Applied Bionics and Biomechanics 9 (2012) 443–456
DOI 10.3233/ABB-2011-0015
IOS Press

Combination of annealing particle filter and belief propagation for 3D upper body tracking

Ilaria Renna [a,b,*], Ryad Chellali [b] and Catherine Achard [a]
[a] *Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie/CNRS, UMR 7222, Paris, France*
[b] *IIT – Italian Institute of Technology Robotics, Brain and Cognitive Sciences Department, Genova, Italy*

Abstract. 3D upper body pose estimation is a topic greatly studied by the computer vision society because it is useful in a great number of applications, mainly for human robots interactions including communications with companion robots. However there is a challenging problem: the complexity of classical algorithms that increases exponentially with the dimension of the vectors' state becomes too difficult to handle. To tackle this problem, we propose a new approach that combines several annealing particle filters defined independently for each limb and belief propagation method to add geometrical constraints between individual filters. Experimental results on a real human gestures sequence will show that this combined approach leads to reliable results.

Keywords: Body tracking, particle filter, belief propagation

1. Introduction

Robot companion is considered as the key element of service and domestic robotics [1]. Mainly for home-care in assisting elderly and disabled people, these robots may perform soon tasks that could help for home nursing or assisted living facility. One of the principal requirements expected from these machines is communication. Indeed, the co-living and collocated robots must be able to exchange information easily and naturally with users: robots might be anthropomorphic by understanding human intents on one hand, and displaying a directly interpretable status on the other hand [41]. This bi-directional information flow passes through the called human-robot-interface (HRI). The later is more and more considered as a research field itself in robotics. Indeed, even if HRI shares sensory-based concepts and tools with classical robotics autonomy [2], it considers communication as a specific autonomic process itself and the robot sensory system must capture users intents. This task concerns both the medium and the content. The first is dealing with the sensory channel through which the information is addressed; the second one is more related to semantics of the exchanged messages. For instance, gestures are generated by arms and recognized by eyes: arms and eyes are mediums while arms movement are human intent signals. In our work, we concentrate on gestures. Mainly, we are aiming to develop a cost-effective system, to be embedded on home robots, able to recognize and to understand human gestures. It is a fact that gestures constitute an important part of means humans employ to communicate with each

*Corresponding author. Ilaria Renna, Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie/CNRS, UMR 7222, Paris, France. Emails: ilaria.renna@upmc.fr (I. Renna), ryad.chellali@iit.it (R. Chellali), catherine.achard@upmc.fr (C. Achard).

others. Gestures are used for everything from pointing at a person to get his attention to conveying information about space and temporal characteristics [3], they can complete other modalities, and some times can be central when speech-based communication is not possible [4] or has to be completed [5]. However, gestures recognition is known to be a complex problem. From robotics and computational points of view, gestures are sequences of movements performed by non-rigid bodies. The recognition of such sequences must first solve the human body pose estimation problem. Once the pose is known over time, one can tackle the classification problem. Exact solutions of the first part of the problem can be found on the shelf. Indeed, a lot of commercial products providing in real time human body poses exist. Unfortunately, the use of such systems is constraining: one needs to rely on an ad-hoc infrastructure such as markers for vision based systems or specific sensors (electromagnetic, IR or US) making these solution inaccessible to a large and cost-effective dissemination for home robots.

In our case, we choose vision sensing based system. Many researchers for a large spectrum of applications have adopted this choice where localization, tracking and 3D modeling are the applications' core. Following that, 'pose estimation' problem can be stated as the localization of moving and/or static articulated and linked parts from a sequence of 2D images. This involves the identification of the set of subparts (head, arms, hands, etc.), the estimation of their respective poses in the three-dimensional space and the tracking of each of them when possible.

This paper reports on the work we achieved concerning low-cost human body pose estimation. It is a part of a larger project dealing with homecare robotics. We use a minimalistic hardware, namely, a single camera and a normal PC allowing the algorithm we developed to run in real time. The outline of the paper is as follows. In section 2, we sum up some related works, while in section 3 we state the problem. In section 4, we introduce the model representing the human body on which we based all the steps of our algorithm. In section 5, we detail the image processing steps leading to the extraction of the 2D features on which the pose estimation is based. Sections 6 and 7, introduces the proposed stochastic algorithm, a combination of Annealed Particle Filters and Belief Propagation concepts and their application to upper body tracking. Experimental results are shown and discussed in section 8, while conclusions are drawn in section 9.

2. Related work

A great number of works deals with vision-based human motion capture and a recent review can be found in [6]. Human pose estimation can be obtained in a tracking context, or directly, by analysing video images. In a general way, these methods can be classified in three categories:

- *Methods based on a probabilistic limb combination*. The first step of these methods is an independent detection of each limb, according to 2D features, that can be achieved thanks to pictorial structure [7, 8], SVM [9] or AdaBoost algorithm [10]. Body pose is then estimated using geometrical constraints obtained with dynamic programming [11], with Integer Quadratic Programming [12] or with Markov networks [13]. Temporal coherence of limbs can then be enforced using Hidden Markov Model [14]. The advantage of these approaches is that, as any assumption is made on the movement, they are robust to complex and fast motion. Conversely their main drawback is that their performances strongly depend on limbs detection quality that is a difficult task, especially in presence of complex background.
- *Examples based learning methods*. These methods compare the observed image with a database of samples. Mapping between 2D observed image features and 2D or 3D pose can be reached using HMM [15], neural networks [16], lookup tables [17], parameter-sensitive hashing [18] or linear [19]/non-linear regression [20]. A limitation of these approaches concerns the generality of the database: appearance variability due to 3D pose configurations, view points, body dimensions, clothes, lighting changes, backgrounds, and so on, yields to an infinite number of 2D images.
- *Model based methods*. For these methods, 3D pose can be estimated with several views [21] or unique view [22] (which is a more challenging problem). However, in some applications, such as human/robot interaction, it is not possible to observe the human with different viewpoints. The first model-based approaches were grounded on a deterministic gradient descent technique [23], or on Kalman filter [24]. As only one tracking hypothesis is considered, these methods are not robust and fail with complex motion. In

probabilistic approaches [25], motion analysis is expressed as a Bayesian inference problem that can be solved with particle filters based strategies [26] where posterior distribution is represented by a set of samples (or particles) with associated weights. This weighted particles set is updated over time taking into account the measurements and a prior knowledge on the system dynamics, and observation models. Particle filters cannot be directly applied to body tracking due to the important degree of freedom (DoF) of the target. In fact, a realistic articulated model of human body is usually composed by at least 20 DoF, and as computational costs increase exponentially with the number of DoF, the exploration of the configuration space has to be optimized. Different methods have been investigated to reduce the large number of particles necessary to solve this high dimensional problem. MacCornick and Isard [27] proposed the partionned sampling of the state space, Deutscher et al. [28] introduced the annealed particle filter and Sminchisescu and Trigss [29] presented the stochastic sampling. Another solution consists in working with the likelihoods of each limb, which are then combined according to the geometrical human body model thanks to Markov chain Monte Carlo (MCMC) [30] or sequential Monte Carlo approaches [31].

Another way to cope with the high dimensionality is to decompose the state space of the 3D human body in several state spaces with lower dimension: one for each body part. Human body is then represented as a graphical model where individual limbs are characterized by nodes and relationships between body parts are represented by edges connecting nodes and encoded by conditional probability distributions. This graphical model allows tracking each subpart individually, and then adding constraints between adjacent limbs thanks, for example, to Belief Propagation (BP) inference algorithm [32, 33, 34, 21, 35]. In this way, the high dimensionality problem is expressed as a set of lower dimension, and thus the complexity of the search task is linear, rather than exponential, according to the number of body parts.

3. Problem statement

Finding human body poses and tracking its subparts is a highly combinatory problem. In fact, a realistic articulated model of the human body is usually composed by at least 20 DoF, and as computational costs increase exponentially with the number of DoF, the configurations space and its exploration have to be optimized. In the following, we give some hints concerning both aspects.

3.1. Configurations space

Configuration space is the space in which each point or vector encodes a human posture. The encoding technique is sensed data-dependent: the vectors are related to the sensing data that one can extract from the images. To achieve this goal, three main approaches are developed. The first one relies on 3D body models. The second one is dealing with 3D based reconstruction and the last one uses *a priori* knowledge, namely some learned human body poses. These approaches are in fact derived from classical computer vision techniques for object recognition or/and reconstruction.

For 3D models based approaches, the object (the human body) is represented by a set of connected and mobile geometrical primitives. Sticks, cylinders or more sophisticated shapes (spheres for instance) represent body parts. Human body pose is then obtained by fitting the projection of the 3D model on the image plane to the 2D image features (contours, skin color, apparent motion, etc.).

Reconstruction based approach uses generally multi-views techniques and projective geometry such as Shape from Silhouette, Shape From motion, Bundle adjustment, etc. to fuse information provided by multi-imaging systems (at least two images).

The last solution deals with learning. Pre-registered data are first obtained from different points of view as well as for different body configurations. This database is then visited to search for the current observation.

Unfortunately, for all previous approaches, the pose estimation problem is an inverse and ill-posed problem. Indeed, images are 2D projective entities that are used to derive 3D entities and the analytic solution does not exist. This derivation is based mainly on regularization and optimization by using redundant information in order to reduce the solutions space. Stereoscopy for instance reduces the solution space through the epipolar constraint. In addition, some other issues like occlusions and mechanical singularities increase the complexity. Finally, additional complexity has to be taken into for the tracking of articulated body due to the sizeable degrees of freedom of the target.

Regarding the previous discussion, the configurations space appears complex because it has a high dimensionality and its relationship with the image space is not bijective.

3.2. Search strategies

One way to solve the tracking problem is to use, as already seen, a probabilistic approach. Indeed, the motion analysis can be expressed as a Bayesian inference problem. As the body parts are dependant, the probability of a given configuration is conditioned by the upper body topology. Among known Bayesian solvers, one is well adapted for our problem, namely, the particle filters based. The strategies supported by this method, allow representing the posterior distribution as a set of samples (or particles) with associated weights [26]. This set is updated over time taking into account the measurements (image features in our case), a prior knowledge on the system dynamics, and observation models. Unfortunately, it is well known that

1- the number of particles required raises exponentially with the dimensionality of the configuration space.
2- to have an accurate and plausible solution, we need a maximum of particles.

To avoid these antagonist requirements, we considered different developed methods that have been investigated with the aim to reduce the particles number. Some techniques proposed the Annealed Particle Filter (APF). This last performs a coarse-to-fine layered search [28]. This modified particle filter uses a continuation principle based on annealing, to introduce the influence of narrow peaks in the fitness function, gradually. This allows reducing by a factor of 10 the number of particles and, as a consequence, to significantly decrease computation times.

To adapt the previous approach to our problem, we represent 3D human body as a graphical model, where individual limbs are characterized by nodes and relationships between body parts are represented by edges connecting nodes and encoded by conditional probability distributions. Additional edges can also be introduced to manage partial or fully occlusions. This graphical model allows to track each subpart individually, and then to add constraints between adjacent limbs. By doing so, it was possible for us to add the Belief Propagation (BP) inference algorithm [21, 35–39]. In this way, the initial high dimensionality problem is expressed as several problems of lower dimension, and thus the complexity of the search task is linear rather than exponential according to the number of body parts.

This article presents the development of a markerless human motion capture system that works with a standard camera coupled with a PC and does not require additional equipments. The system is based on a 3D articulated upper human body model and combines the advantages of above mentioned approaches to decrease the algorithm complexity induced by the high dimensionality of the problem. Rather than track the whole articulated body, each limb is tracked independently thanks to several particles filters (one for each limb); then, a BP method on factor graphs is used to estimate the current marginal of each limb according to geometrical constraints between limbs. Indeed, since belief propagation messages are represented as sums of weighted samples, the belief of each limb is approximated by a collection of samples. So, the association of belief propagation and particles filters algorithms is quite natural. Rather than a simple particle filter, we propose to use the annealing particle filter in this context. This combination of APF and BP allows decreasing the number of particles required per limb (and thus computation times) without modifying the quality of results.

4. The articulate body model

The body is represented by a graphical model [12, 36] of nodes and edges, where each node in the graph corresponds to a body part, and each edge represents the spatial constraints between adjacent connected body parts. Each node has an associated configuration vector defining position and orientation of the body part in the 3D space and a corresponding image likelihood function that models the probability of observing image measurements conditioned on the position and orientation of the part. Each edge has an associated conditional distribution that models the probabilistic relationship between parts. Additional edges related to non collision constraints or to the propagation of state variables across the temporal domain are added. A factor graph [40] is then constructed to decompose the join probability as product of factors.

The considered factors are of three different types: link factors between two nodes at the same time, image likelihood factors between all parts and their

corresponding observations at the same time and time coherence factors that link a part at two consecutive times (Fig. 1). The individual motion of subparts is left to evolve and be detected independently, so that each subpart may be solved individually, while the full body is assembled by inference over the graphical model.

Let us denote with X_μ^k ($\mu \in [1, n]$) the state vector associated to limb μ at time k ($k \in [0, K]$), and with Y_μ^k the corresponding observation. We introduce the following model parameters of each limb μ:

- the interaction potentials (or link factors) $\Psi_{\mu,\nu}(X_\mu^k, X_\nu^k)$, which measure the likelihood between two connected body parts μ and ν;
- the observations probabilities $\Phi_\mu(X_\mu^k, Y_\mu^k)$, which measure the likelihood between the state vector and the corresponding observation (image likelihood factors);
- and the time coherence factors $T(X_\mu^k, X_\mu^{k-1})$, which determine the likelihood for the same limb between two consecutive times.

Then the joint probability at time k can be written as a product of independent factors [35]:

$$P(X|Y) = \prod_{k=0}^{K} \Phi(X^k, Y^k)\Psi(X^k) \prod_{k=1}^{K} T(X^k, X^{k-1}), \quad (1)$$

where

$$\Phi(X^k, Y^k) = \prod_{\mu=1}^{n} \Phi_\mu(X_\mu^k, Y_\mu^k), \quad (2)$$

$$\Psi(X^k) = \prod_{(\mu,\nu) \in S} \Psi_{\mu,\nu}(X_\mu^k, X_\nu^k), \quad (3)$$

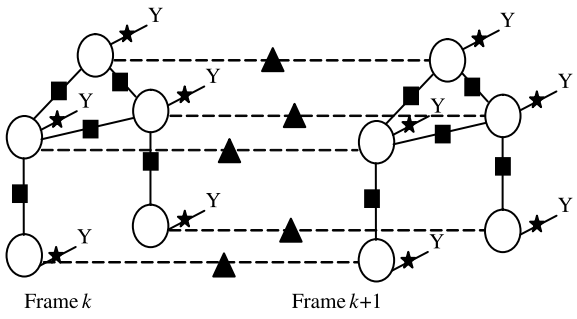

Fig. 1. Factor graph for a five part articulated structure. Circles represent state nodes, squares link factors, triangles time coherence factors and stars image likelihood factors.

$$T(X^k, X^{k-1}) = \prod_{\mu=1}^{n} T(X_\mu^k, X_\mu^{k-1}), \quad (4)$$

and S is the set of all links between connected body parts.

In our application of upper body human tracking, the graph is composed of nine nodes as we consider the motion of 9 limbs including the head, two clavicles, two arms, two forearms and two hands (Fig. 2). At each node is associated a five-dimensional space vector $(x, y, z, \varphi, \theta)$ excepting for the head represented by a four-dimensional one (x, y, z, θ). A 3D point and two angles are enough to localize a limb, such as an arm, modelled by a cylinder because the rotation of the limb around its main axis is not considered. For the head, as we suppose that it is faced to the camera, a 3D point and a single angle are employed. The first step of a tracking iteration consists to track each limb with a particle filter.

5. Particle filtering

The classical filtering problem consists in estimating an unknown signal from observed measurements. In computer vision the observations are image sequences, and the discrete time steps are given by the frame rate. Particle filters approximate conditional densities as a collection of weighted point samples. These approximations are stochastically updated by using Monte Carlo methods on the set of weighted point samples.

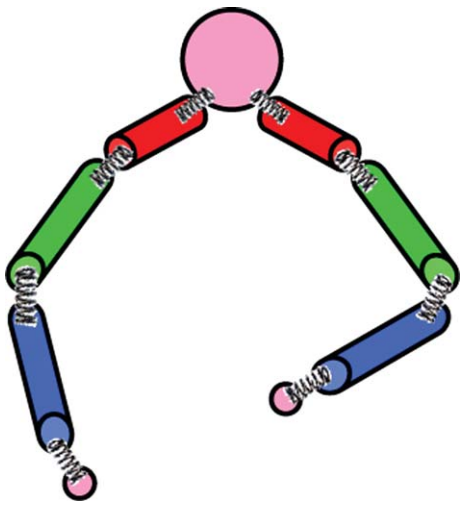

Fig. 2. Soft articulated human model.

Essentially, tracking with a classical particle filter works by performing three steps: (*i*) re-sampling from the weighted particle set obtained at the previous iteration, (*ii*) predicting stochastic movement and dispersion of particles, (*iii*) measuring and consequently updating the particle set. The main method drawback is that the number of particles needed to approximate the conditional densities grows exponentially as dimensionality increases. So, for a high dimensional problem such as human upper body tracking, the complexity becomes substantial.

The marginal probability of each limb is represented by a sum of N weighted particles:

$$P_{\mu}^{k}(X_{\mu}) \propto \sum_{i=1}^{N} \omega_{\mu,i}^{k} \delta(X_{\mu}^{k} - X_{\mu,i}^{k}), \tag{5}$$

where the weights are normalized so that $\sum_{i=1}^{N} \omega_{\mu,i}^{k} = 1$. Each particle $X_{\mu,i}^{k}$ represents a hypothetical position of the limb μ with a corresponding likelihood $\omega_{\mu,i}^{k}$ at time k. Then, the marginal probability density function is obtained recursively in a *prediction* and an *update* stage [26]. The stability and robustness of particle filters can often be improved by various methods. Among them, a particle-based stochastic search algorithm, called Annealed Particle Filtering (APF) [28], was developed. It uses a continuation principle, based on annealing, to introduce gradually the influence of narrow peaks in the conditional density function.

In each time-step, a multi-layered search (starting from layer $m = M$ to layer 1) is conducted. A smoothing of the weighting functions $\omega_{\mu,i}^{k,m} = (\omega_{\mu,i}^{k})^{\beta_m}$ is achieved by a set of values $\beta_M < \beta_{M-1} < \ldots < \beta_1$, where $\omega_{\mu,i}^{k}$ is the original weighting function.

A large β_m produces a peaked weighting function ω_m resulting in a high rate of annealing. Small values of β_m will have the opposite effect. At the same time, the amount of diffusion added to each successive annealing layer decreases according to $[P(X_{\mu,i}^{k} / X_{\mu,i}^{k-1})]^{\alpha_m}$, where the series α_m is such as $\alpha_M < \alpha_{M-1} < \ldots < \alpha_1$.

Method efficiency for recovering full articulated body motion depends on the choice of the tracking parameters α_m, β_m and on the particles number N. The purpose of the annealing filter is meanly to gain robustness and at the same time to reduce the particles number N.

At the end of an annealing iteration, the marginal probability of each limb is represented as a sum of weighted samples as for the standard particle filter. During this first step, the prediction and weighting are performed independently for each limb. Then, before considering the next iterations of annealing or the next frame, the final marginal probability is re-estimated with the belief propagation algorithm on the factor graph to take into account geometrical constraints between limbs.

6. Belief propagation

In the BP algorithm, human body and relationships between body parts are represented by a graphical model, called factor graph. The factor graph is composed of a variable node for each variable X_{μ} (body part) and a factor node on each edge connecting variable node μ to variable node ν. The algorithm recovers the pose of each body part by considering the relationships between every two adjacent body parts. Additional edges have also been introduced to add non-collision constraints between some limbs (this avoids for example that the two hands remain together).

After designing interaction functions and observation functions, BP is used to search for body parts' belief by iteratively updating messages sent from a node to another one. Messages are propagated for all nodes at each frame for a variable number N_{BP} of iterations, and then propagated only once from a frame to the following one [35].

In practice, each message of the BP algorithm is approximated by a set of N weighted samples. The message $m_{\nu \to \mu,i}$ sent from the node ν to the particle i of node μ is written:

$$m_{\nu \to \mu,i}^{k} = \sum_{j} \Psi_{\mu,\nu}(X_{\mu,i}^{k}, X_{\nu,j}^{k}) \Phi_{\nu,j}^{k}(X_{\nu,j}^{k}) \prod_{\nu' \in S(\nu) \setminus \mu} m_{\nu' \to \nu,j}^{k}, \tag{6}$$

where $S(\nu) \setminus \mu$ is the set of the neighbours of node ν except μ, and $\Phi_{\nu,j}^{k}(X_{\nu,j}^{k})$ is the local likelihood of the sample j of node ν. The belief at the node μ is then estimated by:

$$P_{\mu}^{k}(X_{\mu}) \sim \sum_{i=1}^{N} \hat{\omega}_{\mu,i}^{k} \delta(X_{\mu}^{k} - X_{\mu,i}^{k}), \tag{7}$$

where

$$\hat{\omega}_{\mu,i}^{k} = \Phi_{\mu,i}^{k}(X_{\mu,i}^{k}) \prod_{\nu \in S(\mu)} m_{\nu \to \mu,i}^{k}. \tag{8}$$

The resulting combined APF-BP algorithm is described in Table 1.

As we want to be able to track unconstrained human motions, any specific movement model could be used and the prediction is achieved according to Gaussian distributed diffusion. The amount of diffusion for each joint angle j is dependent on the image frames per second (fps). In our experiments σ_j ranges from 10 degrees to about 40 degrees to take in account the different extension field of each limb. For example, variations of rotation angles are smaller for clavicles than for hands.

7. Application to body tracking

The images features used to estimate the image compatibility factors $\Phi^k_{\mu,i}(X^k_{\mu,i}, Y^k_\mu)$ have to be strongly discriminant to allow limbs detection and tracking. For each limb, the 3D model corresponding to the particle i of the state vector is projected on the image plane and the likelihood between this projection and image observations is estimated. This measure is based on oriented edge matching, motion energy and background subtraction. For the head and the two hands, factors are also based on a skin colour probability map [22].

In tracking movements, the background subtraction is useful to focus limbs detection around the body contours avoiding possible mistakes caused by the environment; to surely keep in consideration body contours it is not necessary to have a really precise body detection but rather a fast one.

At each frame, a background subtraction is made by thresholding the absolute value difference between the background image and the actual one for each pixel and comparing each value with the average one (Fig. 3).

The resulting image is used to enhance the motion energy and the skin color probability map.

To measure limb movements so calculate their likelihood two scores are calculated: a gradient score S^μ_R and a movement energy score S^μ_M.

The gradient score is based on contours and is equal to

$$S^\mu_R = \sum_{r \in p(\mu)} f(||\vec{R}(r)||) G[\vartheta_\mu(r) - dir(\vec{R}(r))]. \tag{9}$$

where:

- $p(\mu)$ is the set of points belonging to the projection in the image of the considered limb μ.
- $f(||\vec{R}(r)||)$ is a function of the gradient norm $||\vec{R}(r)||$ at pixel r, that penalizes the highest values of $||\vec{R}(r)||$:

$$f(||\vec{R}(r)||) = \frac{1}{\lambda} ||\vec{R}(r)|| \tanh \frac{\lambda}{||\vec{R}(r)||} \tag{10}$$

where λ is the mean of the gradients norm of each frame.

- $G[\vartheta_\mu(r) - dir(\vec{R}(r))]$ is the Gaussian of the difference between the orientation of each point of the considered limb $\vartheta_\mu(r)$ and the orientation of the gradient $dir(\vec{R}(r))$.

The movement energy score S^μ_M evaluate the probability of having the considered limb in the location in which the movement has been detected.

A first step of motion detection is achieved by thresholding the absolute difference of two successive images, leading to a binary map. Then, a value is given

Table 1
Algorithm resulting from the combination of APF and BP methods

- Initialisation: $X^0_{\mu,i} = X^0_\mu$
- For each time k
  - For $m = M, \ldots, 1$ of the APF
    - Evolution: A new set of particles is drawn according to the system dynamic $[P(X^k_{\mu,i}/X^{k-1}_{\mu,i})]^{\alpha_m}$
    - Weight: Compute the weight $\Phi^{k,m}_{\mu,i} = [P(X^k_{\mu,i}, Y^k_\mu)]^{\beta_m}$
    - N_{BP} iterations of BP $m^{k,m}_{\nu\to\mu,i} = \sum_j \Psi_{\mu,\nu}(X^k_{\mu,i}, X^k_{\nu,j}) \Phi^{k,m}_{\nu,j}(X^k_{\nu,j}) \prod_{\nu' \in S(\nu)\setminus\mu} m^{k,m}_{\nu'\to\nu,j}$ and re-estimation of $\omega^{k,m}_{\mu,i}$ with $\omega^{k,m}_{\mu,i} \propto \Phi^{k,m}_{\mu,i} \prod_{\nu \in S(\mu)} m^{k,m}_{\nu\to\mu,i}$
    - Resample the set of particles according to $\omega^{k,m}_{\mu,i}$
  - End for
- End for

Fig. 3. Environment, frame image, background substraction.

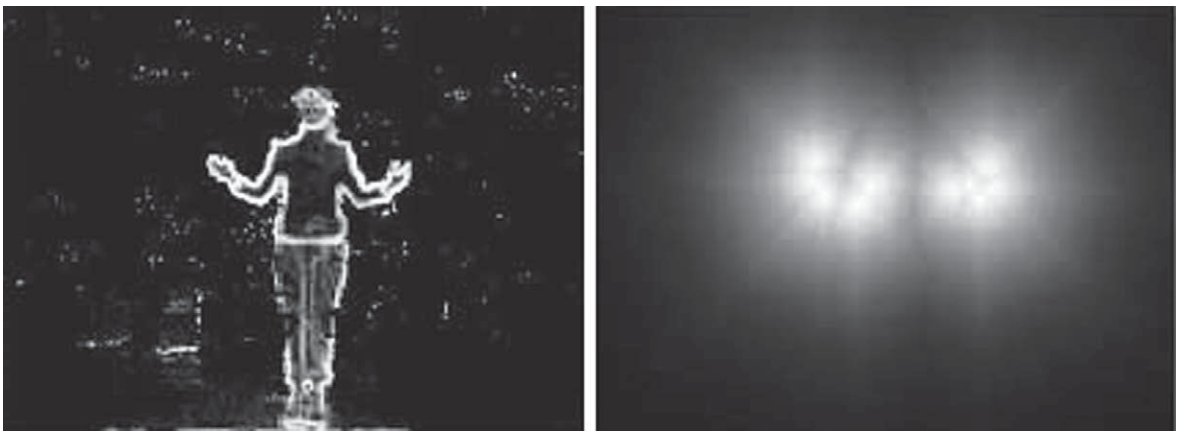

Fig. 4. Map of gradient score and a movement energy score.

at each pixel according to its distance d to motion pixels leading to a movement energy image:

$$I_M(r) = \exp\left(-\frac{d}{\sigma_d}\right). \tag{11}$$

where σ_d is a parameter depending on image resolution. On the so obtained image (Fig. 4), huge values are given to pixels near to moving areas, while low values are set to pixels in static areas. The movement energy score S_M^{μ} is then obtained with:

$$S_M^{\mu} = 1 + \sum_{r \in p(\mu)} P(I_M(r) = 1). \tag{12}$$

where

- as before, the summation takes into account all pixels belonging to the projection of the limb;
- 1 is added to neutralize results on static limbs.

Moreover, to recognize the position of head and hands in each frame a map skin is calculated (Fig. 5) by learning a skin area in the subspace C_bC_r from the YC_bC_r colorspace.

Link factors $\Psi_{\mu,\nu}(X^k_{\mu,i}, X^k_{\nu,j})$ representing the likelihood between two connected body parts are expressed as Gaussians of the distance between the reference points for the articulation between two successive limbs. For hands and head non collision constraints are added to avoid occlusions problems.

Fig. 5. Map of skin.

Fig. 6. Human posture estimations for some usual gestures. These results are obtained with 100 particles per limb, 3 iterations of simulated annealing and 1 iteration of BP.

8. Experiments

This algorithm was tested on sequences of images with a resolution of 360×288 pixels, acquired with a frequency of 15 frames per second with a PC HP, processor Intel® Core™2 Duo, CPU 2.00 GHz, 1.99GB of RAM. In this paper we report the results obtained with two video sequences in which a person makes different everyday movements, like pointing, beckoning or waiting.

Moreover the videos were made in different environments (illumination, background, clothes) and movements are made in different order and fashion. The performance of our algorithm was evaluated by varying relevant model parameters, that are, the number of particles N, the layer number M of APF and the number N_{BP} of cycles of BP, with the proposal of reducing tracking time without losing robustness. Some sequence frames with the corresponding tracking results are shown in Figs 6 and 7. These results are obtained with 100 particles per limb, 3 iterations of simulated annealing and 1iteration of BP. It was very difficult to quantitatively evaluate the results. Furthermore, no ground truth was available for the

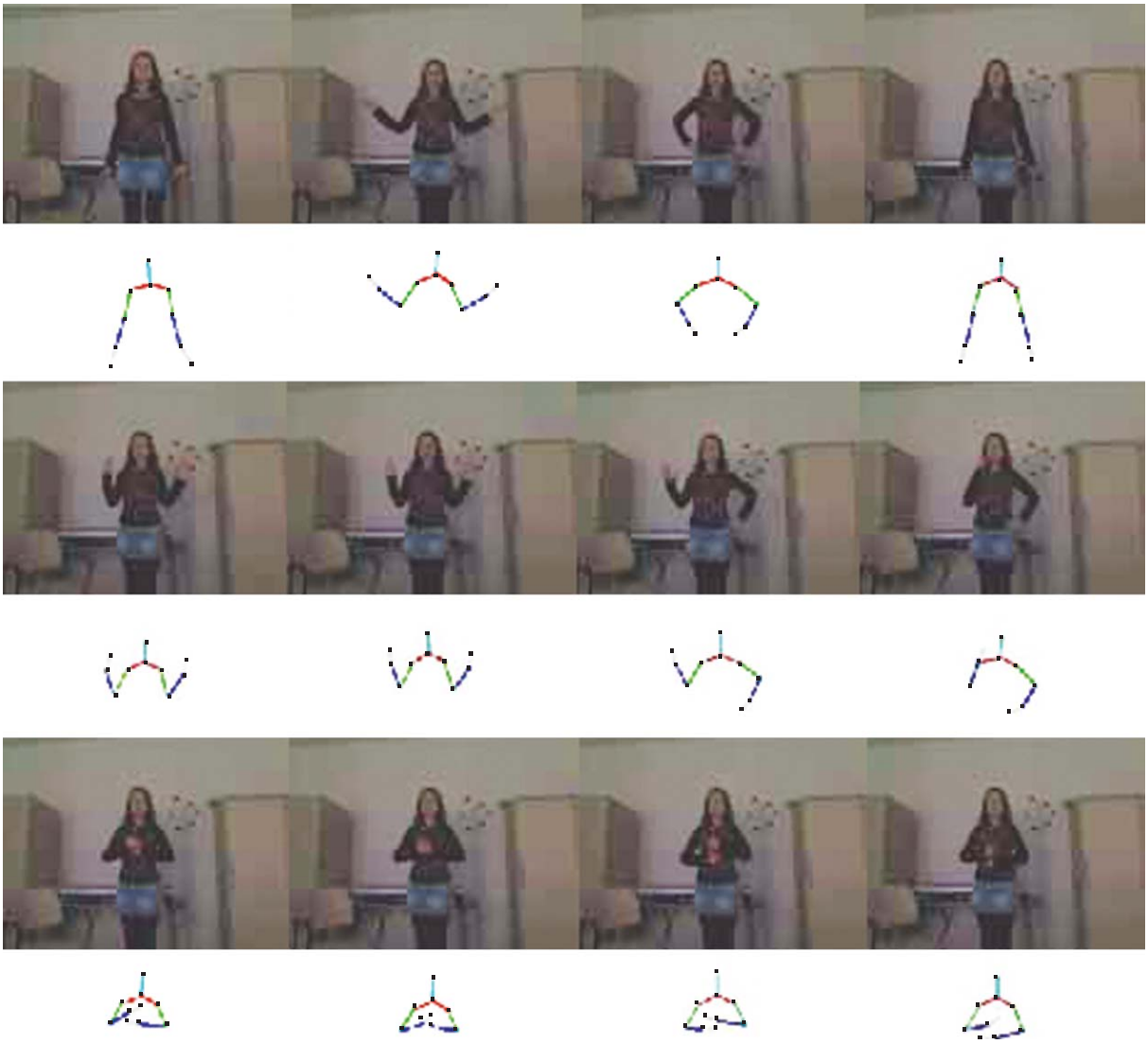

Fig. 7. Human posture estimations for some usual gestures. Parameters as in Fig. 6.

collected data. Hence, for each image of the sequence, we decided to manually click on the body joints in order to obtain the coordinates of interest. Although these data are not completely accurate (for some body positions it is quite difficult to properly locate the coordinates of a joint), they constituted our 2D ground truth. This allowed to evaluate in a quantitative way the accuracy of algorithm results: we computed the distance between the 2D point of the ground truth and the projection on the image of the 3D point representing the limb articulation point. The articulation point was obtained making the mean between the values estimated by the algorithm for two consecutive limbs.

These distances represent the tracking error for each articulation.

In order to quantify the advantages of the combination of simulated annealing and BP, we have performed several tests with different sets of parameters. We considered as results the mean distance errors for all the limbs (all these results have been obtained by averaging 5 realisations of tracking because of stochastic variables in the algorithm). Moreover, it is worth to precise that the algorithm has not been yet optimised. So only the order of magnitude between the different times has been considered. We tested the algorithm making all the possible combination about the number of BP (from 1 cycle to 10), the number of APF (one cycle or three cycles) and the particles number (100 or 300). In a first time we observe that just one BP cycle is enough to obtain a good cohesion between limbs. Moreover when several iterations of BP are performed to reach the convergence, results are damaged probably because human body doesn't exactly validate the articulated model. Nevertheless, it is important to note that without BP cycles, all limbs are disconnected and the tracking diverges quickly.

The best results are obtained with 300 particles and 3 iterations of annealing but computation time is considerable (around 3.5 s). To reduce it, one can either decrease the number of particles or the number of annealing iterations. Errors evaluation shows that it is better to preserve the annealing that allows to use fewer particles without tracking failure and with an acceptable computation time. In fact, in our algorithm 100 particles per limb with 3 annealing loops leads to accurate results with a reasonable error and a small standard deviation on this error. As shown in Table 2, the use of 3 APF layers allows to utilize less particles (just 100) keeping robustness with a lower computation time of 0.75 s compared to a time of 1.43 s when just a layer but more particles are used (results quality is then quite similar). Moreover, it appears that results on shoulder and wrist are more reliable: in fact, shoulders move more lightly than other limbs and wrist tracking is easier thanks to hands skin color.

The outputs are correlated except for some parts because of the random characteristics that affect in partially and locally the solution space.

In Table 3 we show that same results are obtained for the two videos (it is worth to remember that little differences are normal considering the randomized procedures used within the algorithm).

Table 2
Error in pixel for different sets of parameters

Method	Error	Shoulder	Elbow	Wrist	Global error	Computation time
$N=100\ M=3$	Mean	3.3	5.2	4.7	4.4	0.35
$N_{BP}=1$	Standard deviation	1.9	4.3	2.5	2.9	
$N=300\ M=3$	Mean	3.2	5.2	4.3	4.1	2.4
$N_{BP}=1$	Standard deviation	2.0	4.5	2.5	2.6	
$N=300\ M=1$	Mean	3.2	5.4	5.4	4.7	1.43
$N_{BP}=1$	Standard deviation	1.9	4.0	2.7	2.9	

Table 3
Error in pixel using $N=100$, $M=3$, $N_{BP}=1$

Video	Error	Shoulder	Elbow	Wrist	Global error	Computation time
1	Mean	3.3	5.2	4.7	4.4	0.35
	Standard deviation	1.9	4.3	2.5	2.9	
2	Mean	5.84	4.68	4.56	5.70	0.36
	Standard deviation	3.35	5.82	4.2	4.05	

In Fig. 8 we have the right and the left elbows mean positions error (in pixel) versus time. We have some maxima for the error rate. For the right elbow, for example, one maximum is reached at the frame 1700. This corresponds to a specific configuration of occlusion and mechanical singularity. As one can see in

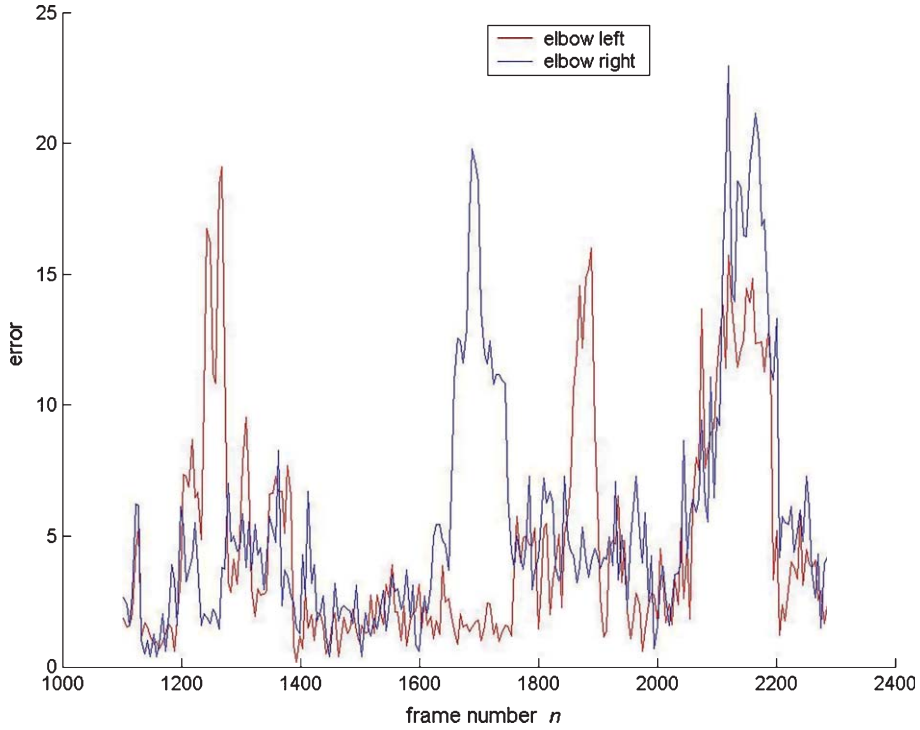

Fig. 8. Elbow left and right mean of 5 Trials.

Fig. 9. Configuration with an occlusion and a singularity.

Fig. 9, the arm, the forearm and the hand limbs are aligned and are perpendicular to the image plane in the first picture, i.e., the projection of each of them is a point. The figure shows that when there is a configuration with occlusion the algorithm is capable to correct itself just after few frames.

9. Conclusion and future work

In this paper we proposed a new approach for upper body pose estimation using an algorithm that combines annealed particle filter and belief propagation methods. This algorithm shows that the use of this combination is effective to reduce the number of particles and, as a consequence, reduces computation time without losing robustness: (i) the use of simulated annealing decreases the number of particles used to track each limb thanks to the introduction of narrow peaks in the fitness function; (ii) BP method assures spatial coherence and an independent tracking for each limb decreasing, moreover, the dimensionality of state space.

This work is part of a larger project dealing with natural gestures-based communication. The ongoing work and the extension of presented here is dealing with both improvement of the tracking process and the semantic interpretation of the arms sequences.

The first point is dealing with the introduction of a geometrical-based regularization in order to improve the tracking and to learn gradually a color model of the target. The other possible improvement concerns the transfer of the algorithm on a highly parallel architecture. Indeed, our formulation is well adapted to such architecture and execution times should be reduced consequently to fit real-time constraints.

The second point is more semantic oriented. It is dealing with the classification and the recognition of the upper body sequences in terms of commands and intents and we are working both on computational and social aspects.

References

[1] K. Dautenhahn, S. Woods, C. Kaouri, M.L. Walters, K.L. Koay and I. Werry, What is a robot companion – friend, assistant or butler? *In Proc IEEE IROS* (2005), 1488–1493.

[2] M. Goodrich and A.C. Schultz, *Foundations and Trends® in Human–Computer Interaction*, vol. 1, no. 3, 2007, pp. 203–275.

[3] A. Kendon, Historical observations on the relationship between research on sign languages and language origins theory. in: *The Study of Signed Languages: Essays in Honor of William C. Stokoe*, David Armstrong, Michael A. Karchmer and John Vickery Van Kleve, eds., Gallaudet University Press, Washginton, DC, 2002, pp. 35–52.

[4] R.M. Krauss, Y. Chen and P. Chawla, Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? in: *Advances in Experimental Social Psychology*, M. Zanna, ed., Academic Press, San Diego, CA, 1996, pp. 389–450.

[5] D. McNeill and S.D. Duncan, Growth points in thinking-for-speaking. in: *Language and Gesture*, D. McNeill, ed., Cambridge University Press, Cambridge, 2000, pp. 141–161.

[6] Thomas B. Moeslund, Adrian Hilton and Volker Kruger, A survey of advances in vision-based human motion capture and analysis, *Computer Vision and Image Understanding* **104**(23) (2006), 90–126.

[7] P.F. Felzenszwalb and D.P. Huttenlocher, Efficient matching of pictorial structures, in: *Computer Vision and Pattern Recognition*, Hilton Head Island, South Carolina, USA, June 13–15, 2000.

[8] D. Ramanan, D.A. Forsyth and A. Zisserman, Strike a pose: Tracking people by finding stylized poses, in: *Computer Vision and Pattern Recognition*, San Diego, California, USA, 2005.

[9] R. Ronfard, C. Schmid and B. Triggs, Learning to parse pictures of people, *European Conference on Computer Vision*, Copenhagen, Denmark, 2002.

[10] M.W. Lee and R. Nevatia, Body part detection for human pose estimation and tracking, *Workshop on Motion and Video Computing*, 2007.

[11] P.F. Felzenszwalb and D.P. Huttenlocher, Efficient matching of pictorial structures, in: *Computer Vision and Pattern Recognition*, Hilton Head Island, South Carolina, USA, June 13–15, 2000.

[12] X. Ren, A.C. Berg and J. Malik, Recovering human body configurations using pairwise constraints between parts, *International Conference on Computer Vision*, Beijing, China, 2005.

[13] G. Hua, M.-H. Yang and Y. Wu, Learning to estimate human pose with data driven belief propagation, *Computer Vision and Pattern Recognition*, San Diego, California, USA, 2005

[14] R. Navaratnam, A. Thayananthan, P.H.S. Torr and R. Cipolla, Hierarchical part-based human body pose estimation, BMVC05.

[15] M. Brand, Shadow puppetry, in: *Proceedings of the International Conference on Computer Vision*, 1999, pp. 1237–1244.

[16] R. Rosales, M. Siddiqui, J. Alon and S. Sclaroff, Estimating 3D Body pose using uncalibrated cameras, *Computer Vision and Pattern Recognition*, Kauai Marriott, Hawaii, 2001

[17] N.R. Howe, Silhouette Lookup for Automatic Pose Tracking, *Workshop on Articulated and Non-Rigid Motion*, Washington, DC, USA, 2004.

[18] J. Tenenbaum, V. de Silva and J. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science* **290** (2000), 2319–2323.

[19] L. Gond, P. Sayd, T. Chateau and M. Dhome, A 3D shape descriptor for human pose recovery, *Articulated Motion and Deformable Objects*, AMDO, Spain 2008.

[20] A. Agarwal and B. Triggs, Recovering 3D human pose from monocular images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(1) (2006), 44–58.

[21] L. Sigal, M. Isard, B.H. Sigelman and M.J. Black, Attractive people: Assembling loose – limbed models using nonparametric belief propagation, *Neural information processing systems* (2003).

[22] P. Noriega and O. Bernier, Multicues 3D monocular upper body tracking using constrained belief propagation, *British Machine Vision Conf.*, Warwick, UK, 2007.

[23] R. Plänkers and P. Fua, Articulated soft objects for multiview shape and motion capture, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(9) (2003).

[24] S. Wachter and H.-H. Nagel, Tracking persons in monocular image sequences, *Computer Vision and Image Understanding* **74**(3) (1999), 174–192.

[25] H. Sidenbladh and M. Black, Learning the statistics of people in images and video, *International Journal of Computer Vision* **54**(1–3) (2003), 183–209.

[26] M. Isard and A. Blake, Conditional density propagation for visual tracking. in: *International Journal of Computer Vision*, vol. 29, no. 1, 1998, pp. 5–28.

[27] J. MacCormick and M. Isard, Partitioned sampling, articulated objects, and interface-quality hand tracking, *European Conference on Computer Vision*, Dublin, Ireland, 2000.

[28] J. Deutscher, A. Blake and I. Reid, Articulated body motion capture by annealed particle filtering, *IEEE Conference on Computer Vision and Pattern Recognition*, 2000.

[29] C. Sminchisescu and B. Triggs, Estimating articulated human motion with covariance scaled sampling, *International Journal of Robotics Research* **22**(6) (2003), 371–393.

[30] M.W. Lee, I. Cohen, Proposal maps driven MCMC for estimating human body pose in static images, *Computer Vision and Pattern Recognition*, Washington, DC, USA, 2004.

[31] T.B. Moeslund, C.B. Madsen and E. Granum, Modelling the 3D pose of a human arm and the shoulder complex utilising only two parameters, *International Journal on Integrated Computer-Aided Engineering* **12**(2) (2005), 159–175.

[32] Jiang Gao and Jianbo Shi, Multiple frame motion inference using belief propagation, *Automatic Face and Gesture Recognition*, 2004.

[33] Tony X. Han, Huazhong Ning, and Thomas S. Huang, E_cient nonparametric belief propagation with application to articulated body tracking. *Computer Vision and Pattern Recognition*, vol. 1, 2006.

[34] Michael Isard. Pampas: Real-valued graphical models for computer vision, in: *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, vol. 1, 2003.

[35] O. Bernier, P. Cheungmonchan, A. Bouguet, Fast nonparametric belief propagation for real-time stereo articulated body tracking, *Computer Vision and Image Understanding*, 2008.

[36] J.S. Yedidia, W.T. Freeman and Y. Weiss, Constructing free energy approximations and generalized belief propagation algorithms. Technical Report 2004-040, MERL, May 2004.

[37] E.B. Sudderth, A.T. Ihler, W.T. Freeman, and A.S. Willsky, Nonparametric belief propagation, *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2003, pp. 605–612.

[38] M. Isard, PAMPAS: Real–valued graphical models for computer vision, *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2003, pp. 613–620.

[39] C. Shen, A. van den Hengel, A. Dick and M.J. Brooks, 2D articulated tracking with dynamic bayesian networks, *Proceedings of the Fourth International Conference on Computer and Information Technology*, 2004.

[40] F.R. Kschischang, B.J. Frey and H.A. Loeliger, Factor graphs and the sum-product algorithm, *IEEE Trans. on Information Theory* **47**(9)(2001), 498–519.

[41] B.R. Duffy, Anthropomorphism and the social robot, *Robotics and Autonomous Systems* **42** (2003), 177–190.