Hindawi
Journal of Control Science and Engineering
Volume 2018, Article ID 2605735, 10 pages
https://doi.org/10.1155/2018/2605735

Research Article

A Study of Chained Stochastic Tracking in RGB and Depth Sensing

Xuhong Liu and Shahram Payandeh

Networked Robotics and Sensing Laboratory, School of Engineering Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada V5A 1S6

Correspondence should be addressed to Xuhong Liu; xuhongl@sfu.ca

Received 21 July 2017; Accepted 19 September 2017; Published 30 January 2018

Academic Editor: Enrique Onieva

This paper studies the notion of hierarchical (chained) structure of stochastic tracking of marked feature points while a person is moving in the field of view of a RGB and depth sensor. The objective is to explore how the information between the two sensing modalities (namely, RGB sensing and depth sensing) can be cascaded in order to distribute and share the implicit knowledge associated with the tracking environment. In the first layer, the prior estimate of the state of the object is distributed based on the novel expected motion constraints approach associated with the movements. For the second layer, the segmented output resulting from the RGB image is used for tracking marked feature points of interest in the depth image of the person. Here we proposed two approaches for associating a measure (weight) for the distribution of the estimates (particles) of the tracking feature points using depth data. The first measure is based on the notion of spin-image and the second is based on the geodesic distance. The paper presents the overall implementation of the proposed method combined with some case study results.

1. Introduction

In this paper, a framework is proposed which can be used to explore the information flow and sharing in a distributed Bayesian tracking using both RGB and depth sensors. The proposed hierarchical (cascaded) particle filter first tracks the human body in the RGB image by exploiting the notion of importance sampling [1], taking into account the physical motion constraints. The information regarding the tracked body obtained from the first layer is then utilized in the second implementation of particle filter using depth image. In this implementation, expected sample distribution for tracking points of interest on the body also takes advantage of the constraints associated with the body movements within the segmented depth image. In addition, we have experimented with two approaches in assigning weight to each sample distribution within the second particle filter implementation. The first metric measure is based on the notion of spin-image at the desired point of interest on the tracked body. The second metric is based on the notion of geodesic distance between a reference point and the desired point of interest on the body.

Tracking the overall movements of the human body combined with tracking specific points of interest located on the tracked body has many applications. These applications can range from virtual/augmented reality (V/AR), surveillance, and motion analysis to human-robot/environment interaction. However, there also exist various challenges associated with tracking when using various types of ambient sensors. Firstly the human body shape and movements are highly variable and the body parts have a number of degrees of freedom. Secondly the tracking environment is usually very complex and can be under different illumination and background conditions. Such environment is a common source of ambiguity which would influence the stability of only RGB based tracking. In our study, we have taken advantage of combined depth and RGB sensors. Time-of-flight sensors are able to provide dense depth measurements at high frame rate, for example, Microsoft Kinect [2].

In tracking the overall coarse shape of gait, [3] uses a novel projected top view of the occupied volumes (virtual top view (VTV)) of the monitoring area. Through segmentation of the VTV, a bounding box can be defined which encloses each person and can be extended to corresponding bounding volume. A particle filtering approach is presented in [4] for 6-DOF object pose tracking using an RGB-D camera. An approach is introduced in [5] that can detect and track people in indoor spaces from a mobile platform without instrumenting the environment using RJ-MCMC particle filter. A novel 3D people detection and tracking approach in RGB-D data is proposed in [6]. The authors combined an online learning of target appearance models using three types of RGB-D features with multihypothesis tracking. However, they primarily detect a region of interest (ROI) of tracking people and some feature points with predefined models.

There also exists a large body of literature aiming at tracking selected limbs of human body. Majority of these approaches model the whole body as articulated interconnected segments. However, existing drawbacks in using articulated model are the high dimensionality of the configuration space and the exponentially increasing computational cost. There are a number of effective 2D tracking methods which have been proposed [7, 8]. These are in general used for applications such as surveillance which do not provide information regarding 3D kinematic model reconstruction.

In the more recent works, some researchers tended to combine different features which can complement each other in order to implement robust tracking. For example, Xu et al. in [9] proposed a novel tracker which extracted both gray and color information as the feature maps to compute the maximum response location via correlation filters. The KCF tracker, which is trained using a single image patch x with size $M \times N$ centered around the target, is utilized for tracking. Another tracking algorithm based on multiple features with an improved scale-updating scheme is proposed in [10]. They integrated HoG, color naming, and intensity feature. Kernel methods on the basis of the STC algorithm are used to fuse these features to implement tracking. The region of object is represented by a bounding box and the experimental results demonstrate that it is promising for various scenarios. Moreover, in [11] Han et al. presented an adaptive multifeature representation for visual tracking. More specifically, they exploited the internal relationship among three complementary features, that is, HoG, color naming, and LBP, by incorporating the idea of cotraining to build an efficient correlation filter framework which is used for tracking. In [12], the author focused on the issue regarding multispeaker tracking by jointly exploiting auditory and visual features in their feature spaces. The visual observation they used is a combination of bounding box provided by a head detector and an audio observation consisting of binaural features extracted from two-channel audio recordings.

This paper proposes a framework for implementing levels of detail in tracking based on chained particle filter. Particle filter (PF) uses a dynamic model to guide the propagation of the state estimation within limited subspace of target measurement [13]. This method provides a robust Bayesian framework for sensor-based tracking of human motion. First we propose an enhanced particle filter implementation based on RGB frames. It offers a faster convergence of the estimate where at each sample the prior distributions of the particles are defined based on the physical constraints associated with the expected movements of the tracking body. A 3D bounding box is then used as a coarse level of detail to represent the location and coarse spatial range of the human body movements for gait analysis. Our second contribution is cascading information from the previous region of interest of the moving body in a form of 3D box with another depth-based particle filter to track points of interests on human body.

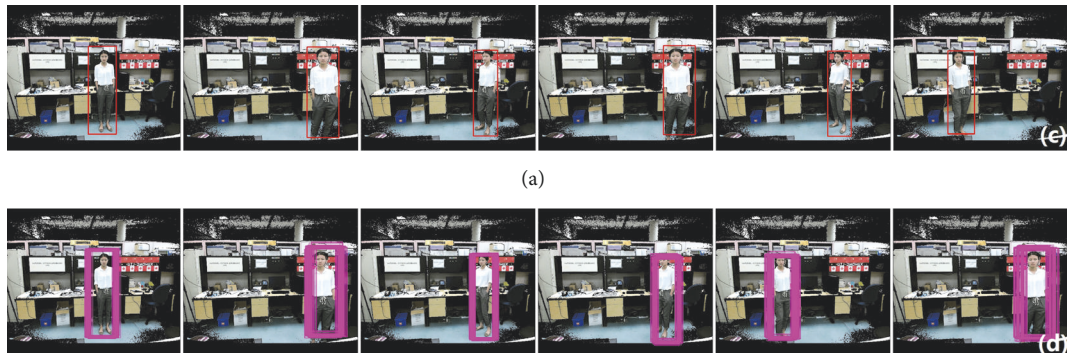
Similar hierarchical framework has been proposed in [14] for human pose recognition from single depth images in [15]. Such techniques, for example, [14], have been applied to tracking faces in low frame-rate video. Our approach incorporates two different types of particle distributions in order to acquire a more consistent result for both tracking coarse shape and points of interest in a real-time tracking system. The work in [15] has proposed a method to identify body extremities by adopting a measurement called geodesic distance. In this paper we have utilized such measure as a weighting factor in sample distribution [16], while at the same time exploring and studying another weighting measure entitled spin-image.

2. First Layer: Tracking in RGB Image

Our implementation of the levels of detail consists of two main cascaded layers of particle filter. The first layer is an enhanced color-based rectangular region tracking and the second is a depth-based particle filter tracking of selected feature points in the body within the bounding volume obtained from the first level. The experimental set-up uses both the color and depth sensors from a single Microsoft Kinect II. The sensor is positioned on the top of a tripod and directly facing the background as well as the person to be tracked. An adult is asked to walk under normal illumination condition and in a natural cluttered environment. Simultaneously, the color and the depth video sequences are captured by Kinect II sensor. The hierarchical tracking is implemented in C++ and runs for single target on a PC with Intel 3.20 GHz CPU.

At the initialization stage, we synchronize RGB frame and depth frame. However, the original frame size of color frame is 1920×1280 pixels, which is different from the depth frame with size of 512×424 pixels. We used calibration functions to map these two coordinate systems into a single one. As a result, a combined 3D synchronized frame is generated which is used in the remaining computation. After such synchronization, each pixel in the final RGB-D frame is associated with both its depth distance value and corresponding RGB color values.

For the first level, we extend the results of [17], which presented a color-based particle filter (CPF) to track a person within a bounding box. CPF performed well since it can capture majority of implementation uncertainties. However, less likely particles are not discarded immediately. On the contrary, they are given some prior weight which can be used in the subsequent steps. This step will result in an added



(a)

(b)

Figure 1: (a) Result of tracking a walking person using the CPF in frame number is 70, 75, 80, 95, 115, and 120. (b) Samples propagated using 100 particles.

computational cost since the evaluation of the likelihood function must be performed at every instance for every sampled particles [1]. This is one of the main reasons for increase in the computational cost in a typical implementation of PF such as condensation algorithm [18]. In this paper, implementation of the CPF in the first level is improved through hierarchical sampling in order to reduce the number of particles needed. To achieve this objective, after initialization of the target region, the region is represented by system matrix A and vector $\mathbf{M}_t$ represented by an 8-tuple state vector $\{x_t, y_t, x_t', y_t', w_t, h_t, w_t', h_t'\}$. In the state vector description, (x_t, y_t) is the center of the region at time t; (x_t', y_t') are the velocities of the box moving in the directions of axes u and v, respectively. (w_t, h_t) are the associated width and height of the box at time t; (w_t', h_t') are the instantaneous changes of the width and height at time t. If the state of the region is known in the current frame, we can obtain it in the next frame. The original state of the bounding box is manually initialized in the first frame of the tracking process. Here in this method a first-order autoregressive dynamic model is constructed to represent the propagation of the state vector:

$$\mathbf{M}_t^i = A\mathbf{M}_{t-1}^i + n_{t-1}^i, \quad n_{t-1}^i \sim \tau_i N\left(\mu_i, \sigma_i^2\right). \tag{1}$$

$\mathbf{M}_{t-1}$ is a part of the recursive estimation computed from the previous time instance. The propagation is also used to solve the problem of sample lost diversity. Hence, an added random value n_{t-1}^i is introduced at each sample around the state of the current instance in order to predict the state at the next instance. $n_{t-1}^i \sim \tau_i N(\mu_i, \sigma_i^2)$ show that this value follows a standard distribution where the mean is μ_i and variance is σ_i^2. Parameter A is the deterministic component of the state model. Both A and n_{t-1} can be modeled based on the knowledge of the scene and the target being tracked.

Since the incremental velocity and direction of a person can be estimated through the movement dynamic model, it can also be utilized to enhance the prior distributing of particles, that is, in the expected direction of movement. More precisely, defining the relationship between velocity, position variation, and property of bounding box is the main idea of the hierarchical sampling which is utilized in this paper. The view of the tracking area is based on perspective geometry. In general and in projected geometry, various movements and activities of tracking human can be regarded as a combination of two different types of motion cases, that is, movements along horizontal (x-direction) and vertical direction (y-direction).

For example, if the person is moving along x direction in the world frame (Figure 1) and by assuming that there is no change in the viewpoint, both the width and height definitions of the ROI remain unchanged. This is due to the fact that the height and width will only change as a function of how closer or further away the subject gets with respect to the camera. Since in this case the subject is moving along the x direction, the height and width of the projected object remain nearly unchanged. Based on this expected observation, more particles should be generated and propagated along the x-direction in order to estimate the variation of the position of the bounding box, along the direction of movement. Since the velocity vector in x and y directions is defined as the state vectors x_{t-1}' and y_{t-1}', the incremental direction of movements can be deduced. After determining the direction of motion of the previous time instance, an increased proportion of the total number of particles can be generated along this direction. This approach for sampling can result in a practical guideline without loss of diversity since the probability of people moving incrementally along a direction is much higher than sporadic movements to some other directions. The previous published methods [17] required generation of a large number of particles which results in the loss of consistency of the human motion estimation between each frame and in general results in a lost tracking. Similar observation can be utilized for the case that the person is moving in the z-direction of the world coordinate.

In the most general case, the person's movement is a combination of the above two cases. A relationship between parameters $\{w, h, w', h'\}$ and $\{x, y, x', y'\}$ can be established through definition of state propagation matrix A [16, 19] which can result in the following relationship:

$$\begin{aligned} w_t' &= w_{t-1}' + a\left(x_{t-1}'\right)^{-1} + b\left(y_{t-1}'\right), \quad a, b \sim N\left(\mu_i, \sigma_i^2\right) \\ h_t' &= h_{t-1}' + a\left(x_{t-1}'\right)^{-1} + b\left(y_{t-1}'\right), \quad a, b \sim N\left(\mu_i, \sigma_i^2\right), \end{aligned} \tag{2}$$

where a follows a standard distribution where both the mean and variance equal 1. Here x'_{t-1} represents the velocity along x direction and y'_{t-1} represents the velocity along y direction at time $t-1$. The greater the velocity in x direction changes, the smaller the width and height vary. Therefore we selected an inverse of its distribution in this function. In contrast, the greater the velocity in the y direction changes, the greater the width and height vary. In this way the previous velocity can be used as a prior knowledge to distribute these particles. Figure 1 shows a sample tracking result for this layer of implementation.

3. Second Layer: Tracking in the Segmented Depth Image

The second layer of tracking consists of several major steps. The segmented input depth data inside ROI is first filtered in order to reduce the dominant noise in the data and to obtain consistent surface point cloud. Here we use nearest-neighbor interpolation algorithm [20] and a median filter with a 3×3 window is used throughout the depth frame. The basic principle of this filter is to replace the value of a certain point with median of its neighboring points. This method uses a two-dimensional window which goes point by point in the whole depth image. Next the human body is further segmented and two representative points are selected in the initial frame on the surface of the human body, for example, p_r and p_t.

The integrated depth segmentation algorithm is demonstrated in the following. Given the bounding box which is acquired to represent the location and coarse spatial range of the human body in the previous layer, the foreground segmentation fully utilized this result and incorporated it with depth information to decrease the computation cost. The key idea of this step is designed to check the depth continuity of neighboring pixels and return all the separated depth clusters inside the ROI. To start with, we performed the depth-first searching (DFS) algorithm to these points and identified the largest depth cluster inside the bounding box area, which is considered to be the human body. DFS is an algorithm which can find the largest connected component in an undirected graph. So after running DFS algorithm, we label each pixel which belongs to the background as 0 and label those pixels as a whole belonging to the human body surface as 1 (Figure 2). This process is iterated in the beginning of preprocessing the depth frame at each time state.

3.1. Human Body Division. In order to enable further appearance-based body part matching between successive occurrences of the tracked person, we extracted surface mesh on a local patch of point cloud. This body division method is explored in order to generate local mesh on the patch including the extremity regions. We initiate the mesh generation by first detecting the head region. This can be done by first finding the minimum width associated with the silhouette of the body. By assuming the natural position of the head region of a walking person, we first deduce the searching area to be the top 1/3 of the entire segmented region. We use the silhouette width along the horizontal direction, to generate the human body silhouette width curve. The variation in the silhouette width curve, representing a silhouette histogram, is shown in the middle column of Figure 3. For example, we can start from the top pixel location of the head region and continuing scan on the horizontal direction until we find the position of a local minimum; it gives an indicator of approaching the neck, which also should be the bottom of the head region.

The computed pixel location in the depth image is defined with respect to a local sensor coordinates (x, y, z) which needs to be mapped to the (X, Y, Z) defined in world coordinates in order to construct the 3D surface mesh of human body. Using the intrinsic parameters of the depth sensor, a 3D point (X, Y, Z) can be calculated from the depth image (x, y, z) using the simplified pinhole camera model shown as follows:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

$$\Downarrow$$

$$X = \frac{Z(x - c_x)}{f_x} \tag{3}$$

$$Y = \frac{Z(x - c_y)}{f_y}$$

$$Z = \text{depth},$$

where (x, y) are pixels of depth image, Z is the distance between Kinect and object, (f_x, f_y) are the focal length parameters of the depth sensor, and (c_x, c_y) are the principal point offset parameters, Figures 4(a) and 4(b).

3.2. Mesh Generation. Using the definition of the 3D bounding box coordinate system, a polygonal surface mesh can be generated as a 3D undirected graph. The undirected graph in this depth segmentation algorithm is defined as $G_h = (V_h, E_h)$, consisting of the set V_h of nodes and the set E_h of edges, which are unordered pairs of elements of V_h. In order to transform a depth clustering of points into a surface mesh, each point becomes a vertex of the graph and edges are created from these vertices. This process goes through all the vertices in the biggest depth clustering and checks each pair of neighboring vertices separately. For each two vertices (p, q) inside the bounding box area, an edge is generated between them if and only if these two rules are both satisfied: (a) their corresponding pixel location is within the segmented depth image within the bounding box coordinate system and (b) their 3D Euclidean distance $d(p, q) = \|p - q\|$ does not exceed a predefined threshold. This threshold is set in order not to connect two points which do not belong to the same object but are next to each other in the image plane, for example, the pixels in the contour of the foreground and their neighboring pixel which belongs to the background. Here the threshold is set as 5 mm, Figure 4(c).

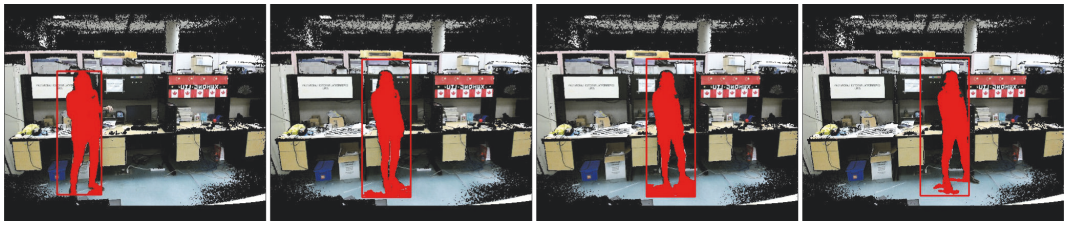


Figure 2: An example sequence associated with two-stage depth segmentation, that is, background subtraction and DFS in order to find the largest connected component.

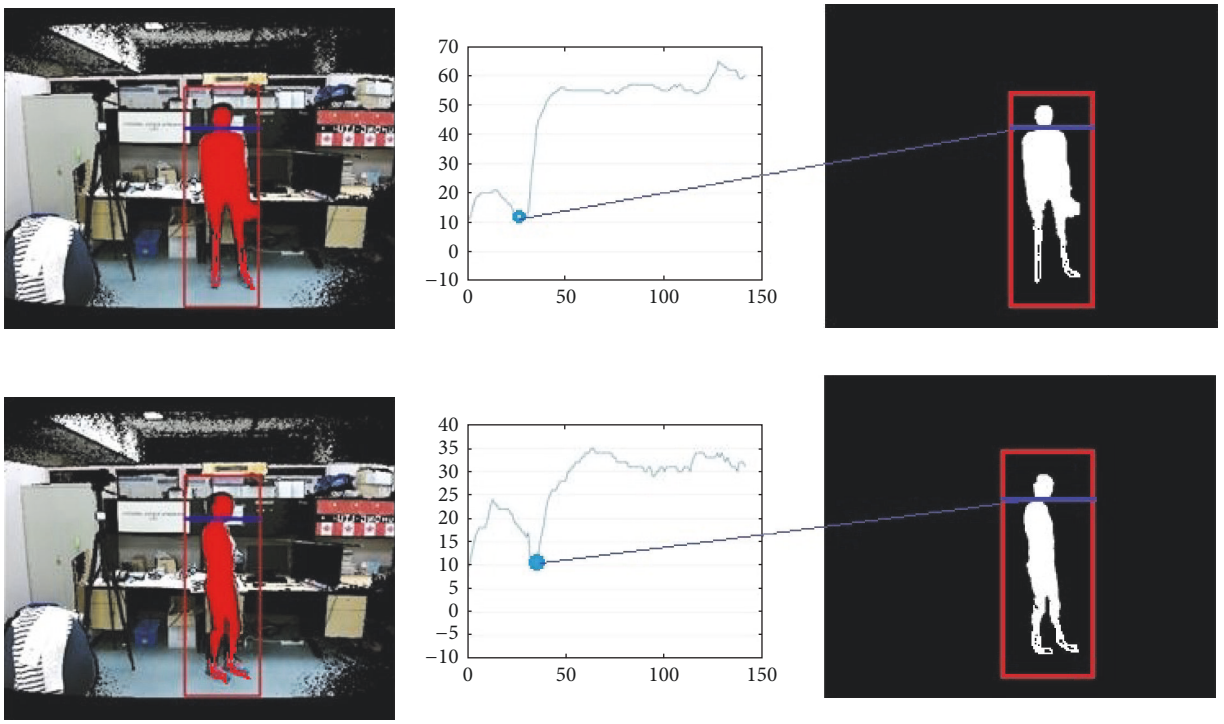


Figure 3: Human head region detection.

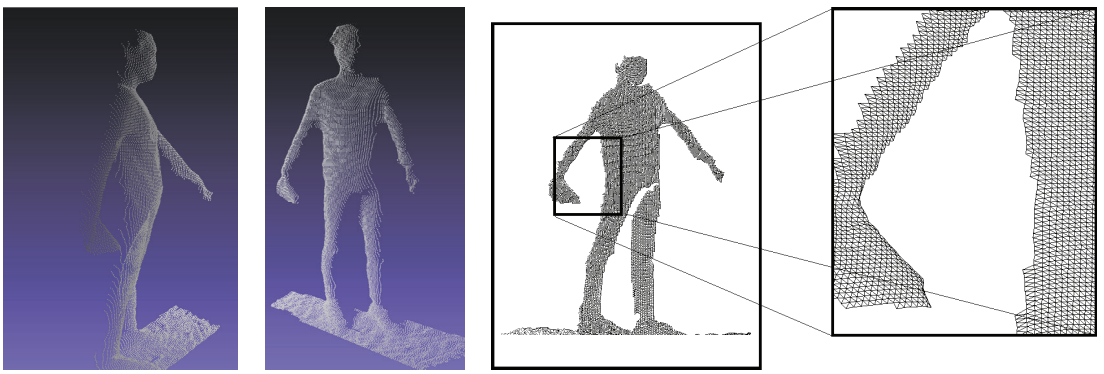


Figure 4: (a-b) Different views of calibrated point cloud of human body in world space; (c) mesh generation on the segmented and calibrated point cloud.

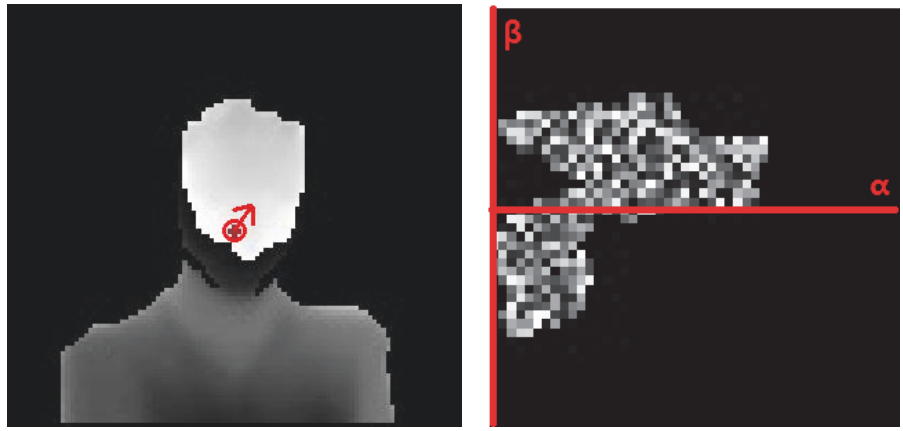


Figure 5: An example of spin-image generated from local depth patch on the segmented body.

3.3. DPF Tracking Using Spin-Image Weighting. The principle behind the depth-based particle filter (DPF) is similar to the implementation in the first layer. Here the idea is to track points of interest which are initially defined on the point cloud of the segmented body, for example, p_r and p_t. Here, at each frame, samples are generated and propagated which are weighted based on the notion of spin-image. Then, the new estimate is defined based on this weighted contribution of the samples.

The spin-image associated with a certain point of a 3D object is used as a reference to weigh samples in the DPF [21]. Spin-image is a 2D image which can capture the information around the neighborhood of a point. An oriented point, which is a static version of oriented particles, is used to generate spin-image of a 3D object. The oriented basis is defined with respect to a basis point as $O = (\mathbf{p}, \mathbf{n})$, where $\mathbf{p}$ is the position of the basis point and $\mathbf{n}$ is the surface normal of the point [22]. This definition transforms the position of each point from a three-dimensional coordinate system into a new position in a two-dimensional coordinate system, Figure (4).

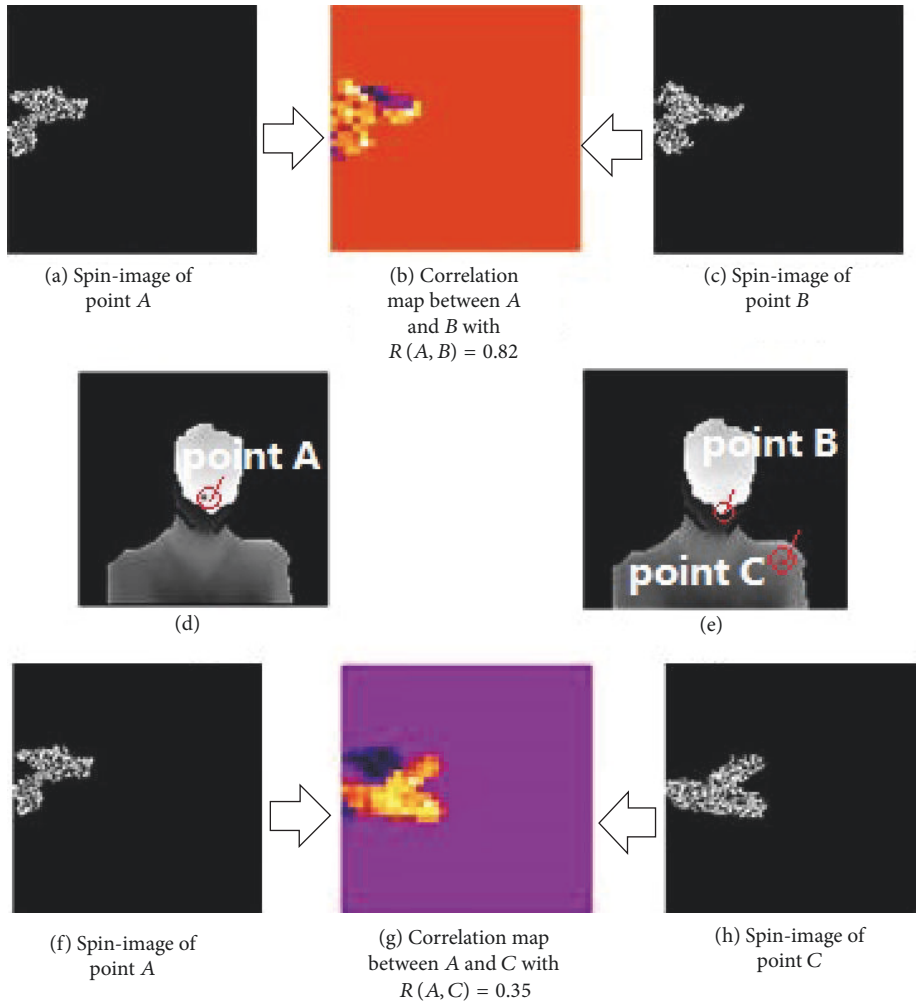
$$S_0(\mathcal{X}) \longrightarrow (\alpha, \beta) = \left(\sqrt{\|\mathbf{x} - \mathbf{p}\| - \mathbf{n} \cdot (\mathbf{x} - \mathbf{p})^2}, \mathbf{n} \cdot (\mathbf{x} - \mathbf{p}) \right). \quad (4)$$

In this equation, $S_0(\mathcal{X})$ is the spin-image map while $\mathbf{x}$ is a 3D point. Two axes in spin-image coordinate are α and β where α represents the perpendicular distance of any other point in point cloud to basis normal and β is signed perpendicular distance of it to basis plane [21]. An example of generating the corresponding spin-image is shown in Figure 5.

In order to measure similarity between images, correlation coefficient is defined and utilized by [22]. Given two spin-images, namely, A and B (with N bins), the correlation coefficient can be calculated through measuring normalized error in each bin of the two images. Correlation coefficient $R(A, B)$ varies from −1 to 1 where −1 implies no correlation and 1 implies complete correlation. Figure 6 illustrates how spin-image which is generated from two different points in the same surface mesh is compared. In Figures 6(a), 6(d), and 6(f), an oriented point A is selected from human body surface mesh with its associated spin-image. In Figures 6(c), 6(e), and 6(h), two different oriented points are selected from the same human body surface mesh with their associated spin-images. Point A and Point B are in similar positions on the surface mesh, whereas point A and point C are in relatively different positions. Thus the spin-images of point A and point B are similar, as is shown by the correlation map in the images. The correlation maps (see Figures 6(b) and 6(g)) are created by plotting the pixel values in one image versus the corresponding pixel values in the other image. This is an effective method of visualizing whether two images are correlated. For points A and B, their spin-images correlation map shows large region of similarity. However, for points A and C, since they come from positions which are not similar, their spin-images are also not similar. The correlation map of their spin-image has relatively small region of overlap and thus shows less similarity.

Each particle sample in the DPF of this level is represented by a 3D point, whose state vector, $\mathbf{P}_t$, is defined as $\mathbf{P}_t = \{x_t, y_t, z_t, x_t', y_t', z_t'\}$, in which (x_t, y_t, z_t) is the location of the 3D point at time t and (x_t', y_t', z_t') is its velocity. At the beginning of each time instance, we first extract all the point locations which belong to the human body. Thus it is possible to figure out whether a generated sample is on the body surface or not. Similar to the previous level, the set of samples is propagated based on the system dynamic model and the movement patterns. At every time instance, points are generated mainly based on the direction of movement of the previous time instance and each sample point is weighted by the correlation coefficient between the spin-image of the sample and the spin-image of the reference point. Samples with the highest weight will be taken as the possible state of the target point. These samples are then propagated and updated to estimate the state at next instance.

Figure 7 shows tracking results of the proposed hierarchical implementation of particle filter using spin-image for its DFP. In (a), the red rectangle shows the tracking results of the CPF obtained from first level. The red solid circle represents the tracking results of the DPF. This point is first initialized on the forehead and then tracked through each



(a) Spin-image of point A

(b) Correlation map between A and B with $R(A, B) = 0.82$

(c) Spin-image of point B

(d)

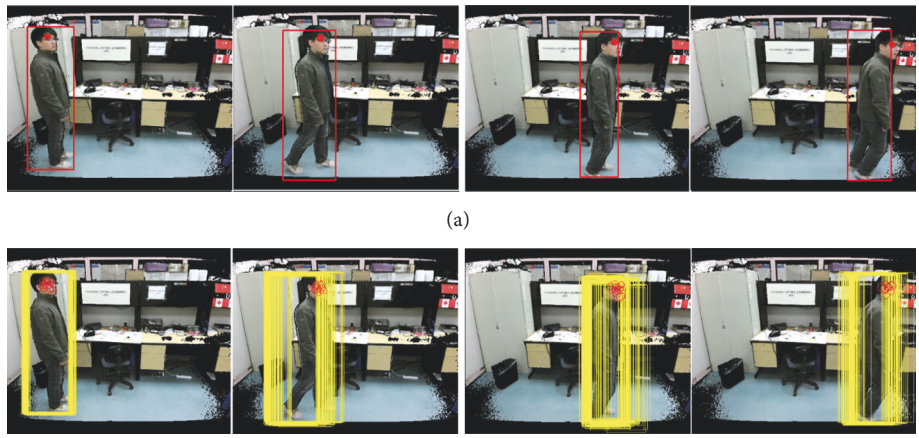
(e)

(f) Spin-image of point A

(g) Correlation map between A and C with $R(A, C) = 0.35$

(h) Spin-image of point C

Figure 6: Comparison of spin-images using correlation map between spin-images of similar and unsimilar points. (a) Spin-image generated from point A; (b) spin-image correlation map (A-B); (c) spin-image generated from point B; (d) the position of point A which is indicated in the figure; (e) the positions of point B and point C which are shown in the figure; (f) spin-image generated from point A; (g) spin-image correlation map (A–C); (h) spin-image generated from point C.



(a)

(b)

Figure 7: Example of the hierarchical implementation of particle filter. (a) State estimation of the designated point of interest on the body (solid circle). (b) Generated samples (circle point) propagated by the proposed DPF.

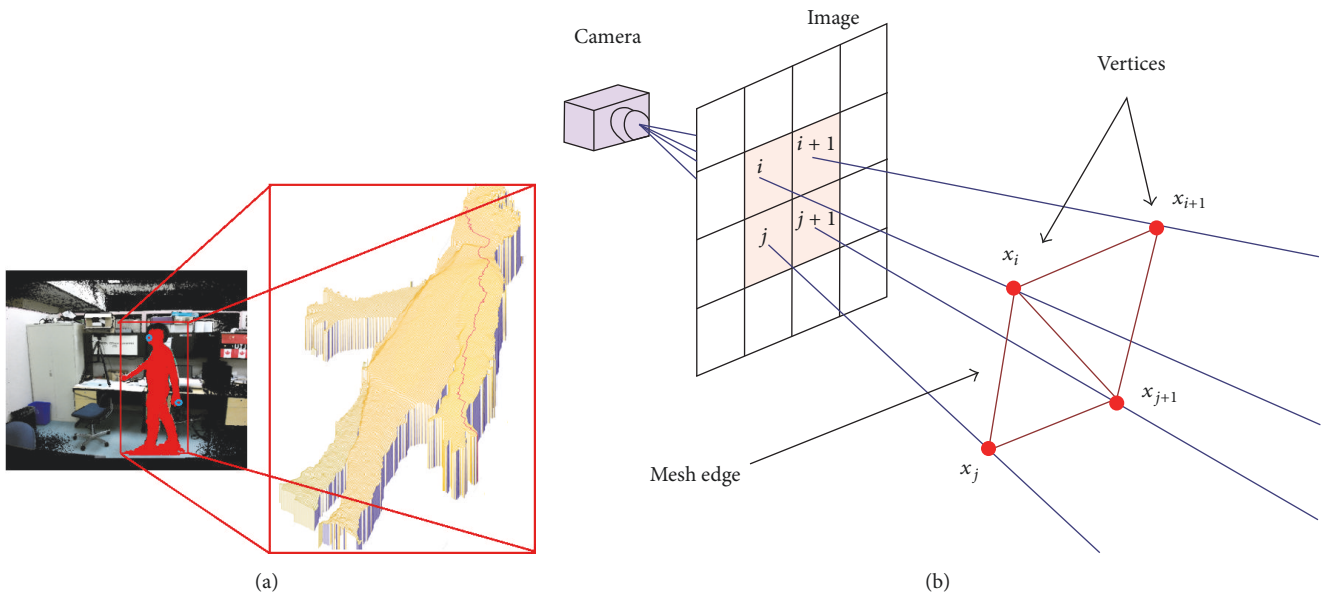


FIGURE 8: (a) Visualization of geodesic distance computation from the point cloud. (b) Example of surface mesh generation using adjacency relationships.

frame. (b) of the figure shows example of samples propagation following our proposed methods. From the results it can be seen that the whole system can track both coarse body area and a designated feature point of interest on the human body during the movement.

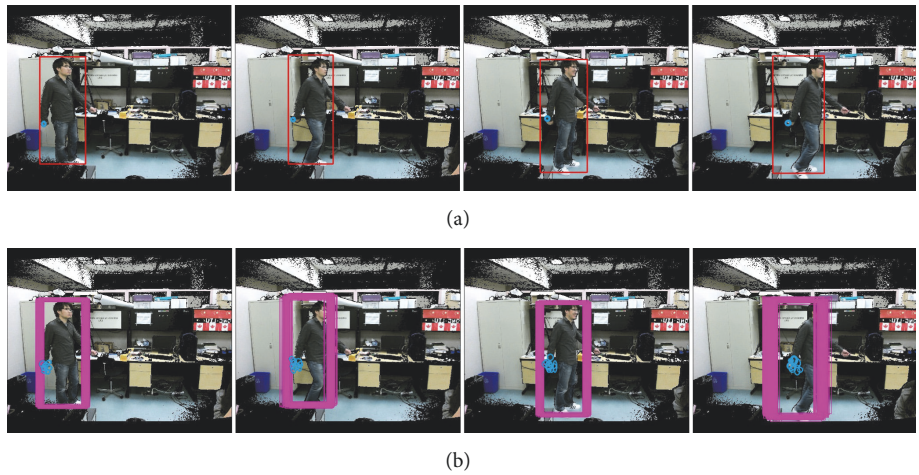
3.4. DPF Tracking Using Weighted Geodesic Distance. In the previous section, given a point of interest on the body, we utilized spin-image in order to associate weight with the sample distribution. The objective of the second method of implementation of DPF is to defined and another approach for associating weight with the sample distribution is studied. Here, the desired feature points on the tracked body are first mapped to the vertices of the constructed surface mesh. During the tracking the distance between these points of interest will remain unchanged along the surface mesh. Hence, during the sample propagation, one can associate a weighting factor with the sample distribution on how much they deviate from the reference geodesic distance that is calculated at the initial time state of the tracking process. Such approach can result in a method robust against mesh deformations, translations, and rotations. Traditional local color-based approaches for defining features are very sensitive to such local deformation. Model-based tracking in defining features are also very restricted mainly due to their high computational cost since the human body can be modeled as an articulated object with high degrees of freedom. Since Euclidean distance between two feature points can vary widely with body movement in 3D space (also being inspired by the concept of Accumulative Geodesic Extrema [15]), we utilized geodesic distance. Geodesic distance between two points on the body, for example, the distance from the nose of a person to the right hand along the body surface, is relatively invariant and independent of different postures.

Constructing surface mesh from point cloud of the whole body allows us to measure geodesic distances between any feature points selected on the body. Geodesic distance [23] is defined as the number of edges in the shortest path connecting two vertices in a graph. Dijkstra's algorithm [24] is performed to compute the geodesic distance between, for example, a reference point and all of the generated particles during the sample propagation process. Figure 8 shows an example of geodesic distance. In this figure, the yellow grids show the reconstructed surface mesh of the extracted human body. The blue circles in this figure show the reference point and the tracking point of extremity. The red lines between these two circles show the geodesic distance along the surface mesh.

Figure 9 shows a sample result associated with hierarchical implementation of particle filter where geodesic distance measure was used as a weighting factor on the sample distribution. In (a), the blue circle represents the result of tracking of a point of interest located at the hand of the subject. (b) demonstrates how the samples are propagated and in particular circles correspond to the sample propagation at the DPF layer.

4. Discussions and Conclusions

In this paper, we proposed an approach for tracking movements of a person in a cluttered environment. The method is based on the notion of a hierarchical particle filter which incorporates two layers consisting of coarse-to-fine tracking subsystems. In the first layer and by considering the computational time needed to converge to the true state, we proposed a sequential approach by defining importance sampling. This method is implemented by modeling the relationship between the movement of the person and method



(a)

(b)

Figure 9: (a) Estimated state of the extremity point (solid circle). (b) Samples (circle point) propagated by the cascaded PF.

of populating the particles in the system dynamic model. In the preprocessing stage of the second layer, we synchronized depth and color frames, extract the human body inside the bounding box, and construct a surface mesh from it. In this layer, we also proposed and utilized two types of measures to associate and weight the sample propagation in the depth-based particle filter implementation for tracking points of interest on the body.

In the second step of our cascaded framework we attempted to use two different features to implement the depth-based tracking. Each feature owns its unique property which has both advantages and drawbacks under the different situation. The depth-based PF using spin-image is a simple and fast adaptive tracking procedure, since this feature is view-invariant and robust with respect to object posture rotation and translation. When the subjects have relative smooth gait patterns the performance of tracking is satisfactory. However, one drawback of this method is that it is hard to capture complex situation patterns when subjects have a significant change in their walking patterns. For the depth-based PF using geodesic distance, one main advantage of this feature is that it is largely invariant to surface mesh deformations and rigid transformations. More precisely, the geodesic distance from the left hand to the right hand of a person along the body surface is unaffected by her/his posture. However, when the surface is not connected, that is, there is a self-occlusion, it fails to calculate the distance along its surface mesh and consequently the focus of the target may fall into other positions on the surface mesh.

Moreover, in our implementation it was observed that spin-image is sensitive to noise generated from the computation of surface normal, and computation of the geodesic distance might result in inconsistent labeling under more complex scenarios such as self-occlusion. In one of our future works, we would like to extend the proposed framework to a network of sensors and also utilize some hardware accelerators (e.g., GPU) in order to achieve a robust tracking of multiple people.

Additional Points

This paper is an extended version of our previous work [19]. This paper further exploits the notion of *importance sampling* which was discussed in [1] for the case of tracking in RGB and depth sensing with added intuitive motivations for increasing sampling rates. In particular, for the case of tracking in the second layer, new details are further included for body segmentation. Our previous work has more details about the implementation issues. Details of experimental set-up are omitted (or shortened) by referring to our previous work and no details are on implementation algorithms.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] S. Payandeh, "On importance sampling in sequential Bayesian tracking of elderly," in *Proceedings of the 10th Annual International Systems Conference, SysCon 2016*, USA, April 2016.

[2] G. Xu and S. Payandeh, "Sensitivity study for object reconstruction using a network of time-of-flight depth sensors," in *Proceedings of the 2015 IEEE International Conference on Robotics and Automation, ICRA 2015*, pp. 3335–3340, USA, May 2015.

[3] X. Dai and S. Payandeh, "Geometry-based object association and consistent labeling in multi-camera surveillance," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 3, no. 2, pp. 175–184, 2013.

[4] C. Choi and H. I. Christensen, "RGB-D object tracking: a particle filter approach on GPU," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1084–1091, Tokyo, Japan, November 20143.

[5] W. Choi, C. Pantofaru, and S. Savarese, "Detecting and tracking people using an RGB-D camera via multiple detector fusion," in *Proceedings of the 2011 IEEE International Conference on*

Computer Vision Workshops, ICCV Workshops 2011, pp. 1076–1083, Spain, November 2011.

[6] M. Luber, L. Spinello, and K. O. Arras, "People tracking in RGB-D data with on-line boosted target models," in *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems: Celebrating 50 Years of Robotics, IROS'11*, pp. 3844–3849, USA, September 2011.

[7] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4S: A real-time system for detecting and tracking people in 2 1/2D," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 1406, pp. 877–892, 1998.

[8] S. Ju, M. Black, and Y. Yacoob, "Cardboard people: A parameterized model of articulated image motion," in *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pp. 38–44, 1996.

[9] Y. Xu, Z. Miao, J. Wang et al., "Combining color features for real-time correlation tracking," *IEICE Transaction on Information and Systems*, vol. E100D, no. 1, pp. 225–228, 2017.

[10] X. Zhou, X. Liu, C. Yang, A. Jiang, and B. Yan, "Multi-Channel Features Spatio-Temporal Context Learning for Visual Tracking," *IEEE Access*, vol. 5, pp. 12856–12864, 2017.

[11] Y. Han, C. Deng, Z. Zhang, J. Li, and B. Zhao, *Adaptive feature representation for visual tracking*, 2017, https://arxiv.org/abs/1705.04442.

[12] Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Exploiting the Complementarity of Audio and Visual Data in Multi-Speaker Tracking," in *Proceedings of the ICCV Workshop on Computer Vision for Audio-Visual*, 2017.

[13] J.-Y. Lee and S. Payandeh, "Haptic teleoperation systems: Signal processing perspective," *Haptic Teleoperation Systems: Signal Processing Perspective*, pp. 1–130, 2015.

[14] Y. Li, H. Ai, T. Yamashita, S. Lao, and M. Kawade, "Tracking in low frame rate video: A cascade particle filter with discriminative observers of different life spans," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1728–1740, 2008.

[15] C. Plagemann, V. Ganapathi, D. Koller, and S. Thrun, "Real-time identification and localization of body parts from depth images," in *Proceedings of the 2010 IEEE International Conference on Robotics and Automation, ICRA 2010*, pp. 3108–3113, USA, May 2010.

[16] X. Liu and S. Payandeh, "Cascaded particle filter for real-time tracking using RGB-D sensor," in *Proceedings of the 2016 IEEE Canadian Conference on Electrical and Computer Engineering, CCECE 2016*, Canada, May 2016.

[17] Y. Lu and S. Payandeh, "Cooperative hybrid multi-camera tracking for people surveillance," *Canadian Journal of Electrical and Computer Engineering*, vol. 33, no. 3-4, pp. 145–152, 2008.

[18] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," in *Proceedings of the CVPR '2000: IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 126–133, June 2000.

[19] X. Liu and S. Payandeh, "Implementation of levels-of-detail in Bayesian tracking framework using single RGB-D sensor," in *Proceedings of the 7th IEEE Annual Information Technology, Electronics and Mobile Communication Conference, IEEE IEMCON 2016*, Canada, October 2016.

[20] L. Rabiner, *Multirate Digital Signal Processing*, Prentice Hall PTR, 1996.

[21] A. E. Johnson and M. Hebert, "Surface registration by matching oriented points," in *Proceedings of the 1997 1st International Conference on Recent Advances in 3-D Digital Imaging and Modeling*, pp. 121–128, May 1997.

[22] A. Johnson, "Spin-images: a representation for 3-D surface matching," Tech. Rep. CMU-RI-TR-97-47, 1997.

[23] M. Mortara, G. Patané, and M. Spagnuolo, "From geometric to semantic human body models," *Computers and Graphics*, vol. 30, no. 2, pp. 185–196, 2006.

[24] G. Bruce, "Shortest-path Algorithm: A Comparison," *Pperation Research*, vol. 24, no. 6, 1976.