3D Human Tracking from Depth Cue in a Buying Behavior Analysis Context

Cyrille Migniot, Fakhr-Eddine Ababsa

3D Human Tracking from Depth Cue in a Buying Behavior Analysis Context

Cyrille Migniot and Fakhreddine Ababsa

IBISC laboratory - University of Evry val d'Essonne, France
{Cyrille.Migniot,Fakhr-Eddine.Ababsa}@ufrst.univ-evry.fr

Abstract. This paper presents a real time approach to track the human body pose in the 3D space. For the buying behavior analysis, the camera is placed on the top of the shelves, above the customers. In this top view, the markerless tracking is harder. Hence, we use the depth cue provided by the kinect that gives discriminative features of the pose. We introduce a new 3D model that are fitted to these data in a particle filter framework. First the head and shoulders position is tracked in the 2D space of the acquisition images. Then the arms poses are tracked in the 3D space. Finally, we demonstrate that an efficient implementation provides a real-time system.

Keywords: Human tracking, kinect, particle filter, buying behavior

1 Introduction

Behavior analysis based on artificial vision method offers a wide range of applications that are currently little developed in the marketing area. In the customer behavior analysis, the camera is often placed on the ceiling of the market. Only the top view of the person is also available. However, 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. The aim of the project ANR-10-CORD0016 ORIGAMI2 that supports this work is to develop real-time and non intrusive tools designed to analyze the shoppers buying act decisions. The approach is, in the first time, based on extracting and following the shoppers' gaze and gesture positions with computer vision algorithmic. It is then based on statistically analyzing the extracted data: the goal of this cognitive analysis is to measure the interaction between the shopper and their environment. This technology will provide consumer goods producers with non biased and exhaustive information on shoppers' behaviors during their buying acts.
To make the tracking possible, the depth cue is required. One of the more popular devices used to provide it is kinect, which has sensors that capture both rgb and depth data. In this paper we integrate the depth cue in a particle filter to track the body parts. The gesture recognition and the behavior of the customer could, in a second time, be analyzed using the Moeslund's taxinomy.

Pose estimation and 3D tracking have received a significant amount of attention in the computer vision research community in the past decade. To do this, the observation is fitted to a model that embodies the possible states. For an articulated target as a person, the particle filter [9] is mostly used. It estimates the current pose from a sample of possible states weighted from a likelihood function that represents the probability that a state of the model corresponds to the observation. A skeleton defines the states of the model. It comprises of a set of appropriately assembled geometric primitives [4, 7, 8] or 3D gaussians [15] to introduce the volume occupied by the body in the 3D space. The main variations on the framework come from the choice of the likelihood function. Skin color [6] and contour (matched to the chamfer distance [16]) are the most useful features. Kabayashi [11] inserts results of classifiers in the likelihood function. Some poses of the skeleton can not be executed by a human body. The sampling can be constrainted by a projection on the feasible configuration space [7] or by stochastic Nelder-Mead simplex search [12].

In the buying behavior analysis context, post treatment is not assesed and real time processing is also appreciated. In the particle filtering, the most expensive operation is the evaluation of the likelihood function because it has to be done once at every time step for every particle. Some adaptations are needed to obtain a real time processing. Gonzales [6] realizes a tracking for each sub-part of the body so as to use only simple models. A hierarchical particle filter [17] simplifies the likelihood function. The annealed particle filtering [4] reduces the required number of particles. Finally Kjellström [10] considers interaction with objects in the environment to constrain the pose of body and remove degrees of freedom.

In this paper, we propose a new human pose tracking by particle filtering in a top view. To obtain a real time processing, the model is broken it up a 2D model representing the head and the shoulders and an 3D model representing the arms. Using the depth cue provided by a kinect drastically reduces the complexity of the first one. For the second one, the pose is constrainted by the position of the shoulders.
Our main contributions are fisrt considering buying act conditions to optimize the tracking, then taking advantage of a recent data acquisition equipment and finally decomposing the model in two parts so as to reduce the filtering complexity and use simultaneously 2D and 3D models.

2 Particle Filter Implementation

We use the Xtion Pro-live camera produced by Asus for the acquisition. All the points that the sensor is not able to measure depth are offset to 0 in the output array. We regard it as a kind of noise. Moreover we only model the upper part of the body. Thus, we threshold the image to only take into consideration the pixels recognized as an element of the torso, the arms or the head. It gives a first segmentation of region of interest (ROI).

The Asus Xtion Pro-live provides simultaneously the color and the depth cues. Nevertheless the color cue is often degraded in practice. Indeed, persons on the supermarket shelves are over-lit. The tracking must be robust and the depth cue is not disturbed by lighting. Thus we only take into consideration the depth cue.

2.1 The Particle Filter

Particle filtering has been a successful numerical approximation technique for Bayesian sequential estimation with non-linear, non-Gaussian models. 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 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 resampled. 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:

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

2.2 The 2D Head-shoulders Model

Top of the head and top of the shoulders make a variation of depth that is well descriptive of the human class [14]. Moreover, the shapes of this two parts in the top view is almost constant and make ellipses. Canton-Ferrer [3] defines the volume of the person by an ellipsoid. While he estimates the position of the person, we estimate its pose. Our model is also made of 2 ellipses whose the dimension is relative to the person stoutness and are computed in the initialization step. The state vector is defined by: $V^{hs} = \{x^h, y^h, \theta^h, x^s, y^s, \theta^s\}$ where $(x^\cdot, y^\cdot)$ gives the position of the center of the ellipse and $\theta^\cdot$ its orientation for the head (h) and the shoulders (s). We first threshold the depth image to separate the pixels that are likely to correspond to the head and the pixels that are likely to correspond to the shoulders. This map is our observation. To define the likelihood function, the

ellipses given by a state vector (a particle) is matched to the chamfer distance map of the thresholded depth image. The interaction between the 2 ellipses at the neck level is introduced by constraints in the propagation step : the position of one part reduces the possible state space of the second.

2.3 The 3D Arms Model

For the arms tracking, we need to realize the tracking in the 3D space. A 3D model of the whole body could be used (figure 1(a)) but the shoulders are best tracked in the 2D space and a complete model is time consuming. A tracking is done for each arm hardly constrained by the 2D estimation of the shoulders position of section 2.2 as illustrated in figure 1(c). In our model, the arm has 5 degrees of freedom: 3 for the shoulder and 2 for the elbow. The pose of the skeleton is defined by the 5 angles of the state vector: $V^a = \{\theta_x^{sh}, \theta_y^{sh}, \theta_z^{sh}, \theta_x^{el}, \theta_z^{el}\}$. Geometrical primitives introduce the volume: arms and forearms are modeled by truncated cylinders, torso by an elliptic cylinder and finally the hands by rectangular planes.

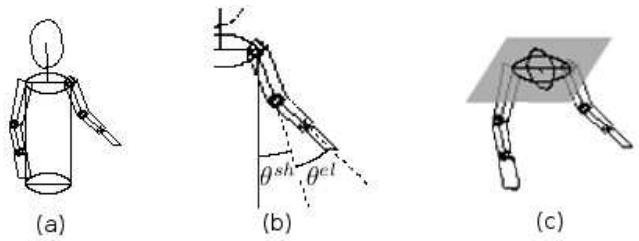

Fig. 1. The 3D models: (a) the 3D model is made of a skeleton with geometrical primitives, (b) the angles of the articulation defines the pose of the person, (c) in the 3D-2D processing the head and the shoulders are tracked in the 2D space of the recorded images whereas the arms are tracked in the 3D space.

The depth variation is well-descriptive of arm. The pixels of the foreground in the depth image are transposed in the 3D space. Then the model state is fitted to these 3D points.
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 state 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 euclidean shortest distance from a point to a 3D model.

3 Performances Analysis

We now present some experimental results. So as to control the movement of the person and to maximize the number of tested poses, we have simulated the behavior of customers in experimental conditions. In fact, the most important variation is the presence of shelves and goods. But, as the camera do not move, an estimation of the background is computed and can be removed to the frames of the sequence.Using experimental conditions is justified because the ROI also obtained are similar to the experimental ones.

The Xtion Pro live camera produced by Asus is installed at 2,9 m of the ground. It provides 7 frames per second. The dimension of a frame is 320×240 pixels. In the first experiment, we recorded two sequences $\mathcal{S}_1$ et $\mathcal{S}_2$ that are made of 450 frames (>1min) and 300 frames (≈43s). The movement of the arms are various and representative of the buying behaviors. The depth cue is extracted with the OpenNI library. The distance of use of the Xtion Pro live camera is between 0,8m and 3,5m. Consequently it can not be used in at-a-distance video surveillance but is relevant for the buying behavior analysis context.
We now estimate the quality of the arms tracking. We have manually annotated the pixels of the arms on the frames of the 2 sequences to create a groundtruth. Then we compute the average distance ε from the projection on the 3D space of each of these pixels to the model state estimated by our method. It has to be minimized to optimize the tracking. The mean variable of the particle filter is the number of particles. If it is increased, the tracking is improved but the computing time is increased too. The processing times are here obtained with a non-optimized C++ implementation running on a 3,1GHz processor. We give in the following the average processing times per frame. We can seen in figure 3 that there are no meaningful improvement over 50 particles (computed in 25 ms). With this configuration there are an average distance of less than 2,5cm between each pixel of the observation and the estimated model state. This processing is real time.
We compare our algorithm with the case where the 3D fitting presented in section 2.3 is applied on a complete 3D model (figure 1(a)) with 17 degrees of freedom. The figure 3 shows that the tracking is less efficient with this configuration. Indeed, the required number of particles is higher because the number of degrees of freedom is higher. Consequently the processing time increases. Moreover, as we realized a part-based treatment, each body part is tracked more efficiently.

In a second experiment, we evaluate the trajectories of the articulations of the arms. Two cameras ARTTRACK1 provides the 3D positions of reflecting balls with the software DTRACK. We have placed captors on the shoulder, the elbow and the wrist of the left arm of a person and we have recorded their positions simultaneously to the kinect acquisition in a sequence $\mathcal{S}_3$ (≈55s). The captors can not be placed accurately on the center of the articulations. The recorded positions are so not a groundtruth. However they can be used to evaluate the trajectories of the articulations that define the arm movement. We show in figure

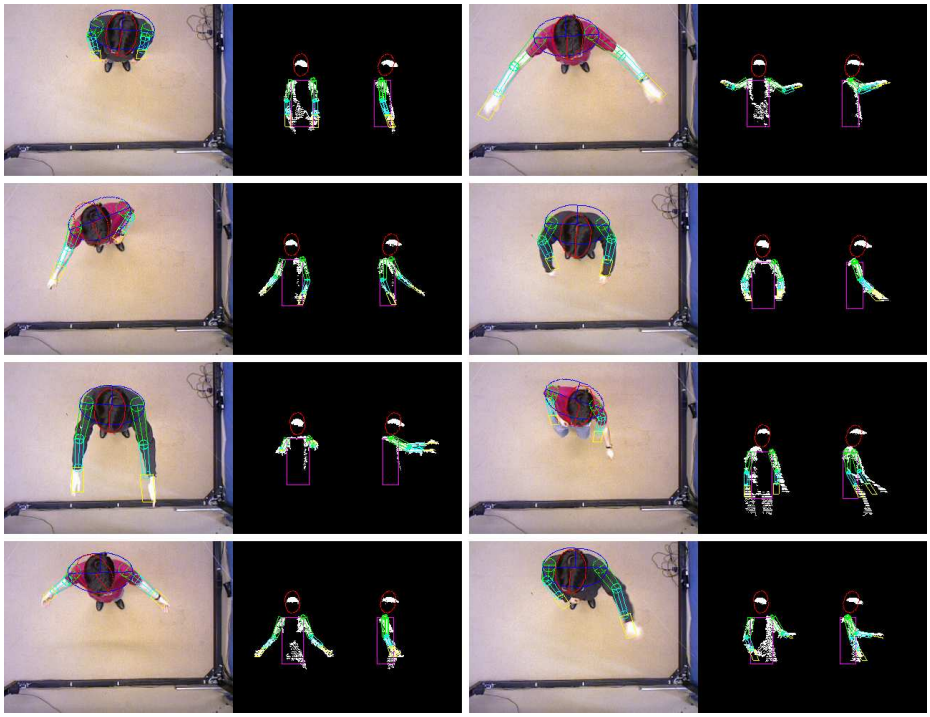

Fig. 2. The tracking provides the pose of the person: visualization of the estimated model state on the recorded frames (in left) and in the 3D space (in right) with the projection of the pixels of the depth image in white.

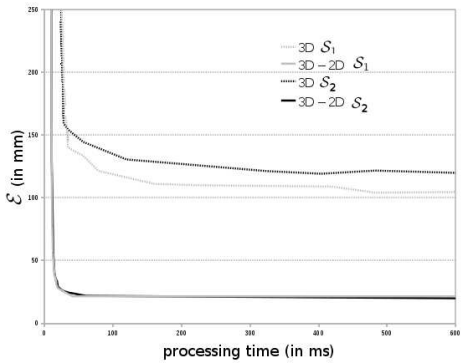

Fig. 3. Performances of the tracking with the various models on the 2 sequences: the tracking is the best with our 3D-2D method.

4 that the our trajectories well fitted the ART ones. Then the difficult case where the person bends down (the sharp peak on the z coordinate) is much

better estimated by our method. Finally our tracking is more robust when the movements are sharp (z coordinate of the wrist). This experiment validates the estimation of the arm movement by our method.

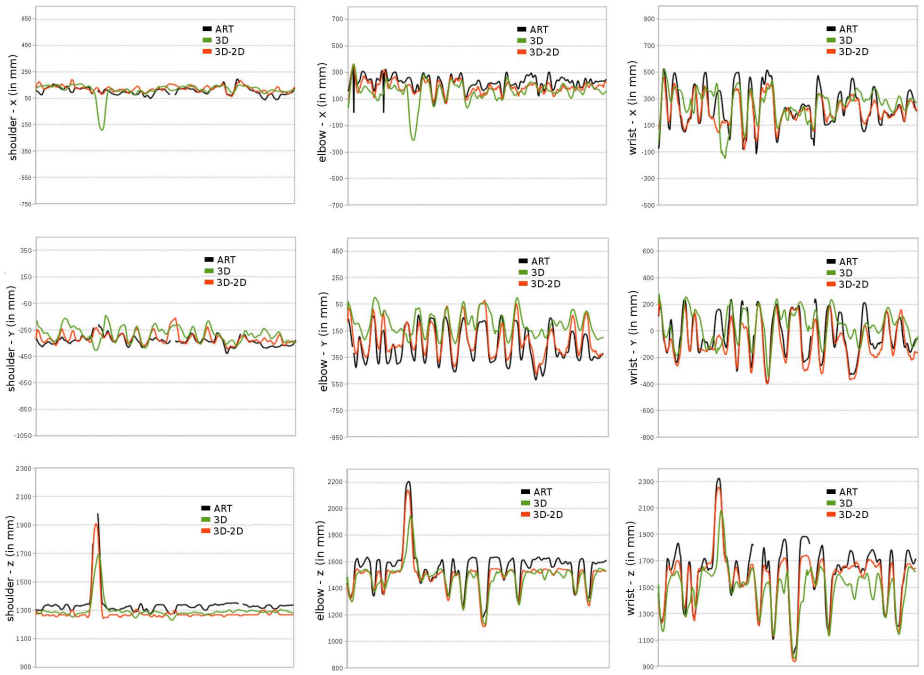

Fig. 4. Trajectories of the 3D coordinates (x,y and z) of the shoulder, the elbow and the wrist of the left arm in the sequence $\mathcal{S}_3$: our 3D-2D method well follows the articulation movements.

4 Conclusion

In this paper we have presented a 3D gesture tracking method that uses the well known particle filter method. To be efficient in the buying behavior analysis context where the camera is placed above the customers, our treatment is adapted to the top view of the person and used the depth cue provided by the new Asus camera. To do this, we have introduced a top view model that simultaneously uses 2D and 3D fitting. The process is accurate and real-time.
In the future, our method could be inserted in an action recognition processing to analyse the customer behavior. Moreover, a camera pose estimation [5, 2, 1] could insert our work in a Augmented Reality context with a moving camera. Finally an additional camera placed at the head level could refine the behavior analysis by a gaze estimation [13].

.....

References

1. Ababsa, F.: Robust Extended Kalman Filtering For Camera Pose Tracking Using 2D to 3D Lines Correspondences. IEEE/ASME Conference on Advanced Intelligent Mechatronics, 1834–1838 (2009)
2. Ababsa, F. and Mallem, M.: A Robust Circular Fiducial Detection Technique and Real-Time 3D Camera Tracking. International Journal of Multimedia 3, 34–41 (2008)
3. Canton-Ferrer, C., Salvador, J., Casas, J. R., and Pardàs. M.: Multi-person Tracking Strategies Based on Voxel Analysis. Multimodal Technologies for Perception of Humans. 91–103 (2008)
4. Deutscher, J. and Reid, I.: Articulated Body Motion Capture by Stochastic Search. International Journal of Computer Vision 2, 185–205 (2005)
5. Didier, J.Y., Ababsa, F. and Mallem, M.: Hybrid Camera Pose Estimation Combining Square Fiducials Localisation Technique and Orthogonal Iteration Algorithm. International Journal of Image and Graphics 8, 169–188 (2008)
6. Gonzalez, M. and Collet, C.: Robust Body Parts Tracking using Particle Filter and Dynamic Template. IEEE International Conference on Image Processing, 529–532 (2011)
7. Hauberg, S. Sommer, S. and Pedersen, K.S.: Gaussian-like Spatial Priors for Articulated Tracking. European Conference on Computer Vision, 425–437 (2010)
8. Horaud, R. Niskanen, M. Dewaele, G. and Boyer, E.: Human Motion Tracking by Registering anArticulated Surface to 3D Points and Normals. IEEE Transaction on Pattern Analysis and Machine Intelligence 31, 158–163 (2009)
9. Isard, M. and Blake, A.: CONDENSATION - Conditional Density Propagation for Visual Tracking. International Journal of Computer Vision 29, 5–28 (1998)
10. Kjellström, H., Kragic, D. and Black, M.J.: Tracking People Interacting with Objects, IEEE Conference on Computer Vision and Pattern Recognition (2010)
11. Kobayashi, Y. Sugimura, D. Sato, Y. Hirasawa, K. Suzuki, N. and Kage, H. and Sugimoto, A.: 3D Head Tracking using the Particle Filter with Cascaded Classifiers. British Machine Vision Conference, 37–46 (2006)
12. Lin, J.Y. Wu, Y. and Huang, T.S.: 3D Model-based Hand Tracking using Stochastic Direct Search Method. IEEE International Conference on Automatic Face and Gesture Recognition, 693–698 (2004)
13. Funes-Mora, K.A. and Odobez, J.: Gaze Estimation from Multimodal Kinect Data. IEEE Conference on Computer Vision and Pattern Recognition, 25–30 (2012)
14. Micilotta, A. and Bowden, R.: View-Based Location and Tracking of Body Parts for Visual Interaction. British Machine Vision Conference, 849–858 (2004)
15. Stoll, C. Hasler, N. Gall, J. Seidel, H.P. and Theobalt, C.: Fast Articulated Motion Tracking using a Sums of Gaussians Body Model. International Conference on Computer Vision, 951–958 (2011)
16. Xia, L. Chen, C.C. and Aggarwal, J.K.: Human Detection Using Depth Information by Kinect. International Workshop on Human Activity Understanding from 3D Data (2011)
17. Yang, C. Duraiswami, R. and Davis, L.: Fast Multiple Object Tracking via a Hierarchical Particle Filter. International Conference on Computer Vision, 212–219 (2005)