J Real-Time Image Proc
DOI 10.1007/s11554-014-0429-7

ORIGINAL RESEARCH PAPER

Hybrid 3D–2D human tracking in a top view

Cyrille Migniot · Fakhreddine Ababsa

Received: 23 August 2013 / Accepted: 2 May 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract This paper addresses the problem of 3D tracking of human gesture for buying behavior estimation. The top view of the customers, which has been little treated for human tracking, is exploited in this particular context. This point of view avoids occlusion except for those of the arms. We propose an hybrid 3D–2D tracking method based on the particle filtering framework, which uses the exclusion principle to separate the observation related to each customer and deals with multi-person tracking. The head and shoulders are tracked in the 2D space, while the arms are tracked in the 3D space: these are the spaces where they are the most descriptive. We validate our method both experimentally, so as to obtain qualitative results, and on-site. We demonstrated that it makes a good estimation for various cases and situations in real-time (≈40 fps).

Keywords Human tracking · Particle filtering · Multi-target tracking · Buying behavior analysis · Xtion Pro-Live

1 Introduction

The 3D human pose estimation is a theoretically interesting and challenging problem. Indeed this is an essential part in a wide range of modern industries, animation for games and movie, video surveillance and marketing issue. More and more areas use computer vision and image analysis to find a solution to their problems and develop methods that process data streams automatically and in real-time. This is the case in particular in marketing. This paper presents the works supported by the ANR project ORIGAMI2. The aim of the project was to develop real-time and non-intrusive tools for analyzing 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 was to measure the interaction between the shopper and his environment. This technology will provide consumer goods producers with non-biased and exhaustive information on shoppers' behaviors during their buying acts within the shelves. The work presented here is focused on the tracking of the customers posture.

We analyze the behavior of customers moving between cheese self-service shelves of a supermarket. As for the video surveillance, the top view of the person is studied. Indeed, it improves the visibility of the movement of the person and better separates the persons moving along the scene. However, in the video surveillance, the camera is traditionally far from the scene and the ground so as to have a wider viewing angle and simultaneously monitor a larger area. But the higher the camera is placed, the smaller the size of the persons in the recorded images is. The resolution is sufficient for the person localization but not for the posture estimation. This is why many approaches have been proposed for position tracking on the top view but not for the pose tracking. On the contrary, in our context, we are looking for the pose. The camera is placed fairly close to the ground so as to correctly view the customers. In the installation project, several cameras are placed in series to follow the customer along the shelves.

For a relevant behavior analysis, the 3D pose of the person is required. But this pose is hardly estimated from a

C. Migniot · F. Ababsa (✉)
IBISC laboratory, 91000 Evry, France
e-mail: Fakhr-Eddine.Ababsa@ufrst.univ-evry.fr;
Fakhr-Eddine.Ababsa@ibisc.fr

C. Migniot
e-mail: Cyrille.Migniot@ibisc.fr

Published online: 03 June 2014

traditional color image. To acquire more informative data, the camera that is placed over the customers is a Xtion PRO-LIVE [44] produced by Asus. It has sensors that simultaneously provide the color and the depth information. Hence a 3D model is fitted to 3D data. This camera is affordable and space-saving for the incorporation in the shelves. As the kinect of Microsoft, it is not suited for outdoor acquisition and has a distance of use for the depth computing between 0.8 and 3.5 m. It is also suitable for our context in a supermarket. On our experiments, the Xtion PRO-LIVE is installed 2.9 m from the floor that corresponds to the top of the shelves.

In human tracking, the images given by the acquisition equipment are the observation that 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 view. Furthermore, the color of the skin and the elements of the face are seldom available in a top view but often used. The well-known Viola–Jones face detector [41] provides for example an accurate estimation of faces localization in the front view. Nevertheless, in this paper, we present a method of posture tracking using a top view.

The second main difficulty of the work is the environment of the customers. First they interact with many elements of the supermarket (the shelves, the goods, etc.). Then several customers often appear in the image and their shape can overlap. Finally, they could be attached with moving elements as shopping trolleys, basket of products or a backpack. We have also cluttered image with disruptive elements and significant occlusions. In order to cope with these difficulties, the observation relative to each target and element of the scene must be separated. In this way each tracker is independent (with no complexity upgrading) and pose of hidden parts could be estimated from the dynamics of the movement.

This paper makes the following contributions:

1. We introduce a hybrid 3D–2D model specially adapted to the top view. The separation of the model allows us to study each part of the body in the space where its shape is the more descriptive and to reduce the computing time.
2. We propose a likelihood function with 2D chamfer distance [40] for the head and shoulders pose estimation and with 3D Euclidean distance for the arms pose estimation.
3. To handle with inter-person occlusion of the arms (head and torso cannot have inter-person occlusion in a top view), we realize multi-person tracking with a tracker per target. To reduce the influence of elements of the environment, we exploit the exclusion principle and separate the observation.
4. The proposed method is validated on more that 25 min of on-site sequences and the estimations are qualitatively and quantitatively evaluated under experimental conditions. We verify particularly the real-time behavior of our algorithm.

The remainder of the paper is organized as follows: after briefly discussing the theoretical background of human gesture tracking in Sect. 2, Sect. 3 provides detailed information related to our hybrid 3D/2D particle filtering. To deal with the data acquired in the supermarket, the exclusion principle is applied and allows multi-person tracking in Sect. 6. A confidence measure that detects wrong estimations is introduced in Sect. 5. Experimental results and analysis are presented in Sect. 6. Finally, we draw our conclusions and provide some future work in Sect. 7.

2 Related works

The multi-target tracking has been studied extensively in the literature for position but not for gesture. Several techniques use the target detection and data association module. The data association procedure matches each detection to the current tracks and chooses the best according to the shape and the color correspondence (with the Bhattacharrya distance [9, 19, 25, 30] or the trajectories of interest points [14]. Brendel [8] built a graph from the detections and solved the data association problem by finding the maximum-weight independent set of the graph. Yang [45] built his graph with an online learned condition random field. They realized their treatments on the complete sequence so as to handle the long occlusions as Andriyenko [1] who minimized an energy function defined on the whole sequence. A shortest path algorithm processed on a graph can separate in each frame the detection of each target [3, 4, 33]. Benfold [2] comparably used a Markov-Chain Monte-Carlo (MCMC) data association within a temporal sliding window. Sometimes, certain parts of the tracking, named tracklets, are linked on a data association process [43] or in a graph [38].

General optical flow based systems [5, 47] infer the successive positions from motion estimation. Schwartz [35] uses it in 3D data for differentiation of body parts that occlude each other. Optical flow is nevertheless less efficient for articulated object tracking.

Object tracking that usually deals with targets of identical appearance can be ranged by learning individual target models: Breitenstein [6] learned target-specific classifiers at run-time to distinguish between the tracking targets; an online boosting is used by Luber [26] on the RGB-D data. A part-based detection and tracking [25, 37, 42] offers

many advantages: it is more tolerant to view point changes and pose variations and it can deal with partial occlusions. A dynamic occlusion handling predicts partial occlusions and selects the visible parts of the body. In [37], if only the head can be seen, it is tracked by a Kalman filter instead of a particle filter in the other cases. In crowded scenarios, Pellegrini [31] modeled people social behavior to more accurately predict their movement.

A body-part learning associated with a random forest allows joints detection used for human pose estimation. This process is more relevant for detection than for tracking: Dantone [11] announced a running of few seconds per image. Shotton [36] realized a local body-part detection to infer joints detection from a huge synthetic dataset. Hu [17] improved by MCMC a first estimation of the joints position computed from face, torso and skin detection. The cue and dataset used for these methods are incompatible with a top view.

Using classifiers is time-consuming. Mitzel [30] used a statistical Poisson process to select the relevant 3D ROI and reduced the detection field. But for a real-time processing, a Kalman or particle filtering is suitable.

The state of all the targets can be modeled in a single particle [20]. The interaction between the targets is considered in the particle formation. Choi's method [10] works with a moving camera. A MCMC solves the joint camera estimation and multi-target problem. Nevertheless, a joint particle filter suffers from exponential complexity in the number of tracked targets. That is inconsistent with a real-time processing. To prevent this, a set of trackers could be associated with each target. To avoid the problem produced by the overlapping of different targets, Gonzales [13] used the exclusion principle introduced by MacCormick and Blake [27]. At each step, the observation (the pixels of the recorded image) is split between the targets. Xing [43] selected the best subset of current observations which corresponds to visible parts to update particle weights. The blocking methods penalize particles that overlap zones with other targets [9]. Hence each tracker is associated with the relevant observation part and can be performed almost independently. We use too the segmentation of the observation because of its low level of complexity and its speed of execution. We have to remain real-time in multi-target operating.

The human tracking in the top view is seldom explored, although it has numerous applications in video surveillance. Heath [14] estimated the 3D trajectories of salient feature points (primarily at the shoulders level) that he used as the observation for the particle filtering. Canton-Ferrer [9] defined the exclusion zone for the blocking by an ellipsoid. For the tracking, he separated 3D blobs and used a particle filter where each particle represents a voxel of the blob. It estimates the centroid of the blob that models the person. These methods only track the position of the person and not its pose.

To obtain a behavior analysis, the estimation of the gesture is required. Nevertheless, a lot of researches have been devoted to develop articulated body pose tracking methods for a single target. To do this, the particle filter [18] has been mostly used. It is a Bayesian sequential importance sampling technique, which recursively approximates the posterior distribution using a finite set of weighted samples. The samples are selected to correspond to the observation and propagated according to the dynamic of the system. In practice the observation is fitted to a model that embodies the possible states. A skeleton defines the states of the model. It comprises of a set of appropriately assembled geometric primitives [12, 15, 16] to introduce the volume occupied by the body in the 3D space. Stoll [39] distributed points of interest along the skeleton and included 3D Gaussians centered on the points of interest to provide the volume. We use a similar model but adapted for the top view.

To describe the human class, various features are used: skin color, shape of the silhouette, movement, etc. Deutscher [12] selected points along the surface of its model that he matched to the observation. The Ω-like shape formed by the head and the shoulders is particularly well descriptive of the human shape [24, 28].

In the buying behavior analysis context, post treatment is not assessed 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 [13] realized a tracking for each sub-part of the body so as to use only simple models. A hierarchical particle filter [46] simplifies the likelihood function. The annealed particle filtering [12] (APF) reduces the required number of particles. Finally Kjellström [22] considered interaction with objects in the environment to constrain the pose of body and remove degrees of freedom.

All poses of the skeleton are not possible in practice. For example, the head cannot rotate over 360°. The sampling can be constrained by a projection on the feasible configuration space [15]. As [13] we decompose the model so as to treat each part in its best representation and to reduce the complexity of the state space.

Finally, transformation can be applied to provide an unimodal likelihood model that allows using a Kalman filter. Larsen [23] used stereo data to disambiguate depth and Brox [7] tracked interest points provided by SIFT.

Data association means post-processing and learning is time-consuming, which is incompatible with a real-time context. In our method, we estimate the pose of the person from a particle filter with an articulated model. We reduce

the computing time by decomposing the model in two parts: the head and shoulders are tracked in the 2D space and the arms are tracked in the 3D space. Indeed, the model is not complex enough to use a layered particle filter as AFP. In order to obtain a suitable result with multiple customers (multiple targets) we use the exclusion principle. A sharing of the observation between the persons is done from their predicted positions.

3 Hybrid 3D–2D human gesture tracking

3.1 The particle filtering

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. In our case, the state of the model represents the pose of the person and the observation is the data acquired by the Xtion PRO-LIVE camera. 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. Hence the particle sample is always located around the expected pose. To automatically define what are the large and the low weights, we use the SIS systematic resampling algorithm [21] resumed in [34].
- *Propagation:* particles are propagated using the dynamic model of the system $p(x_{k+1}|x_k)$ to obtain $\{x_{k+1}^i, \frac{1}{N}\} \sim p(x_{k+1}|y_k)$. This step aims to guess the next state. We chose a constant speed propagation followed by a Gaussian centered on the estimated pose and with a standard variation depending of the variation of the related feature.
- *Weighting:* particles are weighted by a likelihood function related to the correspondence between the model and 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 $\{x_{k+1}^i, \omega_{k+1}^i\} \sim p(x_{k+1}|y_{k+1})$. The unsuitable tested state are hence detected and deleted in the next resampling step.
- *Estimation:* the new pose is approximated by the weighted mean of the particle: $x_{k+1} = \sum_{i=1}^N \omega_{k+1}^i x_{k+1}^i$. This step is only used to provide a single state rather than a probability density. It is independent of the particle sample evolution.

The main variation of the method to be adapted to a particular problem is the definition of the model and the likelihood function. In the following, we explain the choice we have done for each of these points and each tracking process.

3.2 The 2D head-shoulders tracking

In a first time, the head and the shoulders are the only tracked parts. This choice is explained by our assumption that the head and the shoulders are more descriptive in the 2D space of the recorded image while the arms are more descriptive in the 3D space. Indeed, the shapes of the head and the shoulders are fairly constant and they are mainly described by their displacement on the scene. The relevance of this choice will be proved by experimental tests in the Sect. 6.2.

The 2D observation is composed of the color and the depth images recorded by the camera (Figs. 1, 2, 3). On the depth image, the shapes of the head and the shoulders are easily identifiable and fairly constants: it makes two ellipses. Moreover, the Ω-like shape of the top of the body used by Micilotta [28] on the front view leads to sharp depth edges between the two parts. Depending on the distance to the camera (given by the top of the head depth), the distance between the two shoulders is fairly constant. The radius of the ellipse related to the shoulders is only dependent on the person's stoutness. The fitting between the ellipses and the observation must allow some variation of the model: we use a 2D chamfer matching.

The 2D model We use an association of two ellipses corresponding to the usual human anatomic feature to model the head and the shoulders (Figs. 4, 5). Each ellipse is defined by its position in the image and by its orientation: there are 3 degrees of freedom per ellipse. It provides the main information about the person's position for the behavior analysis: the ellipse of shoulders gives the body position in the aisles of the supermarket and the ellipse of head gives the relative orientation that could infer the direction of the gaze. The pose of the arms is estimated after. For a better description, the weighting of each ellipse is realized separately. The two parts are linked by constraints in the propagation step. The pose of the person is constrained by human biomechanics: the head cannot rotate 360° around the neck and the neck is not expandable. The spatial distance between the two centers and the difference of orientation are limited. The particles must not belong to these impossible states.

Fig. 1 The data are acquired in a supermarket. The difficulties are numerous: cluttered image, occlusions, multi-targets and various goods and shopping trolleys

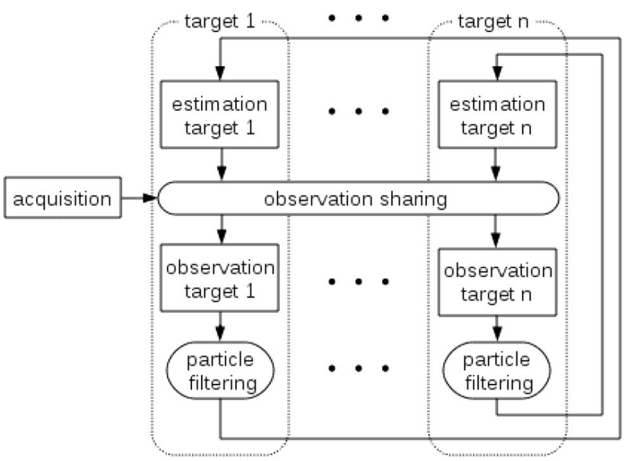

Fig. 2 Overview of the method: the particle filtering is executed independently for each target thanks to the observation sharing

The likelihood function The elliptical shapes done by the head and the shoulders on the top view are fitted to the two ellipses of the model. To do this we use a 2D chamfer distance that provides an efficient comparison between shapes robust to slightly incomplete or occluded edges. First the top of the head is defined as the highest part of the body; thus the element of the candidate blob with the minimal depth value. According to this depth value, the depth image is thresholded on the middle of the head and the end of the shoulder level. Hence we obtain a mask where each pixel is assigned either to the head or the shoulders or the background. The edges of this mask provide an estimation of the interesting shapes of the depth array. The chamfer mask is computed so that the chamfer distance is not computed again for each particle. The chamfer mask provides the shortest distance from each pixel to the edges [40].

The weight assigned to a particle is the mean value of the chamfer mask for the pixel of the ellipses related to the particle. Let Δ_i^{ell} be the set of pixels of the ellipses provided by the state vector of the particle i and the chamfer mask M_{ch}. The weight of the particle i is given by

$$d_i^{\mathrm{ch}} = \underset{p \in \Delta_i^{ell}}{\mathrm{mean}}(M_{\mathrm{ch}}(p)) \tag{1}$$

$$\omega_i^{2D} = \mathrm{e}^{-d_i^{\mathrm{ch}}} \tag{2}$$

3.3 The 3D arms tracking

Contrary to the head and the shoulders, the arms are much more descriptive in a 3D configuration. Indeed, instead the top of the head and the shoulders, that are approximatively at a depth level, the arms are large depth variation and their shape are so distorted by the camera viewscape in the 2D space. The depth array provides a set of 3D points. This set is incomplete and non-continuous. It represents the data visible from the top view that gives a partial representation of the 3D scene (Fig. 6).

The 3D model We introduce a model that is a skeleton whose rigid parts represent the arms and the forearms with the hands. There are, for each arm, 2 separated parts and 2 articulations (Figs. 7, 8). We hypothesise that each shoulder has 3 degrees of freedom and each elbow has 2 degrees of freedom. In the literature, a single degree of freedom is sometime assigned to the elbow but we are experimentally noticed that 2 degrees favorite the transition between arms poses. To represent the volume, geometrical primitives are added: arms and forearms are modeled by truncated cylinders and the hands by rectangular planes.

This model is hardly constrained by the previous one: the 2D position of the shoulders, computed in the previous

Fig. 3 The Xtion PRO-LIVE provides a color and a depth image. For the head–shoulders tracking this data makes the observation

Fig. 4 The 2D model is made of two ellipses

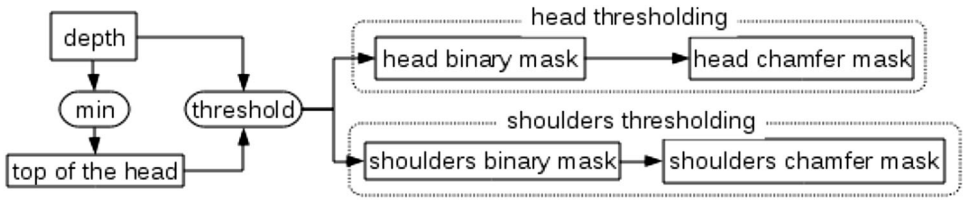

Fig. 5 The depth image is thresholded at the head and the shoulder levels computing from the top of the head position to realize the chamfer mask

Fig. 6 The Xtion PRO-LIVE camera provides simultaneously a color and the depth images. A set of 3D points of visible part of the scene from the top view is computed from the depth image

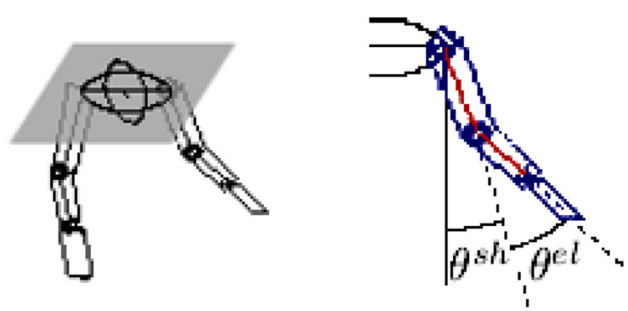

Fig. 7 The arms pose is defined by a 3D model constrained by the location of the shoulders computed by the 2D tracking (in the *left*). This model (in the *right*) is made of an articulated skeleton (in *red*) that comprises of a set of appropriately assembled geometric primitives (in *blue*). The degrees of freedom correspond to the angles at the shoulder and elbow levels

part for the 2D tracking, gives the location and the orientation of the person in the scene. The tracking studies the movement of the arms in relation to the torso. These relative movements are more descriptive of the person's actions.

All the movements of this articulated model are not possible in practice: for example, the wide angle of flexion at the elbow is almost 180° because of the olecranon

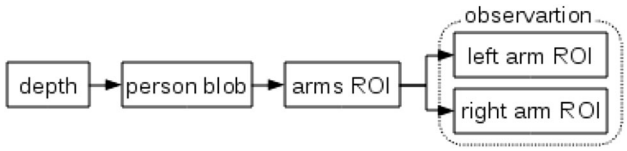

Fig. 8 Overview of the 3D observation pre-processing: the blob corresponding to the studied person is extracted from the depth array; then the head and the shoulder parts are deleted and finally the pixels corresponding to each arm are separated according to the shoulders ellipse orientation

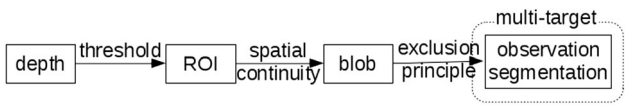

Fig. 9 Overview of the observation segmentation: thresholding of the depth array and spatial continuity limit the studied observation; the exclusion principle separates it in case of multi-target tracking

presence. Hence the state space is limited by anatomical constraints. As bones of the skeleton are rigid, the estimation of the hand (and elbow) position can be directly computed from the values of angles of the articulations.

The likelihood function The goal is that the set of 3D points of the observation be the closest of the 3D model defined by the particle. As we have 3D data, evaluated data are expressed in an unit that corresponds to physical information: the distance in meter. Moreover, the 3D space has the advantage of keeping the distance constraints. Indeed, in the recorded 2D image, an object is bigger if it is close to the camera than that if it is far. The arm part length can also lead the processing.

As the 3D process is addressed to the arms, we reduce the observation to this part. First, a depth threshold deletes the elements too low to belong to the upper part of a body: that provides some regions of interest (ROI). Then, the elements of the body in the ROI are spatially continuous. We only select the continuous ROI that corresponds to the tracked person (we know its location thanks to the previous 2D tracking) and that we named blob. Finally, we delete the set of pixels that corresponds to the head, the torso and the shoulders from the estimation given by the previous 2D pose estimation. Let Δ_{arm} be this new set of 3D points. This pre-processing is summarized in the left part of the diagram in Fig. 9.

The model describes well the person pose if the elements of the observation are closed to the model in the 3D space. The weighting of the particle i is given by

$$d_i^{3\text{D}} = \underset{p \in \Delta^{\text{arm}}}{\text{mean}}(\mathcal{D}^{3\text{D}}(p, \mathcal{M}_i)) \tag{3}$$

$$\omega_i^{3\text{D}} = \text{e}^{-d_i^{3\text{D}}}, \tag{4}$$

where $\mathcal{M}_i$ is the 3D model pose defined by the particle i and $\mathcal{D}^{3\text{D}}$ is the 3D distance from a point to a 3D volume defined by

$$\mathcal{D}^{3\text{D}}(p, \mathcal{M}) = \min_{i \in \mathcal{M}}(d^{3\text{D}}(p, i)), \tag{5}$$

where $d^{3\text{D}}$ is the Euclidean distance between two points in 3D space.

4 Exclusion principle inclusion

Concatenating the data of the all targets in a single-state vector increases dramatically the complexity of the system and the computing time. Applying the exclusion principle to track each target independently is relevant to obtain a real-time processing. Instead of sharing the impact of each element in the state vector, it is the observation that is shared. Indeed, this is the observation that provides the interaction between the various elements of the scene. If the influence of other objects is deleted from the observation, a tracking can be executed independently.

For each moment k and target t, a prediction of the model state $\hat{x}_{t,k}$ is computed from the previous estimation $x_{t,k-1}$ and the dynamic of the system (a constant speed model in our case). Each prediction defines a 3D shape related to a model state. Let $\mathcal{T}$ be the set of tracked targets. We assign each pixel of the observation (ROI) to the target whose model state is the nearest in the 3D space. The assignation $a(p)$ of the pixel p is defined by

$$a(p) = \underset{t \in \mathcal{T}}{\text{argmin}}(\mathcal{D}^{3\text{D}}(p, \hat{x}_{t,k})) \tag{6}$$

Some examples of observation sharing are shown in Fig. 17.

5 A confidence measure

Sometimes, some particular poses are difficult to estimate. It is the case when the fist is held high because the visible part of the forearm in the top view is hidden by the hand. In our method, the model is matched to the observation but the observation is not matched to the model. Consequently, if the observation does not provide enough information, the estimation is degraded. Let Δ_{skeleton} be the pixels of the skeleton of the estimated model projected in the 2D space of the recorded images. The measure $\mathcal{C}$ evaluates the matching of the observation to the model. It is defined by the average distance in the 2D image from pixels in Δ_{skeleton} to the nearest element of the ROI. It is expressed in number of pixels. With a good correspondence, it is set to 0.

This variable is much less descriptive than our likelihood function. Moreover, combining it with our likelihood function is time-consuming and does not improve the tracking quality. Nevertheless, $\mathcal{C}$ effectively detects the cases where our estimation, in exceptional circumstances,

is defective. Thus we use $\mathcal{C}$ as a confidence measure. If for a frame it is higher a threshold (10 pixels), the user knows that the estimation of this frame is not taken into consideration. It could be important in future work for an action recognition process.

6 Performances

6.1 Experimental device

To demonstrate the effectiveness and robustness of the proposed method, we performed simulation of the behavior of customers under experimental conditions. We have installed the Xtion PRO-LIVE camera produced by Asus [44] at 2.9 m of the ground. The dimension of a frame is 320×240 pixels. Xtion PRO-LIVE has a frame rate of 30 frames per second. One of our main objectives is to obtain a real-time algorithm. That is why we propose in the following the accuracy performances in correspondence with the processing time it requires. The processing times given in the following are the average time required for a frame and obtained with a non-optimized C++ implementation running on a 3.1 GHz processor. In the particle filtering, increasing the number of particles improves the accuracy but increases the computing time. We present, therefore, criteria that represent a data relative to the accuracy of the tracking and that is easily understandable by the user. We promote the distance in meter in the real space between our estimations and the true position as for example in Eq. 7. Thus we display qualitative criteria as a function of the processing time to evaluate the tracking quality. The number of particles is an internal parameter. It is not a criterion to select the best method.

To evaluate various positions and configurations, our tests are realized on 4 sequences (Fig. 10):

- $\mathcal{S}_1$ (>1 min) contains little displacements of the person but large and various movements of the arms.
- $\mathcal{S}_2$ (≈43 s) is similar to the first one but with flashy colors.
- $\mathcal{S}_3$ (≈54 s) where the left arm realizes various movements, is recorded with the ART protocol which will be presented later.
- $\mathcal{S}_4$ (≈32 s) contains two persons with frequent occlusion of the arms.

These sequences are available for comparison in https://evra.ibisc.univ-evry.fr/origami/.

Ground-truth To quantitatively evaluate our algorithm efficiency, we need a ground truth to compare to our results. First, for 2D evaluation, we have manually annotated the position of the center of the shoulders on the frames of sequences $\mathcal{S}_1$ and $\mathcal{S}_2$. Then, for the evaluation of the observation segmentation with multi target, we have manually annotated the pixels of the frames of the sequence $\mathcal{S}_3$ that belong to each target. The quantitative evaluation of the 3D estimation focuses on the arms. The set of pixels of the arms Δ_{arms} are manually annotated in the 2D space for all the frames of the sequences $\mathcal{S}_1$ and $\mathcal{S}_2$.

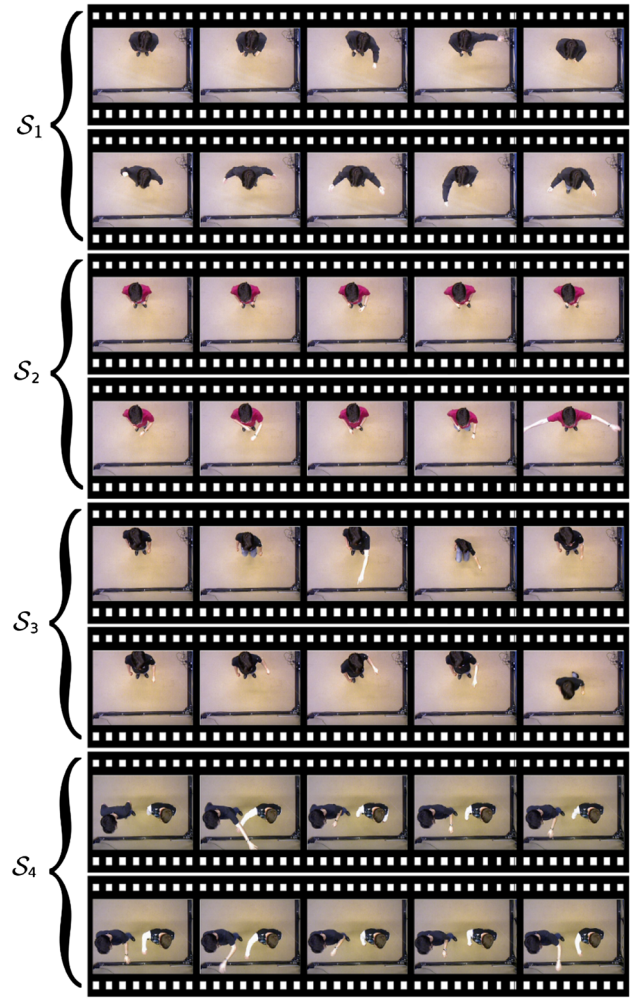

Fig. 10 Set of experimental sequences used for the qualitative and quantitative evaluation of our algorithm

We propose an other mean of comparison for the 3D evaluation. To check that our estimation of the arms pose is a good description of the movement, we want to compare the trajectories of the articulations. Thanks to the software DTRACK, the 3D positions of reflecting balls can be followed by an association of two cameras ARTTRACK1. Reflecting balls are placed on the shoulder, the elbow and the wrist of the left arm of a person. Then the balls' positions are recorded by this ART process simultaneously to the Xtion PRO-LIVE acquisition in the sequence $\mathcal{S}_3$. The ART recorded positions cannot be strictly considered as a ground-truth since the captors cannot be placed on the

centers of the articulations. But we use it as a reference to evaluate the trajectories.

Complexity and time processing The parameter that induces the complexity of our algorithm is the number of particles N. Each of the 3 steps of the particle filtering is realized one time per particle and frame. The complexity is, therefore, defined as $\mathcal{O}(N)$. Chamfer mask computing, preprocessing and estimation are processed one time per frame. Their processing times are negligible in comparison with the processing time of the particle filtering steps. As the particles of different targets are independent, the processing time is proportional to the number of target currently tracked.

The required number of particles is defined by the number of degrees of freedom of the state vector. Indeed, the particles must be enough to be distributed properly in the state space. In our method, the model is separated in two parts in particular to decrease the number of degrees of freedom. We have compared our algorithm with two full 3D methods [29] with a model of 17 degrees of freedom. The second one is a part-based processing where the state vector is shared for the sampling and the propagation step. Their large state space induces a bigger required processing time. Therefore, a real-time processing requires to minimize the number of degrees of freedom. That has influenced our choice to separate our model in two parts. In the following, these two methods are referred as (3D) and (3D–PB) and our method is referred as (3D–2D).

For an efficient functioning of our algorithm we have experimentally noted a processing time of approximately 25 ms per frame. This corresponds to a frame rate of 40 fps and a throughput of 3.85 Mpps. This value increases with the number of current tracked targets. Nevertheless, our algorithm could be optimized, for example with a parallelization of the process for each target.

6.2 Head and shoulders tracking

The 2D shapes well describe the human class for the head and the shoulders. We check that, on these body parts, our 2D tracking is more effective that a basic 3D tracking. To do this, we compare our results with the two 3D tracking methods [29]. Our method provides a MOTP less than 20 mm while the full 3D methods provide a MOTP of around 50 mm. Separating the model increases the 2D accuracy.

The tracking from a top view is less studied. Canton-Ferrer [9] and Heath [14] realized a person location tracking from color image on a top view. They obtained a MOTP of around 150 mm while we obtain less than 20 mm. The use of the depth array leads to more robustness.

In Fig. 11, we have compared our 2D tracking with the ellipse fitting of Pilu [32] processed on the depth array and with a chamfer fitting processed on the edge (obtained by a Canny processing). We notice that our method provides a much more accurate tracking in a smaller processing time.

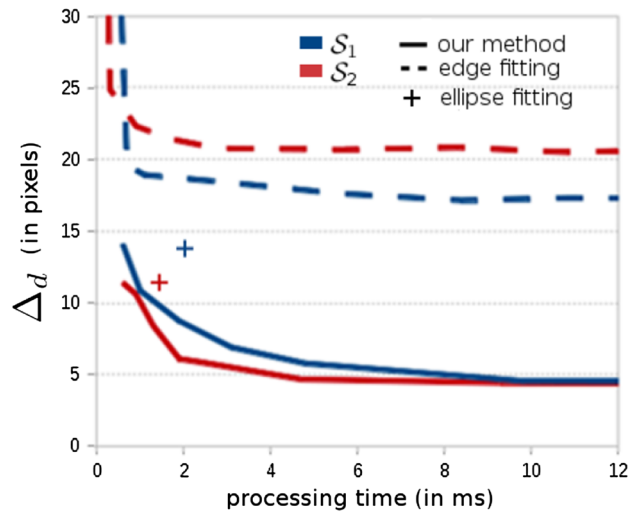

Fig. 11 Variation Δ_d of the 2D position of the shoulder ellipse from the ground-truth to our method or to an ellipse fitting [32] on the depth array or to a chamfer fitting on the edges: our 2D tracking is more efficient and using only color array significantly reduces the accuracy

6.3 Arms tracking

The shape of the arms is not descriptive in the 2D space where the articulated aspect is not represented. The arms likelihood function must be defined in the 3D space. In Fig. 12, the models estimated by our method are projected in the 2D recorded image and in the 3D space. Our estimation is qualitatively well fitted to the observation for various positions.

Let ε be the average 3D Euclidean distance between the points of Δ_{arms} and the estimated model. A low value of ε indicates a good correspondence between the model and the observation and so a good estimation.

$$\varepsilon = \underset{p \in \Delta^{\text{arms}}}{\text{mean}}(\mathcal{D}^{3\text{D}}(p, \mathcal{M}_{x_{k+1}})) \tag{7}$$

The evolution of the accuracy/processing time curves contains 3 parts (Fig. 13). First a low time processing comes from a reduction of the number of particles. Under a certain number, the spread of the state space is inappropriate and the accuracy falls rapidly. In the contrary, over a certain number, adding particles does not substantially improve the tracking. The time processing so increases for no improvement. Between these two parts is the operating area where the system works properly. The compromise between accuracy (to optimize the tracking) and time-processing (to obtain real-time) may be correct if it is taken in this range. For our 3D–2D method, the operating area is

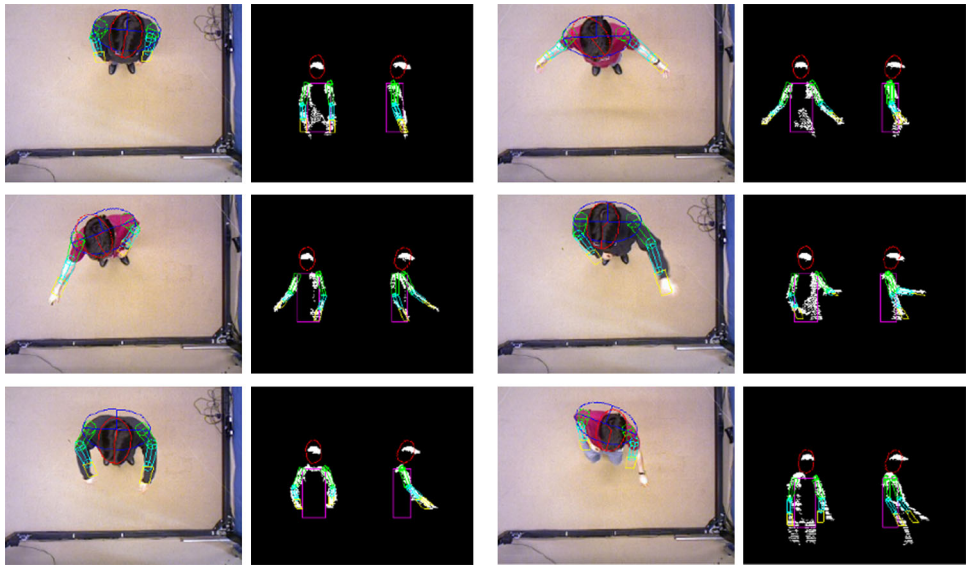

Fig. 12 The tracking provides the pose of the person: on the left the model in the color image and on the right the model in the 3D space (the pixels in white correspond to the points given by the depth image)

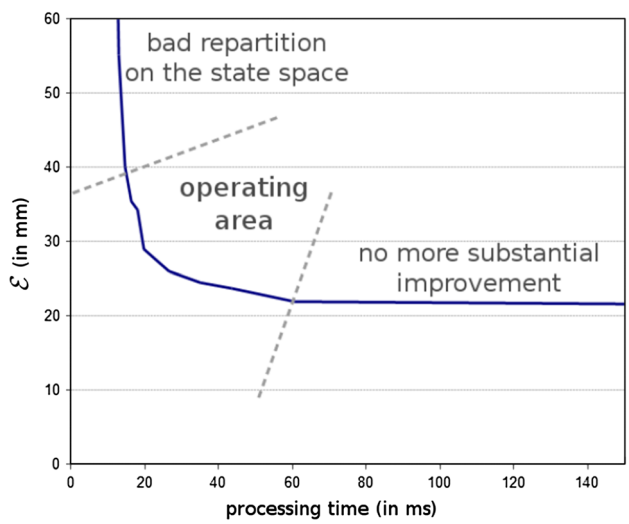

Fig. 13 Zoom in the performances of the 3D–2D tracking methods in average for the sequences $\mathcal{S}_1$ and $\mathcal{S}_2$. There are 3 main areas: if the number of particles is too small, their spread on the state space is insufficient; but increasing it over a certain number does not improve the accuracy; the last area is the operating area where the compromise accuracy/processing time should be chosen

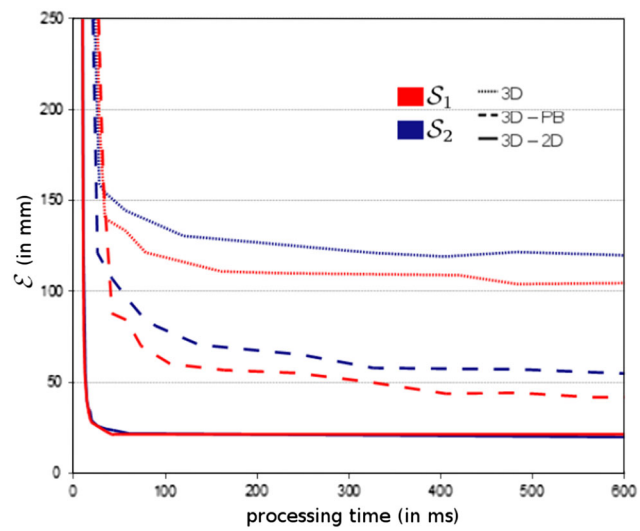

Fig. 14 Performances of the 3D tracking methods in sequences $\mathcal{S}_1$ and $\mathcal{S}_2$: the 3D–2D method we propose is substantially better. We obtained an operating area for a smaller processing time and with an accuracy 4–6 times better

in a tight range providing good tracking in a very competitive processing time.

As we can seen in the Fig. 14 and in Table 1, our hybrid 3D–2D process explicitly provides the best tracking quality. Then, studying the body parts has a real influence. For our 3D–2D processing, there is no meaningful improvement over 50 particles (processed in approximately 25 ms). In this configuration, the average distance between a point of the observation and the estimated model is less than

Table 1 The average best compromise for $\mathcal{S}_1$ and $\mathcal{S}_2$ that corresponds to the inflexion point of the curves in Fig. 14. Our hybrid 3D–2D method is faster and more accurate

Method	Processing time (ms)	ε (mm)
3D	80	130
3D–PB	80	75
3D–2D	25	25

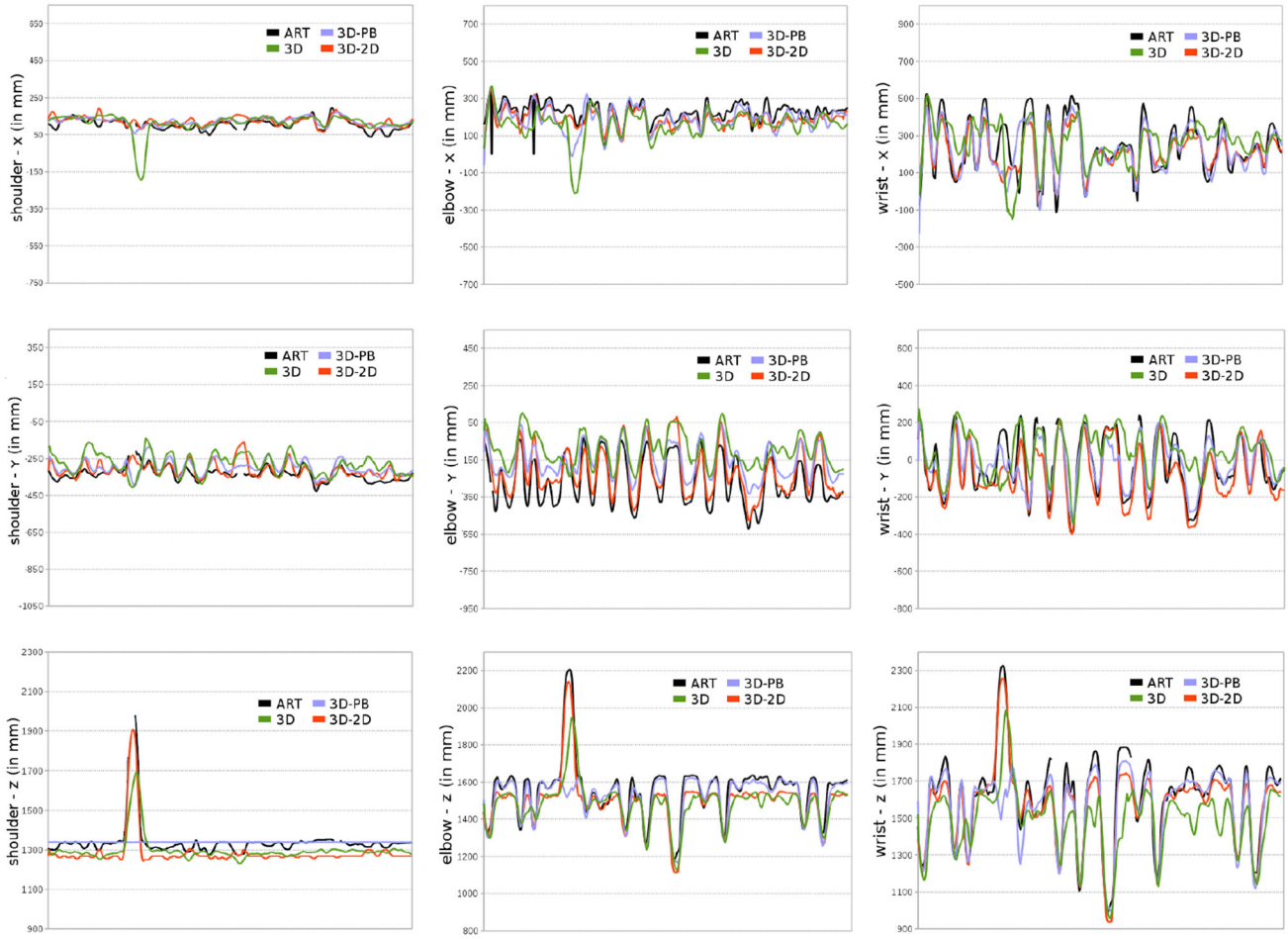

Fig. 15 Trajectories of the 3D coordinates (in row) of the shoulder, the elbow and the wrist (in column) of the left arm in the sequence $\mathcal{S}_3$. Shoulders are attached to the torso and have a slower movement. The full 3D model has more difficulties to follow this movement. That proves that the shoulders deserve to be tracked separately (in the 2D space). On the contrary, the wrist has the sharpest movements because it cumulates the movements of the arm and the forearm. We notice that these large displacements are well tracked by our method

2.5 cm. The processing is thus efficient in real-time. Our method is less time-consuming than the full 3D tracking because the number of degrees of freedom is smaller. The sampling step requires a smaller number of particles and the likelihood function is computed fewer times. The tracking is more accurate because it is more robust to sharp movements and to small ROI.

Body parts trajectories Fig. 15 shows that our trajectories of the parts of the arm are well fitted to the ART ones. Then the difficult cases where the person bends down (the sharp peak on the z coordinate) are much better estimated by our method. This movement is scarce and the dynamic of the system penalizes it in a state space with a high number of degrees of freedom. Separating the model allows to be more robust to these particular cases because more attention is given to important feature as the height of the person. Finally, our tracking is more robust when the movements are sharp (z coordinate of the wrist). This experiment confirms our results.

The confidence measure We focus on the arms movement. We computed a confidence measure $\mathcal{C}$ for each arm with the estimation for all the frame of the sequences $\mathcal{S}_1, \mathcal{S}_2$ and $\mathcal{S}_3$. Figure 16 shows that the wrong estimations are evident with this criterion. We notice that there is a very small number of wrong estimations: for the 3 sequences, there are 1.52 % of wrong estimation.

The major problems have to do with the brandished fists (sequence $\mathcal{S}_1$). There are also difficulties with the too sudden movements (sequence $\mathcal{S}_3$) at the propagation step. But it provides generally a wrong estimation for only one frame.

6.4 Multi-person tracking

To evaluate the accuracy of the observation segmentation between the two targets of the sequence $\mathcal{S}_3$, we have computed the mean number of correctly assigned pixels for each frame. The overlapping rate is more than 99 %. Thus

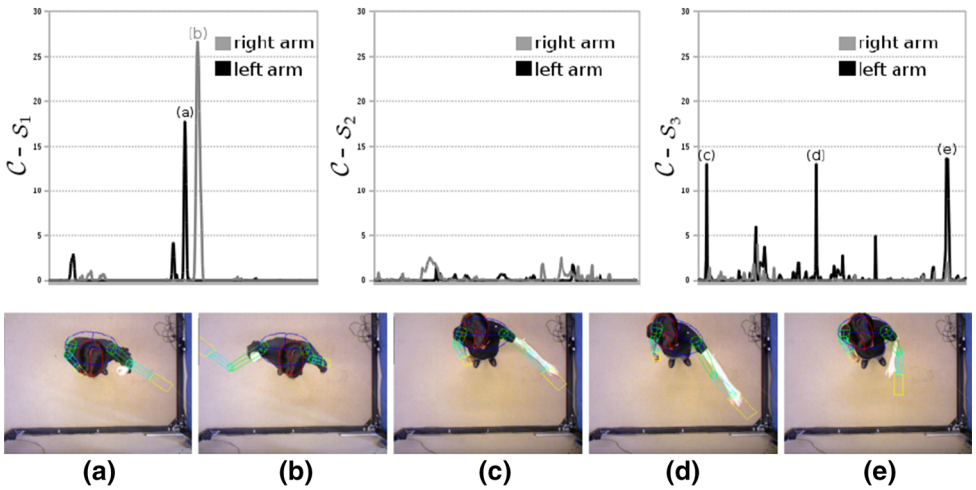

Fig. 16 The confidence measure detects the wrong estimations: when the fist is held high (**a**, **b**) or when the movement is too fast (**c**–**e**). The wrong estimations represent only 1.5 % of the frames

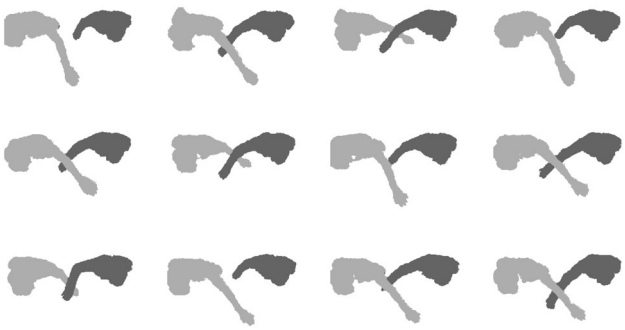

Fig. 17 Segmentation of the observation for several frames of the sequence $\mathcal{S}_4$ obtained with the exclusion principle presented in Sect. 4

very few wrong pixels are considered as observation. Moreover, they are spatially very close to the well-assigned pixels so the estimation is not much affected. We are in a situation very close to the single target one.

6.5 On-site tracking

In experimental conditions, we have the control of what happens and we can simulate the situation we want. It is more convenient to realize quantitative evaluation. Nevertheless, our method has to be used in a supermarket. We have thus realized too tests on-site. Many new difficulties appear: in practice the position where the arms are along the body represents the great majority of cases. In this position the arms are hardly discernible in the top view. But this case is not relevant in the buying action recognition step. Then, to not enter in collision, persons can adopt a strange movement. Moreover, all the shelf corridor cannot be recorded by a single camera. Several cameras have

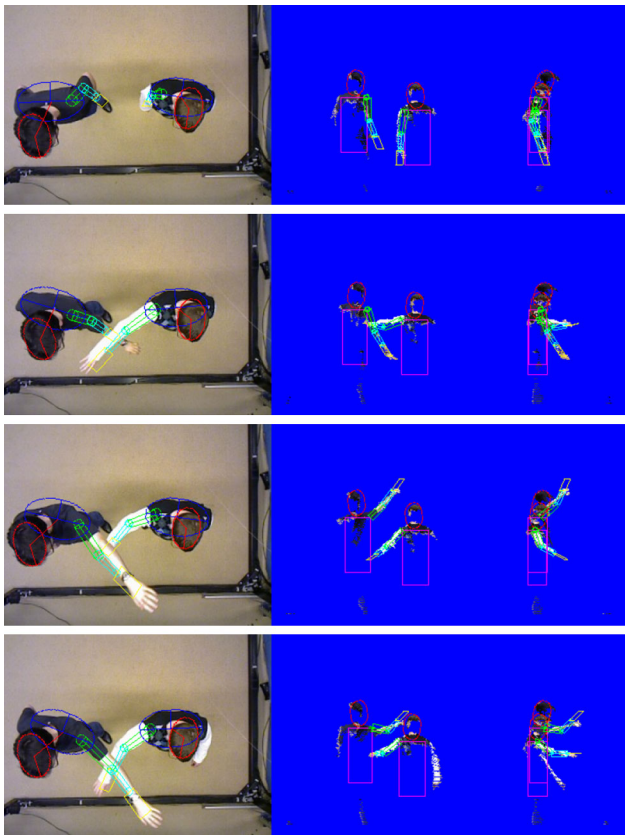

Fig. 18 Multi-person tracking obtained on the sequence $\mathcal{S}_4$

to be placed and synchronized to follow the displacement of a customer on the whole corridor.

We tested approximately 25 min of on-site recording with various customers and behaviors. These sequences contain multi-person view, interaction with trolley and other objects, fast movement, etc. Figures 17, 18, 19 show

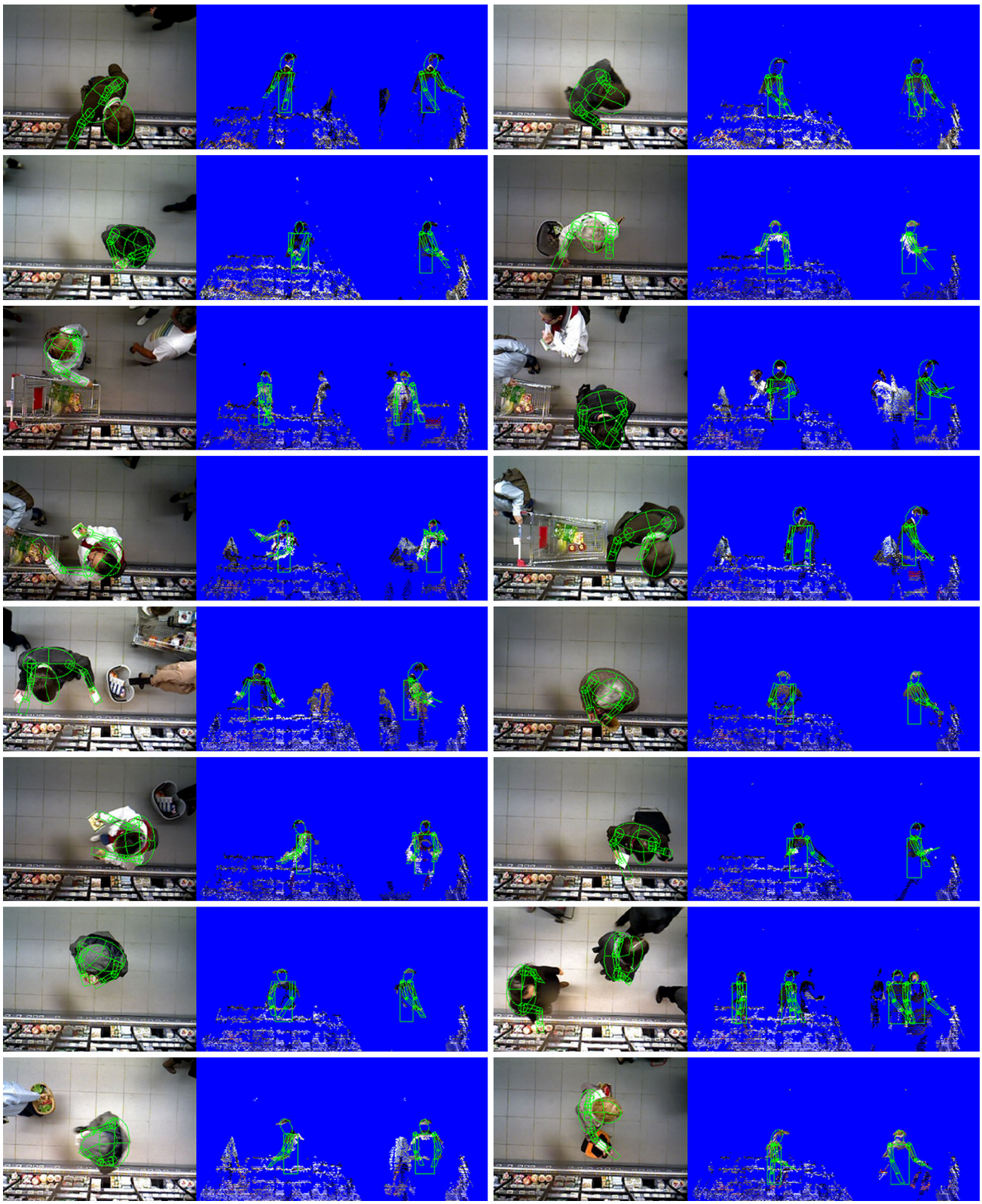

Fig. 19 Tracking on sequences recorded on site in a supermarket

that our method provides a correct pose estimation in all these hard situations. These sequences are not available because we are awaiting regulatory approval for CNIL.

The processing is real-time when the number of customers is not too high (that represents the most of the cases). Otherwise, the tracking of each target could be parallelized. Moreover, we have noticed experimentally that the pose of the person is well estimated with a process on only 7 frames per second.

6.6 Perspective for the action recognition

The ultimate aim of the project is to obtain a behavior analysis. The buying acts have to be recognized. That is the mean perspective of our work and it will be developed later. However, to validate our work, we have started considering how it could be used to recognize buying acts.

We have isolated some particular buying acts: catching an object in the shelves and looking at handhold goods (to read the label or research the cost). Then we have studied the evolution of the tracking obtained by our method on these acts so as to find the most well-descriptive configuration for each of them.

For the object catching, the 3D trajectories of the wrist and the elbow are the most descriptive features (Fig. 20). Indeed the regular curves of these body parts are always noticed for this action. The velocity of the movement provides clues to recognize impulse and reasoned purchase.

For the goods looked at, the space given by the 10 angles of the arms are the most descriptive feature (Fig. 21). Generally, it corresponds to a pose and is roughly constant over time even if the person keep walking when he looks at the goods. The duration of the this pose determines the time the customer has studied the goods and could infer doubts.

The definition of the various behaviors to recognize and the way to use our results are in progress.

Fig. 20 Examples of trajectories of the elbow and the wrist in the *YZ* space relatively to the shoulder for action of catching a good in the on-site sequences. The trajectories of the articulation points in the 3D space are well descriptive of this buying act

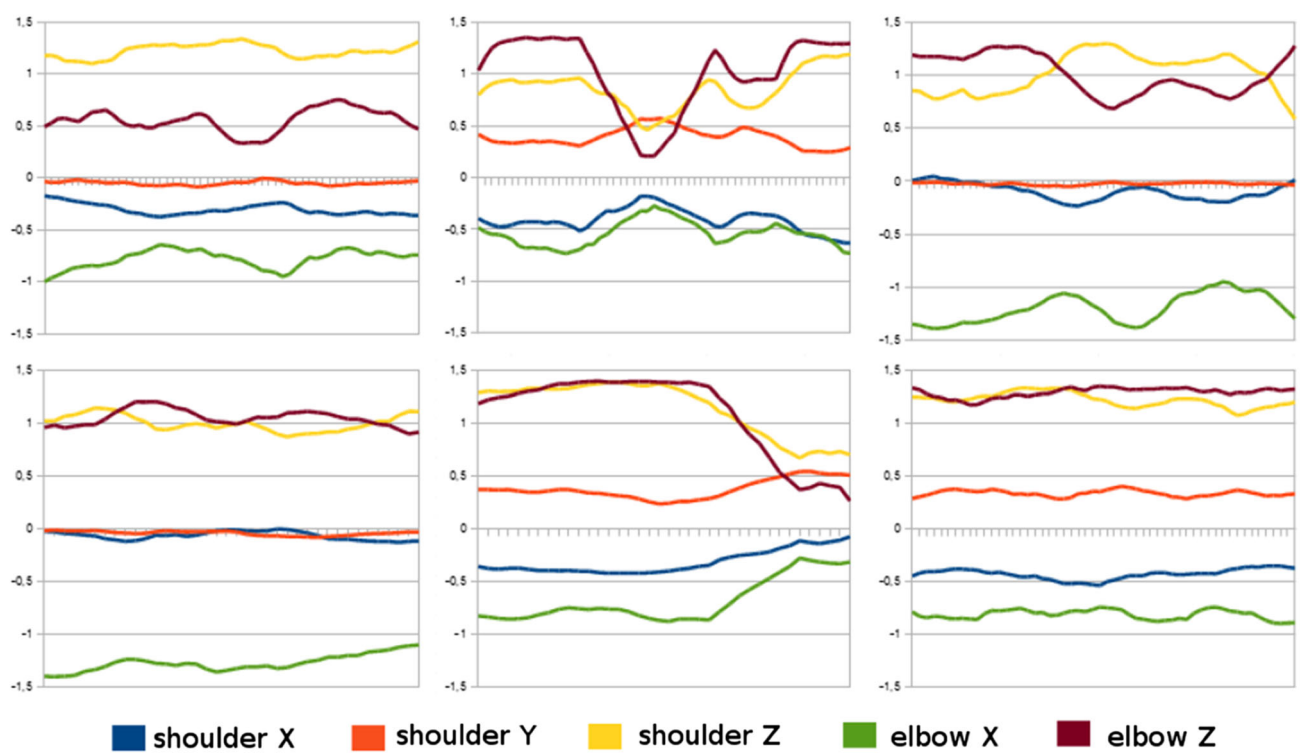

Fig. 21 Evolution of the angles of the arm model for various goods watching action in the on-site sequences: the space of the degrees of freedom of the arm is the best to describe this buying act

7 Conclusion

In this paper we have proposed a method for gesture tracking by particle filtering using the Xtion PRO-LIVE. We consider the particular case of the top view. The depth cue is hence needed in our process. To be efficient in real-time, we separate the body into two parts. The head and the shoulders are tracked in the 2D space of the recorded images and the arms are tracked in the 3D space. Experiments show that it is in these spaces that the tracking is actually the best for each of the two parts. They also confirm the effectiveness of our method in a real-time processing. Moreover, a confidence measure is associated with each frame so as to detect the possible wrong estimations. Finally, applying the exclusion principle makes possible the multi-person tracking process with no additional complexity.

Our algorithm is run efficiently in 25 ms per frame (i.e. at the frame rate of 40 fps and a throughput of 3.85 Mpps). That is suitable with a real-time operation and on-line applications.

In the future, the tracking will be inserted in an action recognition application. As we have seen in the Sect. 6.6, our method provides relevant information that has to be used for the complete buying act recognition process. The definition of the behaviors of interest will be the first step. Then descriptor should research how to best describe each of the action from our results. Finally, a classification will realize the recognition.

Acknowledgments This work is supported by project ANR-10-CORD0016 ORIGAMI2.

References

1. Andriyenko, A., Schindler, K.: Multi-target tracking by continuous energy minimization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1265–1272 (2011). doi:10.1109/CVPR.2011.5995311
2. Benfold, B., Reid, I.: Stable multi-target tracking in real-time surveillance video. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3457–3464 (2011). doi:10.1109/CVPR.2011.5995667
3. Ben Shitrit, H., Berclaz, J., Fleuret, F., Fua, P.: Tracking multiple people under global appearance constraints. In: IEEE International Conference on Computer Vision, pp. 137–144 (2011). doi:10.1109/ICCV.2011.6126235.
4. Berclaz, J., Fleuret, F., Turetken, E., Fua, P.: Multiple object tracking using K-shortest paths optimization. IEEE Trans. Pattern Anal. Mach. Intell. **33**, 1806–1819 (2011). doi:10.1109/TPAMI.2011.21
5. Botella, G., Martn H., J.A., Santos, M., Meyer-Baese, U.: FPGA-based multimodal embedded sensor system integrating low- and mid-level vision. Sensors. **11**, 8164–8179 (2011). doi:10.3390/s110808164
6. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Online multiperson tracking-by-detection from a single, uncalibrated camera. In: IEEE Trans. Pattern Anal. Mach. Intell. **33**, 1820–1833 (2010). doi:10.1109/TPAMI.2010.232
7. Brox, T., Rosenhahn, B., Gall, J., Cremers, D.: Combined region and motion-based 3D tracking of rigid and articulated objects. IEEE Trans. Pattern Anal. Mach. Intell. **32**, 402–415 (2009). doi:10.1109/TPAMI.2009.32
8. Brendel, W., Amer, M., Todorovic, S.: Multiobject tracking as maximum weight independent set. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1273–1280 (2011). doi:10.1109/CVPR.2011.5995395
9. Canton-Ferrer, C., Salvador, J., Casas, J.R., Pardàs, M.: Multi-person tracking strategies based on voxel analysis. In: Multimodal Technologies for Perception of Humans, pp. 91–103 (2008). doi:10.1007/978-3-540-68585-2_7
10. Choi, W., Savarese, S.: Multiple target tracking in world coordinate with single, minimally calibrated camera. In: European Conference on Computer Vision, vol. 4, pp. 553–567 (2010). doi:10.1007/978-3-642-15561-1_40
11. Dantone, M., Gall, J., Leistner, C., Van Gool, L.: Human pose estimation using body parts dependent joint regressors. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3041–3048 (2013). doi:10.1109/CVPR.2013.391
12. Deutscher, J., Reid, I.: Articulated body motion capture by stochastic search. Int. J. Comput. Vis. **61**, 185–205 (2005). doi:10.1023/B:VISI.0000043757.18370.9c
13. Gonzalez, M., Collet, C.: Robust body parts tracking using particle filter and dynamic template. In: IEEE International Conference on Image Processing, pp. 529–523 (2011). doi:10.1109/ICIP.2011.6116398
14. Heath, K., Guibas, L.J.: Multi-person tracking from sparse 3D trajectories in a camera sensor network. In: ACM/IEEE International Conference on Distributed Smart Cameras, pp. 1–9 (2008). doi:10.1109/ICDSC.2008.4635679
15. Hauberg, S., Sommer, S., Pedersen, K.S.: Gaussian-like spatial priors for articulated tracking. In: European Conference on Computer Vision, pp. 425–437 (2010). doi:10.1007/978-3-642-15549-9_31
16. Horaud, R., Niskanen, M., Dewaele, G., Boyer, E.: Human motion tracking by registering an articulated surface to 3D points and normals. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 158–163 (2008). doi:10.1109/TPAMI.2008.108
17. Hu, Z., Wang, G., Lin, X., Yan, H.: Recovery of upper body poses in static images based on joints detection. Pattern Recognit. Lett. **30**, 503–512 (2009). doi:10.1016/j.patrec.2008.12.005
18. Isard, M., Blake, A.: CONDENSATION_conditional density propagation for visual tracking. Int. J. Comput. Vis. **29**, 5–28 (1998). doi:10.1023/A:1008078328650
19. Jiang, Z., Huynh, D.Q., Moran, W., Challa, S., Spadaccini, N.: Multiple pedestrian tracking using colour and motion models. In: IEEE International Conference on Digital Image Computing: Techniques and Applications, pp. 328–334 (2010). doi:10.1109/DICTA.2010.63
20. Khan, Z., Balch, T.R., Dellaert, F.: Efficient particle filter-based tracking of multiple interacting targets using an MRF-based motion model. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 254–259 (2003). doi:10.1109/IROS.2003.1250637
21. Kitagawa, G.: Monte Carlo filter and smoother for non-gaussian nonlinear state space models. J. Comput. Graph. Stat. **5**, 1–25 (1996). doi:10.2307/1390750
22. Kjellstrom, H., Kragić, D., Black, M.J.: Tracking people interacting with objects. In: IEEE Conference Computer Vision and Pattern Recognition, pp. 747–754 (2010). doi:10.1109/CVPR.2010.5540140
23. Larsen, A.B.L., Hauberg, S., Pedersen, K.S.: Unscented Kalman filtering for articulated human tracking. In: Scandinavian

Conference on Image Analysis, pp. 228–237 (2011). doi:10.1007/978-3-642-21227-7_22
24. Lee, M.W., Cohen I., Jung, S.K.: Particle filter with analytical inference for human body tracking. In: IEEE Workshop on Motion and Video Computing, pp. 159–165 (2002). doi:10.1109/MOTION.2002.1182229
25. Liem, M., Gavrila, D.: Multi-person tracking with overlapping cameras in complex, dynamic environments. In: British Machine Vision Conference, pp. 1–10 (2009). doi:10.5244/C.23.87
26. Luber, M., Spinello, L., Arras, K.O.: People tracking in RGB-D data with on-line boosted target models. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3844–3849 (2011). doi:10.1109/IROS.2011.6095075
27. MacCormick, J., Blake, A.: A probabilistic exclusion principle for tracking multiple objects. Int. J. Comput. Vis. **39**, 57–71 (2000). doi:10.1023/A:1008122218374
28. Micilotta, A.S., Bowden, R.: View-based location and tracking of body parts for visual interaction. In: British Machine Vision Conference, pp. 849–858 (2004). doi: 10.5244/C.18.87
29. Migniot, C., Ababsa, F.: Part-based 3D multi-person tracking using depth cue in a top view. In: International Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (2014)
30. Mitzel, D., Sudowe, P., Leibe, B.: Real-time multi-person tracking with time-constrained detection. In: British Machine Vision Conference, pp. 1–11 (2011). doi:10.5244/C.25.104
31. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.J.: You'll never walk alone: modeling social behavior for multi-target tracking. In: IEEE International Conference on Computer Vision, pp. 261–268 (2009). doi:10.1109/ICCV.2009.5459260
32. Pilu, M., Fitzgibbon, A.W. and Fisher, R.B.: Ellipse-Specific Direct Least-Square Fitting. IEEE Int'l Conf. on Image Processing. 3: 599–602 (1996). doi:10.1109/ICIP.1996.560566
33. Pirsiavash, H., Ramanan, D., Fowlkes, C.C.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1201–1208 (2011). doi:10.1109/CVPR.2011.5995604
34. Ristic, B., Arulampalam, S., Gordon, N.: Beyond the Kalman Filter: Particle Filters for Tracking Applications. Artech House, London (2004)
35. Schwarz, L.A., Mkhitaryan, A., Mateus, D., Navab, N.: Human skeleton tracking from depth data using geodesic distances and optical flow. Image Vis. Comput. **30**, 217–226 (2012). doi:10.1016/j.imavis.2011.12.001
36. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. Mach. Learn. Comput. Vis. **411**, 119–135 (2013). doi:10.1007/978-3-642-28661-2_5
37. Shu, G., Dehghan, A., Oreifej, O., Hand, E., Shah, M.: Part-based multiple-person tracking with partial occlusion handling. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1815–1821 (2012). doi:10.1109/CVPR.2012.6247879
38. Song, B., Jeng, T.Y., Staudt, E., Roy-Chowdhury, A.K.: A stochastic graph evolution framework for robust multi-target tracking. In: European Conference on Computer Vision, pp. 605–619 (2010). doi:10.1007/978-3-642-15549-9_44
39. Stoll, C., Hasler, N., Gall, J., Seidel, H.-P., Theobalt, C.: Fast articulated motion tracking using a sums of gaussians body model. In: IEEE International Conference on Computer Vision, pp. 951–958 (2011). doi:10.1109/ICCV.2011.6126338
40. Thiel, E., Montanvert, A.: Chamfer masks : discrete distance functions, geometrical properties and optimization. In: IAPR International Conference on Pattern Recognition, pp. 244–247 (1992). doi:10.1109/ICPR.1992.201971
41. Viola, P.A., Jones, M.J.: Robust real-time face detection. IEEE Int. J. Comput. Vis. **57**, 137–154 (2004). doi:10.1023/B:VISI.0000013087.49260.fb
42. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. Int. J. Comput. Vis. **75**, 247–266 (2007). doi:10.1007/s11263-006-0027-7
43. Xing, J., Ai, H., Lao, S.: Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1200–1207 (2009). doi:10.1109/CVPR.2009.5206745
44. Xtion PRO-LIVE. http://www.asus.com/Multimedia/Xtion_PRO_LIVE/
45. Yang, B., Nevatia, R.: An online learned CRF model for multi-target tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2034–2041 (2012). doi:10.1109/CVPR.2012.6247907
46. Yang, C., Duraiswami, R., Davis, L.S.: Fast multiple object tracking via a hierarchical particle filter. IEEE Int. Conf. Comput. Vis. **1**, 212–219 (2005). doi:10.1109/ICCV.2005.951
47. Zhang, Z., Hou, Y., Wang, Y., Qin, J.: A traffic flow detection system combining optical flow and shadow removal. In: IEEE Conference on Intelligent Visual Surveillance, pp. 45–48 (2011). doi:10.1109/IVSurv.6157021

Cyrille Migniot received his Ph.D. degree in Image Processing from the University of Grenoble (France) in 2008. He was a post-doctoral researcher at the IBISC Laboratory, Evry (France) in 2012. He is now holding a post-doctoral position in INRIA, Grenoble (France).

Fakhreddine Ababsa received his Ph.D. degree in Robotics from the University of Evry Val d'Essonne (France) in 2002. Since 2004, he is assistant professor in electrical and computer sciences at the University of Evry Val d'Essonne and researcher at the IBISC Laboratory (ex. Complex System Laboratory). His current interests are focused on robust estimation, motion tracking, human-machine interaction and sensor fusion with applications to scientific problems, in particular the development of real-time augmented reality systems.