

[25] A. Galata, N. Johnson, and D. Hogg. Learning Variable-Length Markov Models of Behavior. *Computer Vision and Image Understanding*, 81(3):398–413, 2001.

[26] Neil D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *In NIPS*, page 2004, 2004.

[27] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian process dynamical models. In *Advances in Neural Information Processing Systems (NIPS) 18*, pages 1441–1448, Vancouver, Canada, December 2005. MIT Press.

[28] Shaobo Hou, A. Galata, F. Caillette, N. Thacker, and P. Bromiley. Real-time body tracking using a gaussian process latent variable model. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct. 2007. ISSN 1550-5499.

[29] C. Stauffer and W.E.L. Grimson. Learning Patterns of Activity Using Real-Time Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 747–757, 2000.

[30] L.Q. Xu, JL Landabaso, and M. Pardas. Shadow Removal with Blob-Based Morphological Reconstruction for Error Correction. *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 2, 2005.

[31] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[32] D.J.C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[33] N. Vaswani. Particle filtering for large-dimensional state spaces with multimodal observation likelihoods. *Signal Processing, IEEE Transactions on*, 56(10):4583–4597, Oct. 2008. ISSN 1053-587X.

[34] A. O. Balan, L. Sigal, and M. J. Black. A quantitative evaluation of video-based 3d person tracking. In *ICCCN '05: Proceedings of the 14th International Conference on Computer Communications and Networks*, pages 349–356, Washington, DC, USA, 2005. IEEE Computer Society. ISBN 0-7803-9424-0.

[35] Mario A. T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:381–396, 2002.

[36] H. Kato and Mark Billinghurst. Marker tracking and hmd calibration for a video-based augmented reality conferencing system. In *Proceedings of the 2nd International Workshop on Augmented Reality (IWAR 99)*, San Francisco, USA, October 1999.

[37] L. Sigal and M. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical report, 2006.