3D Human Tracking in a Top View Using Depth Information Recorded by the Xtion Pro-Live Camera

Cyrille Migniot, Fakhr-Eddine Ababsa

3D Human Tracking in a Top View Using Depth Information Recorded by the Xtion Pro-Live Camera

Cyrille Migniot and Fakhreddine Ababsa

IBISC - team IRA2 - University of Evry val d'Essonne, France,
{Cyrille.Migniot,Fakhr-Eddine.Ababsa}@ibisc.fr

Abstract. This paper addresses the problem of the tracking of 3D human body pose from depth image sequences given by a Xtion Pro-Live camera. Human body poses could be estimated through model fitting using dense correspondences between depth data and an articulated human model. Although, most of the time for the video surveillance, the camera is placed above the persons, all the tracking methods use the front view. Indeed the human shape is more discriminative in this view. We propose a new model to be fitted to the top view in a particle filter framework for a real-time markerless tracking. The model is composed of two parts: a 2D model providing the human localization and a 3D model providing its pose. There are few wrong estimations and they are efficiently detected by a confidence measure. ...

1 Introduction

In recent years, there is a body of research on the problem of human parts detection, pose estimation and tracking from 3D data. The image is fitted to an articulated model that embodies the possible human movements. The great majority of the methods in the literature use a model adapted to a front view of the person because the shape of a person is much more discriminative on this orientation. Furthermore the color of the skin and the elements of the face is seldom available in a top view. Nevertheless, in the application of the video-surveillance, the camera is frequently installed above the persons. The tracking on a top view need to use depth feature. One of the more popular devices used to provide it is kinect, which has sensors that capture both rgb and depth data. Here, we simulate the installation of a Xtion Pro-Live cameras (depth-color camera launched by Assus in 2013) in the ceiling of a supermarket. The goal is to analyze the behaviors of the customers during their buying acts within the shelves. The first step is tracking the pose of a customer. Then it will be recognized by an other treatment.

The main challenge in articulated body motion tracking is the large number of degrees of freedom to be recovered. For non-linear or non-Gaussian problems, the particle filter algorithm [1] has became very popular. Based on the Monte-Carlo simulation, it provides a suitable framework for state estimation in a non-linear,

non-Gaussian system. At moment k, let x_k be the state of the model and y_k be the observation. Particle filter recursively approximates the posterior probability density $p(x_k|y_k)$ of the current state x_k evaluating the observation likelihood based on a weighted particle sample set $\{x_k^i, \omega_k^i\}$. Each of the N particles x_k^i corresponds to a random state propagated by the dynamic model of the system and weighted by ω_k^i. There are 4 basic steps:

- **resampling:** N particles $\{x_k'^i, \frac{1}{N}\} \sim p(x_k|y_k)$ from sample $\{x_k^i, \omega_k^i\}$ are re-sampled. Particles are selected by their weight: large weight particles are duplicated while low weight particles are deleted.
- **propagation:** particles are propagated using the dynamic model of the system $p(x_{k+1}|x_k)$ to obtain
  $\{x_{k+1}^i, \frac{1}{N}\} \sim p(x_{k+1}|y_k)$.
- **weighting:** particles are weighted by a likelihood function related to the correspondence from the model to the new observation. The new weights ω_{k+1}^i are normalized so that : $\sum_{i=1}^{N} \omega_{k+1}^i = 1$. It provides the new sample $\{x_{k+1}^i, \omega_{k+1}^i\} \sim p(x_{k+1}|y_{k+1})$.
- **estimation:** the new pose is approximated by:

$$x_{k+1} = \sum_{i=1}^{N} \omega_{k+1}^i x_{k+1}^i \tag{1}$$

The articulation of the person is often taken into consideration and modeled by a skeleton whose the rigid segments represent the body parts. To represent the volume occupied by the person, the skeleton comprises of a set of appropriately assembled geometric primitives [2–4]. The most usefull features chosen to describe the human class are the skin color [5], contour [6] and results of classifiers [7].

To reduce the computing time, Gonzales [5] splits its tracking to each sub-part of the body, Yang [8] simplifies the likelihood function with a hierarchical particle filter and Deutscher [2] reduces the required number of particles with an annealed particle filtering.

The movements of the skeleton can be constrainted by interaction with objects in the environment [9, 10]. All poses of the skeleton are not possibles in practice. For example the head can not rotate over 360°. The sampling can be constrained by a projection on the feasible configuration space [3].

The human tracking in the top view is a seldom explored issue that has numerous applications in the video surveillance. Heath [11] estimates the 3D trajectories of salient feature points (primarily at the shoulders level) that he uses as the observation for the particle filtering. Canton-Ferrer [12] defines the exclusion zone for the blocking by an ellipsoide. For the tracking, he separates 3D blobs and used a particle filter where each particle represents a voxel of the blob. That estimates the centroid of the blob that model the person. These methods realize the tracking of the position of the person and not the gesture. We propose a method that well-follows the pose of the arms that defines the gesture of the

person.

In this paper, we describe a human gesture tracking from a top view by particle filtering. Relevant information is given by the depth provided by a Xtion Pro-Live camera. Rincón [13] realizes a fisrt tracking to estimate the global location of the person and a second to recover the relative pose of the limbs. Similarly, but for the upper part of the body, we broke it the model up a 2D model representing the head and the shoulders and an 3D model representing the whole body (figure 1). The first one is relatively easy to obtain from the depth array. For the second one, the computational cost is reduced by constraining the space of possible poses with prior information given by the head and shoulders location.

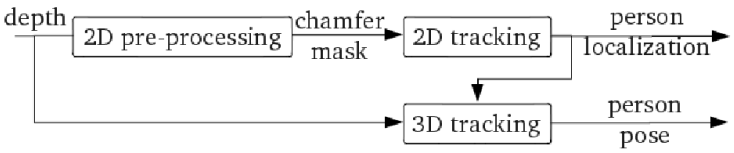

Fig. 1. Overview of the method: the 2D tracking estimates the location of the person while the 3D tracking estimates the pose.

Our main contributions are first using a specially top view adapted model, secondly taking advantage of the depth signal to define likelihood function, thirdly decomposing the model in two parts so as to reduce the complexity of filtering and use simultaneously 2D and 3D models and finally introducing a confidence measure so as to detect the wrong pose estimations.

2 Particle Filter Implementation

To acquire the depth image, the Xtion Pro-live camera produced by Asus is installed at 2,9 m of the ground. This depth image provides a set of points in the 3D space (in grey in the figure 4) that represent the visible part of the surface of the person on a top view. Moreover, using 3D data allows introducing spatial constraints: a same object has various sizes in the 2D space according to its distance to the camera while it has always the same size in the 3D space. Two trackers using particle filter are presented in the following. For the initialisation step, the shoulders location is given by a detection process while the pose of the arms is fixed to be parallel to the body. A small number of frames are sufficient to find the right arm pose. We use a simple constant-vitesse dynamic model with a gaussian dispersion for the propagation.

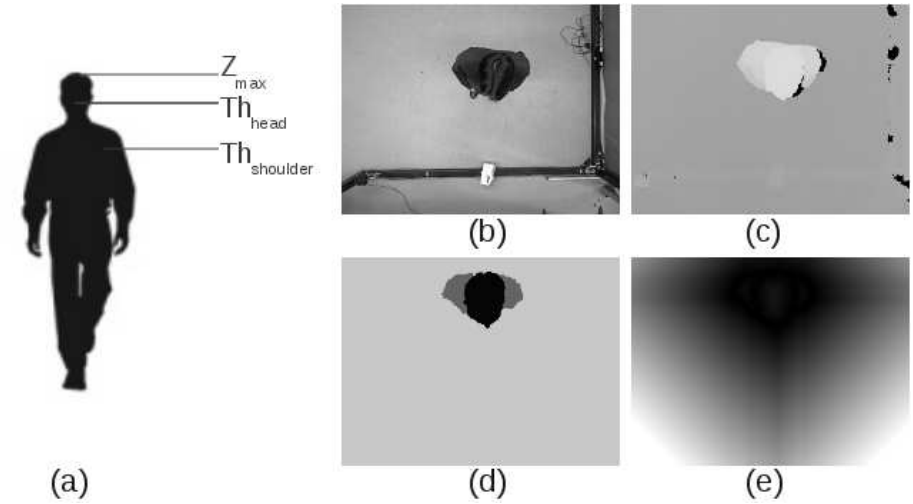

Fig. 2. The camera records simultaneously the color (b) and the depth (c) images. The depth image is thresholded (d) in the head and the shoulders levels (a). The ellipses of the model are fitted to the chamfer distance of this map (e).

2.1 The head-shoulders model

As Micilotta [14], we use the Ω-like shape produced by the head and the shoulders. The top view of the head and the shoulders is modeled by two ellipses. Each ellipse defines 3 degrees of freedom.
The depth cue gives the distance of an element to the camera, thus it gives its distance to the ground. We threshold the depth image in two levels corresponding to the middle of the head and the end of the shoulders (figure 2(d)). The likelihood function is computed by the matching of the ellipses given by the particle with the chamfer distance of the thresholded depth image (figure 2(e)). The chamfer distance is robust to the person stoutness variation. Constraints are inserted in the propagation step in order to link the two ellipses.

2.2 The complete 3D model

Modelisation This model is hardly constrained by the previous one. The 2D position of the shoulders gives the location and the orientation of the person in the scene. The 3D model determines the arms movement. We make the hypothesis that an arm has 5 degrees of freedom: 3 for the shoulder and 2 for the elbow. The skeleton of the model is also build and defined by a 5-dimensions state vector for each arm. The state space is limited by biomechanical knowledge about human motion. To represent the volume, geometrical primitives are added. Arms and forearms are modeled by truncated cylinders, torso by an elliptic cylinder and finally the hands by rectangular planes (figure 3).

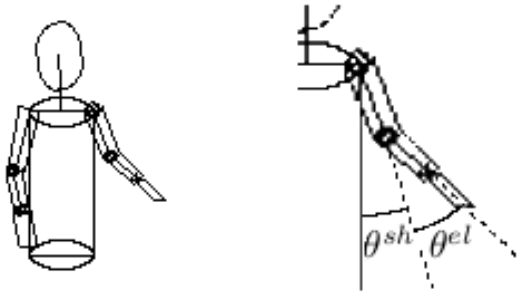

Fig. 3. The complete 3D model (left) is defined from the angles of the state vector (right).

Likelihood function As for the previous model, the depth cue is used. Nevertheless, not only the thresholded 2D representation is exploited. Indeed the depth variation is well-descriptive of arm. Thus a 3D chamfer distance is chosen. Let Δ be the pixels of the foreground of the depth image excluding the head and the shoulders detected previously and $\mathcal{M}$ be the 3D model given by a particle. The likelihood function related to particle i is defined by:

$$\omega_i = \underset{p \in \Delta}{average}(d_{3D}(p, \mathcal{M}^i)) \tag{2}$$

where d_{3D} is the shortest distance from a point to a 3D model.

3 Performances

We have simulated the behavior of customers in experimental conditions to evaluate the ability of our method to track a person. We have recorded two sequences $\mathcal{S}_1$ et $\mathcal{S}_2$ that are made of 450 frames (>1min) and 300 frames (≈43s) with various arm movements.

To qualitatively evaluate our results, we display in figure 4 the 3D model given by the tracking and projected on the color image and in the XZ and YZ planes. The pixels that can be seen in the depth image are represented in gray in the XZ and the YZ planes. We can also notice that the model is well-fitted to the person.
We use the depth cue as a ground truth to provide a quantitative evaluation. The pixels of the foreground of the depth image excluding the head and the shoulders (given by the 2D tracking) are shared between the two arms through the shoulders orientation. The evaluation measure we used is the average of the distance between the transposition in the 3D space of these pixels and the 3D model estimated by the tracking.
A high number of particles increases the accuracy of the tracking but it increases

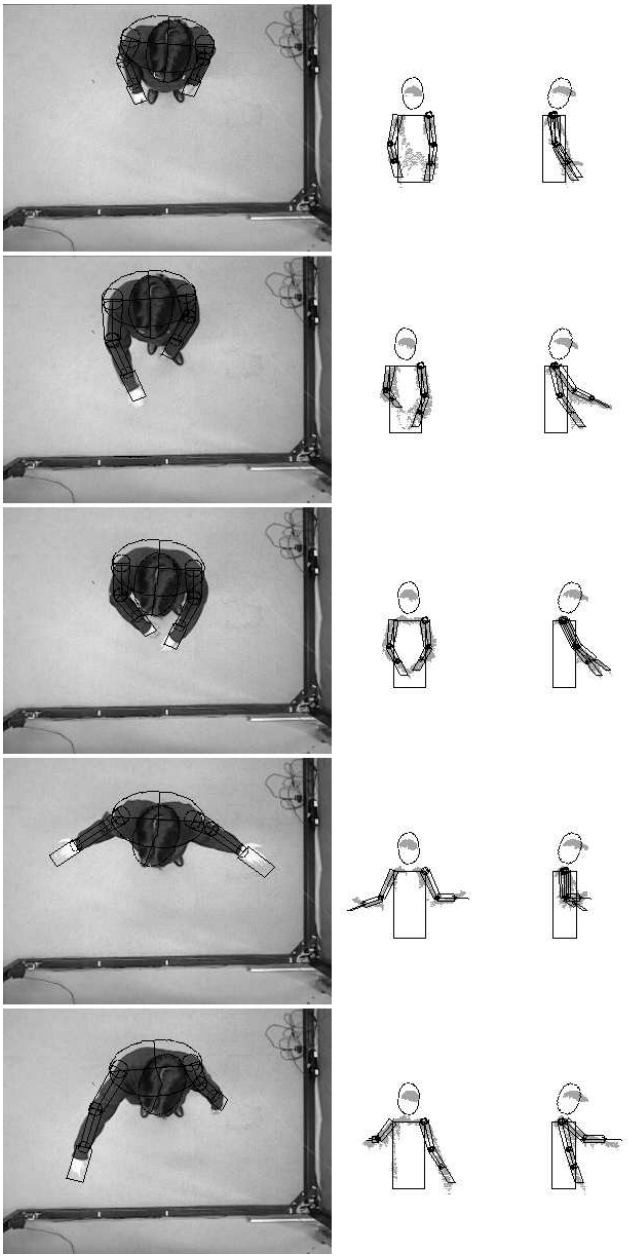

Fig. 4. On the left the model on the color image and on the right the model in the 3D space projected on XZ and YZ planes (the pixels in gray correspond to the points given by the depth image).

the time processing. A compromise may be found. The processing times are here obtained with a non-optimized C++ implementation running on a 3,1GHz processor. The figure 5 shows that there are no more meaningful improvement over 75 particles. Under this limit, it provides a real time processing. The tracking is significantly degraded under 25 particles. With 75 particles on 912 evaluations of the arm in the sequence, we obtain an evaluation measure with an average of 25mm and a standard deviation of 5mm.

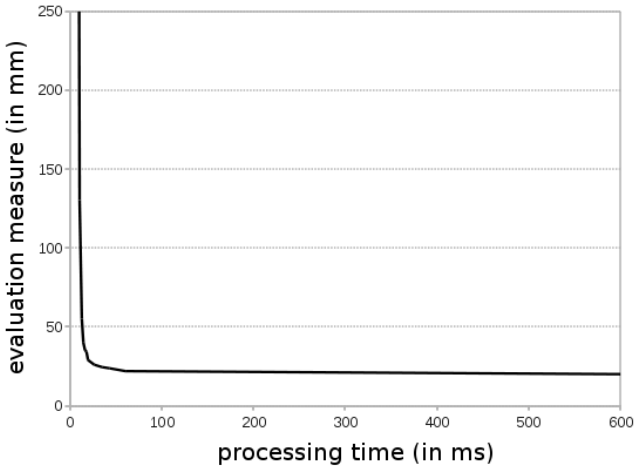

Fig. 5. Evolution of the tracking quality as a function of the processing time for the two sequences: when the number of particles increases, the accuracy of the tracking increases but the processing frequency decreases. Our process is real time.

Sometimes, some particular poses are difficult to estimate. It is the case when the fist is held high (figure 6). To detect these false estimations we introduce a confidence measure. The skeleton of the estimation is projected on the 2D plane of the depth image. The confidence measure is the average distance from the pixels of the projected arms to the foreground. It is equal to 0 when the model is well-fitted to the observation and up otherwise. We have computed this confidence measure on the sequences $\mathcal{S}_1$, $\mathcal{S}_2$ et $\mathcal{S}_3$ (figure 6). In case of wrong estimation, this criterium provides higher values that are easy to detect. The number of wrong estimations is small (1,52% for the three sequences) and they mostly persist for only one frame.
To evaluate the trajectories of the parts of the arm, we follow the 3D positions of the shoulders, the elbows and the wrists on a third sequence $\mathcal{S}_3(\approx 55\text{s})$ with large movement of the right arm. We use two ARTTRACK1 cameras and the software DTRACK to follow reflecting balls (figure 7(a)). The position of these captors are recorded simultaneously with the Xtion Pro-Live acquisition. We can see in figure 7(bcd) that the movement recorded by the captor is relatively closed

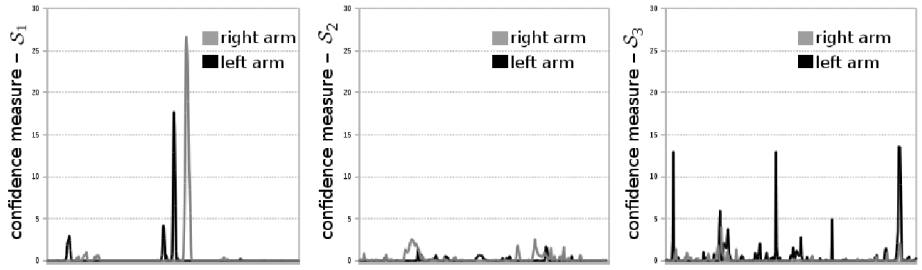

Fig. 6. The confidence measure shown for the two arms: the local maxima correspond to false estimations of the pose. They are relatively scarce and easy to detect.

to the ones computed by our method from the Xtion Pro-Live acquisition. The ART movements are most extensive because the captors can't be placed precisely on the articulation.

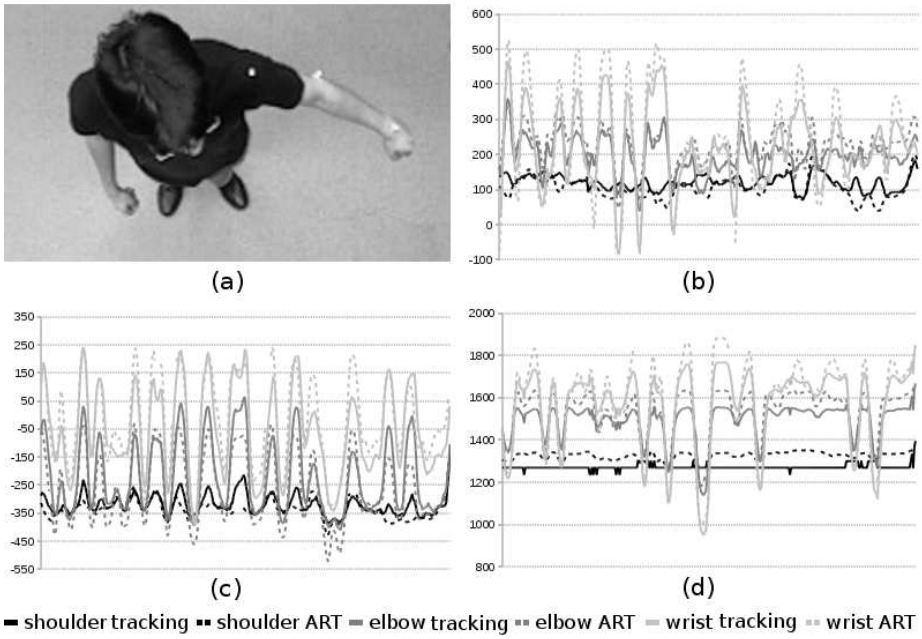

Fig. 7. Captors are used to follow the 3D positions of the shoulder, the elbow and the wrist of an arm (a). The trajectories of these captors (ART) are closed to the ones provided by our method (tracking) for the x (b), y (c) and z (d) coordinates.

4 Conclusion

In this paper we have proposed a new 3D tracking method that uses the particle filter to the particular case of the top view. A new Asus camera is used to take advantage of the depth cue. To do this, we have introduced a top view model that simultaneously used 2D and 3D fitting based on a chamfer distance. Moreover, for the behavior recognition context, a confidence measure is associated to each frame so as to detect the possible wrong estimations. The process is efficient and real-time.
The tracking is the first step of the behavior analysis. Future works would use our tracking for action recognition. A camera pose estimation [15–17] could insert our work in a Augmented Reality context with a moving camera. Finally a coupled tracking and segmentation method would give more information for the following of the process.

References

1. Isard, M., Blake, A.: Condensation - conditional density propagation for visual tracking. International Journal of Computer Vision **XXIX** (1998) 5 – 28
2. Deutscher, J., Blake, A., Reid, I.: Articulated body motion capture by annealed particle filtering. In: Computer Vision and Pattern Recognition. (2000)
3. Hauberg, S., Sommer, S., Pedersen, K.S.: Gaussian-like spatial priors for articulated tracking. In: European Conference on Computer Vision. (2010) 425 – 437
4. Horaud, R., Niskanen, M., Dewaele, G., Boyer, E.: Human motion tracking by registering anarticulated surface to 3d points and normals. IEEE Transaction on Pattern Analysis and Machine Intelligence **XXXI** (2009) 158 – 163
5. Gonzalez, M., Collet, C.: Robust body parts tracking using particle filter and dynamic template. In: IEEE International Conference on Image Processing. (2011) 529 – 532
6. Xia, L., Chen, C., Aggarwal, J.K.: Human detection using depth information by kinect. International Workshop on Human Activity Understanding from 3D Data (2011)
7. Kobayashi, Y., Sugimura, D., Sato, Y., Hirasawa, K., Suzuki, N., Kage, H., Sugimoto, A.: 3d head tracking using the particle filter with cascaded classifiers. In: British Machine Vision Conference. (2006) 37 – 46
8. Yang, C., Duraiswami, R., Davis, L.: Fast multiple object tracking via a hierarchical particle filter. In: International Conference on Computer Vision. (2005) 212 – 219
9. Kjellström, H., Kragic, D., Black, M.J.: Tracking people interacting with objects. In: IEEE Conference on Computer Vision and Pattern Recognition. (2010)
10. Oikonomidis, I., Kyriazis, N., Argyros, A.: Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. IEEE International Conference on Computer Vision (2011) 2088–2095
11. Heath, K., Guibas, L.: Heath, k., and guibas, l.j. In: ACM/IEEE International Conference on Distributed Smart Cameras. (2008) 1–9
12. Canton-Ferrer, C., S.J.C.J.R., Pardàs, M.: Multi-person tracking strategies based on voxel analysis. In: Multimodal Technologies for Perception of Humans. (2008) 91–103

13. Del Rincón, J., Makris, D., Nebel, J.: Tracking human position and lower body parts using kalman and particle filters constrained by human biomechanics. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics **62** (2011) 26–37
14. Micilotta, A., Bowden, R.: View-based location and tracking of body parts for visual interaction. British Machine Vision Conference (2004) 849–858
15. Didier, J., Ababsa, F., Mallem, M.: Hybrid camera pose estimation combining square fiducials localisation technique and orthogonal iteration algorithm. International Journal of Image and Graphics **8** (2008.) 169–188
16. Ababsa, F., Mallem, M.: A robust circular fiducial detection technique and real-time 3d camera tracking. International Journal of Multimedia **3** (2008) 34–41
17. Ababsa, F.: Robust extended kalman filtering for camera pose tracking using 2d to 3d lines correspondences. IEEE/ASME Conference on Advanced Intelligent Mechatronics (2009) 1834–1838