Data Fusion in 3D Vision Using a RGB-D Data Via Switching Observation Model and Its Application to People Tracking

Du Yong Kim, Ba-Tuong Vo, and Ba-Ngu Vo

Abstract— **In this paper, we propose a new method for 3D people tracking with RGB-D observations. The proposed method fuses RGB and depth data via a switching observation model. Specifically, the proposed switching observation model intelligently exploits both final detection results and raw signal intensity in a complementary manner in order to cope with missing detections. In real-world applications, the detector response to RGB data is frequently missing. When this occurs the proposed algorithm exploits the raw depth signal intensity. The fusion of detection result and raw signal intensity is integrated with the tracking task in a principled manner via the Bayesian paradigm and labeled random finite set (RFS). Our case study shows that the proposed method can reliably track people in a recently published 3D indoor data set.**

I. INTRODUCTION

In recent years, 3D vision research using RGB-D data has been gaining popularity due to the development of compact and cheap sensors such as the Kinect [1]. Various techniques for multi-person tracking application using RGB-D data have been proposed in [5], [6].

In [5] 3D multi-person tracking using RGB-D data is achieved by applying histogram of oriented gradient (HOG) [2] detectors to RGB and depth data. Specifically, the HOG detector is applied to RGB data and the same method is proposed for depth data via the so-called histogram of oriented depth (HOD) [4]. Missed-detections are inevitable because the offline-trained HOG/HOD detectors are not perfect. To circumvent the missed-detection problem, the authors of [5] incorporate the online-boosting method of [3] when the tracking algorithm is not able to produce trajectories.

In [6], several detectors (e.g., upper body detector, depth-based shape detector) are combined to generate multiple cues. Each of these detectors has its own advantage in certain scenarios and the combination of detectors is expected to improve the performance over any individual detectors. The fusion of detector outputs is achieved by using Markov chain Monte Carlo simulation.

We exploit the observation that, in RGB-D data processing, more reliable bearing information can be obtained from the RGB data with a carefully designed detector whereas the depth information is only measured by the IR module. As with all detection-based techniques, missed-detection is inevitable and poses a significant challenge. To address this challenge, we propose a hybrid observation (RGB-D data from Kinect) model so as to utilize the appropriate observation models for the right scenarios to achieve reliable data fusion for 3D people tracking.

When the proposed hybrid observation model is used in multi-person tracking, data association is required to account for uncertainty in association of measurements to targets. Traditional tracking algorithms such as multiple hypothesis tracking (MHT) [9] and joint probability data association (JPDA) filter [10] are suitable in this respect, however, they depend on various heuristics and require additional track management for the appearance/disappearance of objects. On the other hand, the RFS multi-target tracking framework addresses all of the aforementioned uncertainties in a mathematically principled and systematic manner via the optimal Bayes multi-object filter [7], [8]. While efficient approximations to the optimal Bayes multi-object filter such as the probability hypothesis density (PHD) filter [7], [11], [12], the cardinalized PHD filter [7], [13], and the multi-Bernoulli filter [7], [14] are able to manage target appearance/disappearance, they do not produce tracks for individual targets. To address this problem, we adopt the labeled RFS multi-object filtering framework introduced in [16], [17].

Relevant research of RFS-based filters for multi-object tracking with video data can be found in [19], [20], [21], [22], [23]. However, to the best of our knowledge, this paper is the first attempt to apply RFS-based filtering to RGB-D data for 3D application. Specifically, this paper proposes a hybrid observation model for the labeled RFS multi-object filter which then leads to a reliable 3D people tracking algorithm for RGB-D data.

The remainder of the paper is organized as follows. Section 2 defines the states and measurements in 3D coordinates along with explanations of how 3D measurements are obtained from the Kinect sensor's RGB-D data. In Section 3 we detail the proposed observation model and explain how the proposed observation model is applied in the labeled RFS tracker. In Section 4, experimental results are demonstrated in a 3D people tracking application. Concluding remarks are presented in Section 5.

II. STATE AND OBSERVATION IN RGB-D DATA

A. Random finite set description for state and observation

Suppose that at time k, there are N_k objects and M_k observations (i.e., detections). Since multiple object states and their detections are not ordered and their numbers are not fixed, it is natural to represent the multi-object state and measurement as finite sets of points. To this end, we follow the RFS framework, where the states of the N_k objects and

Du Yong Kim is with the School of Electrical Electronic and Computer Engineering, University of Western Australia, Crawley, Australia duyongkim@gmail.com

Ba-Tuong Vo and Ba-Ngu Vo are with the Department of Electrical and Computer Engineering, Curtin University, Bentley, Australia {ba-tuong.vo, ba-ngu.vo}@curtin.edu.au

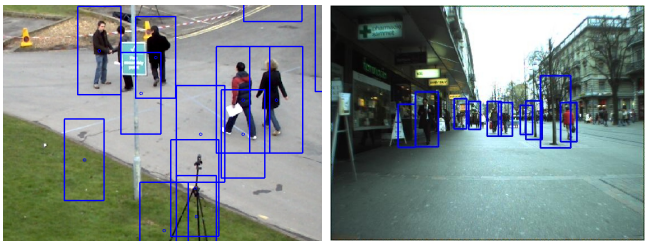

Fig. 1. HOG detections including clutters and missed-detections

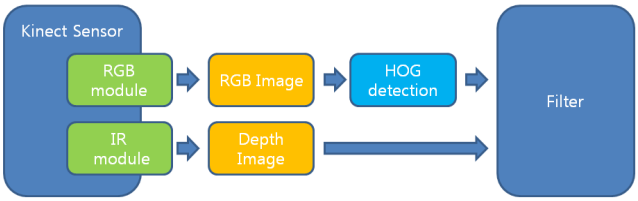

Fig. 2. Sensor schematic diagram

the M_k observations can be represented as a finite set of states and a finite set of observations, respectively:

$$X_k = \{x_{k,1}, ..., x_{k,N_k}\} \in \mathscr{F}(\mathbb{X}), \\ Z_k = \{z_{k,1}, ..., z_{k,M_k}\} \in \mathscr{F}(\mathbb{Z}), \quad (1)$$

where $\mathscr{F}(\mathbb{X})$ and $\mathscr{F}(\mathbb{Z})$, respectively, are the finite subsets of the state space $\mathbb{X}$ and observation space $\mathbb{Z}$. Here, the state vector is a 6-D vector, i.e., $x_k = [p_{X,k}, v_{x,k}, p_{Y,k}, v_{y,k}, p_{D,k}, v_{d,k}]^{\mathrm{T}}$ which represents the 3-D position and corresponding velocities. From a given image at time k, we measure the 3-D position $z_k = [p_{X,k}, p_{Y,k}, p_{D,k}]^{\mathrm{T}}$ by using a designated detector.

The above RFS representation is the basis of a number of algorithms from the RFS framework. We use the labeled RFS filter [16] in order to provide trajectories of people.

B. RGB-D observation for people detection

In this paper, the HOG detector is used to detect pedestrians in RGB data [2]. As the name suggests, the HOG describes the edge orientation information within the fixed size of the block and stores the information as a histogram. For the specific type of the object (e.g., people) the HOG has unique histogram that allows it be distinguished from the HOG of other types of objects. By using the HOG descriptor, a classifier, usually support vector machine (SVM), is used to determine whether the given HOG describes a target of interest or not. Due to the classification error, as shown in Fig. 1, HOG detections may contain false positives whose HOG features are quite similar to that of pedestrians, but are actually originated from non-targets. Conversely, there may be missed-detections if the HOG feature of pedestrian is not discriminative enough.

Depth information in RGB-D data is obtained from the IR camera of the Kinect sensor. The depth image resolution has the same pixel resolution as the RGB camera of the Kinect and the depth information at the certain pixel location is

Fig. 3. HOG detection results and depth intensity in RGB-D data

coded with 11 bits. The relation between the coded bit and the depth metric in meters is given in Eq. (3) in the next section.

III. Switching Observation Model for RGB-D Data Fusion

A. Proposed observation models for RGB-D data

In this section, we describe the likelihood functions for RGB detection (i.e, HOG) and depth intensity in the proposed hybrid observation model. Fig. 2 shows the sensor schematic diagram of the proposed observation model where RGB data is processed by the HOG detector and the depth image is directly used without any detection process. As an example, Fig. 3 illustrates the HOG detection and the depth intensity image. Note that the field of view in Fig. 3 is generated by two Kinect sensors.

For the RGB detection, assuming that the state is observed in additive Gaussian noise, the measurement likelihood of a detection at location z_k of an object with state x_k is given by

$$g_k^{(1)}(z_k|x_k) = \mathscr{N}(z_k; Hx_k, \Sigma) \quad (2)$$

where $\mathscr{N}(z; m, P)$ denotes a normal distribution with mean m and covariance P, $H = [1\ 0\ 1\ 0\ 1\ 0]$, and Σ is the covariance matrix of the observation noise. From RGB data input the designated detector outputs the Cartesian coordinates of the detections in 2D. This mathematical model, which describes the point wise detector response may be dependent on the type of features considered. For the depth measurement, we simply use the metric depth obtained from the raw depth value at the location of the RGB detection. Here, the raw depth value from the IR sensor is coded with 11 bits, thus the metric depth of the raw depth value can be calculated from Eq. (3). The likelihood function Eq.(2) can also be used for the HOD detector [4], which we test the HOD in our experiments.

To achieve reliable data fusion performance in 3D, we exploit the depth intensity from the IR module in the Kinect as a secondary observation. The rationale for using depth intensity as a secondary observation is that bearing information in the depth image is not reliable compared to the HOG detection with RGB image data. We used the conversion equation (3) to obtain the metric depth value from the raw depth information of the IR sensor as follows.

$$T_j(w_j(x_k)) = \frac{8 \cdot B \cdot F_x}{W_{max} - w_j(x_k)}, \quad (3)$$

where $T_j(\cdot)$ is the converted depth metric in meters at the pixel j, $w_j(x_k)$ is the raw depth information coded with 11 bits at the pixel j given the state x_k, $B = 0.075m$ is the distance between the IR projector and the IR camera, F_x is the focal length of the IR camera in the horizontal direction, and $W_{max} = 1084$ is the maximum raw depth value in bits.

Denoting the image of depth information by d_k, the likelihood of the depth intensity is given by

$$g_k^{(2)}(d_k) = \prod_{i=1}^{\mathbb{M}} \varphi_i(d_{k,i}) \prod_{i \in \mathbb{T}(x)} \frac{\phi_i(d_{k,i}, x_k)}{\varphi_i(d_{k,i})}, \tag{4}$$

where $\mathbb{M}$ is the number of pixels in the given image, $d_{i,k}$ is the depth intensity value at pixel i, $\phi_i(d_{k,i}, x_k) = \mathcal{N}(d_{k,i}; \bar{T}_i(w_i(x_k)), \sigma_t)$ is the target originated intensity distribution, $\varphi_i(d_{k,i}) = \mathcal{N}(d_{k,i}; 0, \sigma_b)$ is the non-target intensity distribution, $\mathbb{T}(x)$ denotes a set of pixels in illuminated by a target with state x, σ_t and σ_b are the variances of the Gaussian noises for target/non-target intensities, respectively, $\bar{T}_i(w_i(x_k))$ is the averaged metric depth of the target region in meters at pixel i illuminated by the target with given x_k. Note that the target region and the non-target intensity distribution are obtained from the background subtracted image similar to the method used in [18].

The observation likelihood model given by Eq. (4) is typically used for the track-before-detect problem in the literature [15], [18]. The difference between the track-before-detect method and ours is that our proposed method only uses the depth image data when the miss-detections occur.

The observation model given by Eq. (4) can be applied to another type of intensity such as a classification score of the HOG feature (i.e., raw signal before thresholding). However, the classification score distribution is dependent on the training process of the HOG detector. Consequently, it may not be suitable for autonomous robotics application where off-line training simply does not cover enough scenarios. On the other hand, the IR camera does not require training and the depth information is inherently accessible. Therefore, we propose to use the depth intensity image as a secondary observation to compensate the imperfection in HOG detections.

B. Application to people tracking in 3D

In this subsection, we describe how the proposed observation model is used with the labeled RFS filter. First, to make the paper self-contained, we briefly summarize the labeled RFS filter's prediction and update equation.

The multi-object state is a finite set $\mathbf{X} \in \mathscr{F}(\mathbb{X} \times \mathbb{L})$, where $\mathbb{L}$ is a discrete set called the label space. The label for given object state is defined as $\ell = (t, i)$ where t is the time of birth and i is the index of individual objects at that time. The set of labels of $\mathbf{X}$ is denoted by $\mathscr{L}(\mathbf{X}) \triangleq \{\ell : (x, \ell) \in \mathbf{X}\}$.

We consider multi-object density of the form:

$$\pi_{k-1}(\mathbf{X}) = \Delta(\mathbf{X}) \sum_{(I,\xi) \in \mathscr{F}(\mathbb{L}_k) \times \Xi_{k-1}} \omega_{k-1}^{(I,\xi)} \delta_I(\mathscr{L}(\mathbf{X})) \left[p_{k-1}^{(\xi)}\right]^{\mathbf{X}} \tag{5}$$

where $\Delta(\mathbf{X}) = \delta_{|\mathscr{L}(\mathbf{X})|}(|\mathbf{X}|)$ is the distinct label indicator. Here, $\mathbb{L}_k$ is the label space at time k, $\Xi_{k-1} \triangleq \Theta_0 \times ... \times \Theta_{k-1}$ denote the space of association map history to time $k-1$, where Θ_t denotes the space of all association maps at time t (an association map is a function $\theta : \mathbb{L} \to \{0, 1, ..., |Z|\}$ such that $\theta(i) = \theta(i') > 0$ implies $i = i'$). Each $I \in \mathscr{F}(\mathbb{L}_{k-1})$ represents a set of tracks labels at time k . The pair (I, ξ) represents the *hypothesis* that the set of tracks I has a history ξ of association maps. The weight $\omega_{k-1}^{(I,\xi)}$ represents the probability of hypothesis (I, ξ) and $p_{k-1}^{(\xi)}(\cdot, \ell)$ is the probability density of the kinematic state of track ℓ for the association map history ξ.

In the following, subscript 'B' and 'S' are used to distinguish the processes for new-born object and survived object, respectively. $\mathbb{B}_k$ is the label space of new-born objects, $1_{\mathbb{L}_k}(Y)$ is the indicator function meaning that whether the label set Y is a subset of $\mathbb{L}_k$ or not.

1) Prediction:

The predicted multi-object density to time k is given by

$$\pi_{k|k-1}(\mathbf{X}) = \Delta(\mathbf{X}) \sum_{(I,\xi) \in \mathscr{F}(\mathbb{L}_k) \times \Xi_k} \omega_{k|k-1}^{(I,\xi)} \delta_I(\mathscr{L}(\mathbf{X})) \left[p_{k|k-1}^{(\xi)}\right]^{\mathbf{X}} \tag{6}$$

where

$$\begin{aligned}
\omega_{k|k-1}^{(I,\xi)} &= w_{B,k}(I \cap \mathbb{B}_k) \omega_{S,k}^{(\xi)}(I \cap \mathbb{L}_{k-1}), \\
p_{k|k-1}^{(\xi)}(x, \ell) &= 1_{\mathbb{L}_{k-1}}(\ell) p_{S,k}^{(\xi)}(x, \ell) + (1 - 1_{\mathbb{L}_{k-1}}(\ell)) p_{B,k}(x, \ell), \\
p_{S,k}^{(\xi)}(x, \ell) &= \frac{\left\langle p_{S,k-1}(\cdot, \ell) f_{k|k-1}(x|\cdot, \ell), p_{k-1}^{(\xi)}(\cdot, \ell) \right\rangle}{\eta_S^{(\xi)}(\ell)}, \\
\eta_{S,k}^{(\xi)}(\ell) &= \left\langle p_{S,k-1}(\cdot, \ell), p_{k-1}^{(\xi)}(\cdot, \ell) \right\rangle, \\
\omega_{S,k}^{(\xi)}(L) &= [\eta_{S,k}^{(\xi)}]^{L_{k-1}} \textstyle\sum_{I \subseteq \mathbb{L}_{k-1}} 1_I(L) [q_S^{(\xi)}]^{I-L} \omega_{k-1}^{(I,\xi)}, \\
q_S^{(\xi)}(\ell) &= \left\langle 1 - p_{S,k-1}(\cdot, \ell), p_{k-1}^{(\xi)}(\cdot, \ell) \right\rangle.
\end{aligned}$$

Here, $f_{k|k-1}(x|\cdot, \ell)$ is the state evolution density for each object state. In the implementation, we used the constant velocity model as follows.

$$\begin{aligned}
x_k &= \mathbf{F} x_{k-1} + \Gamma v_k \\
\mathbf{F} &= I_3 \otimes \begin{bmatrix} 1 & T_s \\ 0 & 1 \end{bmatrix}, \Gamma = I_3 \otimes \begin{bmatrix} T_s^2/2 \\ T_s \end{bmatrix},
\end{aligned} \tag{7}$$

where T_s is the measurement scan interval, and $v_k \sim \mathcal{N}(v_k; 0, Q)$ is a i.i.d., Gaussian process noise vector with $Q = \sigma_v^2 I_3$ where σ_v is the standard deviation of the process noise acceleration. In the experiments, these values are set by considering the maximum speed of the object regarding to the frame rate; I_3 is the 3×3 identity matrix.

2)Update: When the measurement is available the updated multi-object density is given by

$$\begin{aligned}
&\pi_{k|k}(\mathbf{X}|Z_k) = \\
&\Delta(\mathbf{X}) \sum_{(I,\xi) \in \mathscr{F}(\mathbb{L}_k) \times \Xi_k} \sum_{\theta \in \Theta_k} \omega_k^{(I,\xi,\theta)}(Z_k) \delta_I(\mathscr{L}(\mathbf{X})) \left[p_{k|k}^{(\xi,\theta)}(\cdot|Z_k)\right]^{\mathbf{X}}
\end{aligned} \tag{8}$$

where Θ_k is the space of mappings $\theta : \mathbb{L}_k \to \{0, 1, ..., |Z_k|\}$, such that $\theta(i) = \theta(i') > 0$ implies $i = i'$, and

Fig. 4. Missed-detections in selcted frames (first row: HOG detection with RGB data, second row: HOD detection with depth data)

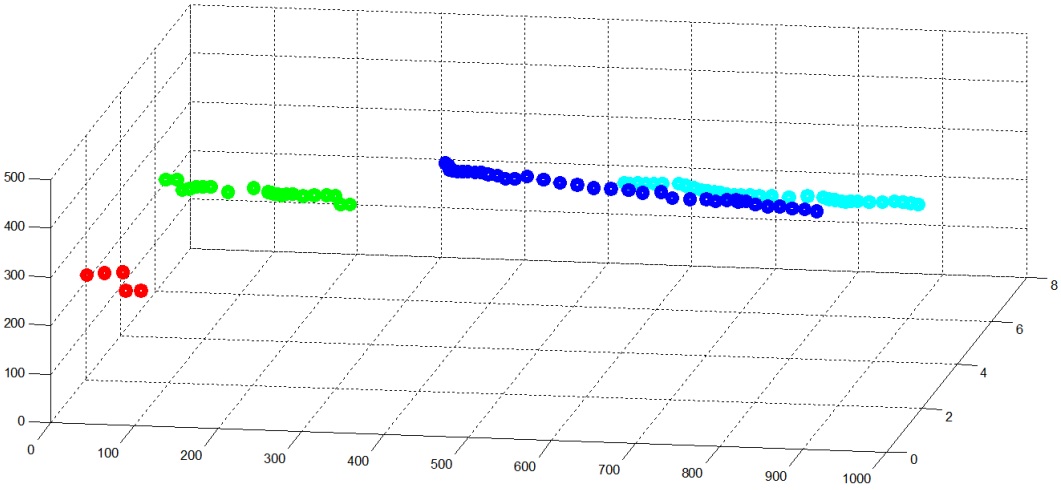

Fig. 5. Tracking results in 3D

$$\omega_k^{(I,\xi,\theta)}(Z_k) \propto \delta_{\theta^{-1}(\{0:|Z_k|\})}(I)\omega_{k|k-1}^{(I,\xi)}[\eta_{Z_k}^{(\xi,\theta)}]^I,$$
$$p_{k|k}^{(\xi,\theta)}(x,\ell|Z_k) = \frac{p_{k|k-1}^{(\xi)}(x,\ell)\psi_{Z_k}(x,\ell;\theta)}{\eta_{Z_k}^{(\xi,\theta)}(\ell)},$$
$$\eta_{Z_k}^{(\xi,\theta)}(\ell) = \left\langle p_{k|k-1}^{(\xi)}(\cdot,\ell), \psi_{Z_k}(\cdot,\ell;\theta)\right\rangle,$$
$$\psi_{Z_k}(x,\ell;\theta) = \delta_0(\theta(\ell))(1-p_{D,k}(x,\ell))\frac{p_{D,k}(x,\ell)g_k^{(2)}(d_k)}{\kappa_k(z_{\theta(\ell)})} + (1-\delta_0(\theta(\ell)))\frac{p_{D,k}(x,\ell)g_k^{(1)}(z_{\theta(\ell)}|x,\ell)}{\kappa_k(z_{\theta(\ell)})},$$

Note that $\psi_{Z_k}(x,\ell;\theta)$ is the generalized likelihood function which integrates the two types of likelihood (i.e., Eq.(2) and Eq.(4)).

For 3D people tracking, we have incorporated the proposed observation model, i.e., Eq. (2)-(4) into the labeled RFS filter. The first observation likelihood model is used when the detection response from the HOG detector is available. When there is a missed-detection in the HOG detector, we switch to the second observation model Eq. (4) in order to improve the tracking performance.

At this point, it is required to know in which track the missed-detection occurred. In a single person tracking case, it is straightforward to declare the missed-detection. However, in a multi-person tracking problem, it is difficult because track labels and data association are related. This problem is addressed by using the labeled RFS filter of [16], [17] because it provides the track label and data association information.

In summary, we propose a new form for the function $\psi_{Z_k}(x,\ell;\theta)$ in [16] where ℓ is the track label and θ is the association map so that the filter can deal with the missed detections by switching to an alternative measurement characterized by a corresponding likelihood. As the associ-

ation mapping results are available from the labeled RFS filter [16], it automatically utilizes one of HOG detection results or the depth intensity. With this mechanism, we can address missed-detections intelligently to provide more accurate tracking results. In the experiment, we consider the linear Gaussian dynamic motion model and implement the generalized labeled multi-Bernoulli recursion equations based on particle approximation.

IV. Experimental results

In this section, we evaluate the performance of the proposed labeled RFS filter with switching observation model on the RGB-D dataset of [4]. In our experiments, we also considered the HOD detector for comparison purpose. It will be shown that the HOD detections are not reliable in bearings as opposed to the HOG detections because the bearing information from the IR sensor module is not accurate.

The data set is generated by three vertically mounted Kinect sensors assembled on a sensory tower at 1.5m height. The data acquisition rate, i.e., frame rate, is 30Hz. In this dataset, temporal missed-detections occur in the detector. The HOG detector output depends on the detection threshold and missed-detections are inevitable when we only use detections with high certainty. Likewise, there are also missed-detections in the HOD detector. We display frames for which both the HOG and HOD have missed-detections, respectively, in Fig. 4. In these scenarios the combination of HOG and HOD, called Combo-HOD [4], is not able to alleviate the missed-detection problem as illustrated in the results.

The missed-detection problem can be effectively addressed by the proposed observation model. When missed-detections occur in several time steps the tracker with single observation model (including Combo-HOD) breaks down in the sense that it treats the same object as a new track after few time steps. This problem is alleviated by the proposed observation model as illustrated in Fig. 6.

We compare three approaches: 1) proposed model with HOG detector + depth intensity; 2) Combo-HOD; 3. HOD detector + depth intensity; to show the superior performance of the proposed observation model. Observe that the combination of two types of detectors (i.e., Combo-HOD) could slightly improve the tracking performance compared to that of the single type feature (i.e., method 3). However, it could not deal with mis-detections when both detectors failed (also described in Fig. 4). The proposed hybrid observation model using only the depth data (i.e., HOD + depth intensity) were tested to verify that bearing information in HOD detector is not good as the HOG detector.

Another possible implementation is the Combo-HOD with the proposed hybrid model. Our experience suggests possible improvement in the speed of the HOG detection with the help of the scale-informed search as noted in [4]. However, the tracking performance did not improve.

From the experimental results, it is recommended that more reliable bearing information from RGB data should be used and the depth intensity can be effectively used as a secondary observation to cope with temporal mis-detection. The tracking results of our method is also displayed in 3D space in Fig. 5.

An important advantage of the proposed approach is that it is general enough to apply to any type of detectors and raw signal intensities. In the current tested data set, the overlapping view of sensor is not effectively considered. This is a future direction of research. Additionally, object occlusion problem is another important issue that needs further investigation.

V. Conclusion

In this paper, we have presented a hybrid observation model that fuses the bearing and range information in RGB-D data for detection and tracking in 3D space. The proposed observation model is particularly effective when there are RGB detection losses. To this end, the observation model switches between the final detection response and the raw signal density. The proposed observation model is verified on a 3D multi-person tracking example. We used the labeled multi-object Bayes filter as a tracking framework because it provides target trajectories and allows the proposed observation model to be seamlessly integrated within the filter. The proposed method has been successfully tested with a recently published real-world RGB-D dataset.

References

[1] *Xbox 360 Kinect Sensor Manual,* Microsoft, Oct. 2010.

[2] N. Dalal and B. Triggs, "Histogram of oriented gradients for human detection," In *CVPR*, 2005.

[3] H. Grabner and H. Bischof, "On-line Boosting and Vision," In *CVPR*, 2006.

[4] L. Spinello, and K.O. Arras, "People Detection in RGB-D Data," *IEEE/RSJ International Conference on Intelligent Robots and Systems,* (IROS'11), San Francisco, CA, USA, 2011.

[5] M. Luber, L. Spinello, and K.O. Arras, "People Tracking in RGB-D Data With Online Boosted Target Models, " *IEEE/RSJ IROS 2011,* (IROS'11), San Francisco, CA, USA, 2011.

[6] W. Choi, C. Pantofaru, and S. Savarese, "Detecting and Tracking People using an RGB-D Camera via Multiple Detector Fusion," *Workshop on Challenges and Opportunities in Robot Perception, at ICCV 2011,* (ICCVW'11), Barcelona, Spain, 2011.

[7] R.P.S. Mahler, *Statistical Multisource-Multitarget Information Fusion*, Norwood, MA: Artech House, 2007.

[8] R.P.S. Mahler, "Multitarget Bayes Filtering via First-Order Multitarget Moments," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1152-1178, 2003.

[9] D. Reid, "An algorithm for tracking multiple targets," *IEEE Trans. AC*, vol. AC-24, no. 6, pp. 843–854, 1979.

[10] Y.Bar-Shalom and T.E. Fortmann, *Tracking and Data Association*, Academic Press, San Diego, 1998.

[11] B.N. Vo, S. Singh, and A. Doucet, "Sequential Monte Carlo methods for multitarget filtering with random finite sets," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 41, no. 5, pp. 1224-1245, 2005.

[12] B.N. Vo, and W.-K. Ma, "The Gaussian Mixture Probability Hypothesis Density Filter," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4091-4104, 2006.

[13] B.-T. Vo, B.-N. Vo, and A. Cantoni, "Analytic implementations of the Cardinalized Probability Hypothesis Density Filter," *IEEE Trans. Signal Processing*, Vol. 55, No. 7, part 2, pp. 3553-3567, 2007.

[14] B.T. Vo, B.N. Vo, and A. Cantoni, "The Cardinality Balanced Multi-Target Multi-Bernoulli Filter and Its Implementations," *IEEE Trans. Signal Processing*, vol. 57, no. 2, pp. 409-423, 2009.

[15] B.N. Vo, B.T. Vo, N.-T. Pham, and D. Suter, "Joint Detection and Estimation of Multiple Objects From Image Observations," *IEEE Trans. Signal Processing*, vol. 58, no. 10, pp. 5129-5141, 2010.

Fig. 6. Selected frames of tracking results (first row: HOG detection with depth intensity (proposed), second row: Combo-HOD, third row: HOD detection with depth intensity)

[16] B.-T. Vo, and B.-N. Vo, "A Random Finite Set Conjugate Prior and Application to Multi-Target Tracking," Proc. 7th Int. *Conf. Intelligent Sensors, Sensor Networks & Information Processing* (ISSNIP'2011), Adelaide, Australia, December 2011.

[17] B.-T. Vo, and B.-N. Vo, "Labeled Random Finite Sets and Multi-Object Conjugate Priors," *IEEE Trans. Signal Processing*, vol. 61, no. 13, pp. 3460-3475, 2013.

[18] R. Hoseinnezhad, B.N. Vo, and B.T. Vo, "Visual Tracking in Background Subtracted Image Sequences via Multi-Bernoulli Filtering," *IEEE Trans. Signal Processing*, vol. 61, no. 2, pp. 392-397, 2013.

[19] N.-T. Pham, W. Huang, and S. Ong, "Probability hypothesis density approach for multi-camera multi-object tracking," In Proc. *ACCV'07*, vol.I, Tokyo, Japan, November, 2007, pp. 875-884.

[20] E. Maggio, M. Taj and A. Cavallaro, "Efficient multitarget visual tracking using random finite sets," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1016-1027, 2018.

[21] Stephan Reuter, Kaus Dietmayer, "Pedestrian tracking using random finite sets," Proc. 14th Int. *Information Fusion (FUSION)*, 2011.

[22] T. M. Wood, C. A. Yates, D. Wilkinson, G. Rosser, "Simplified Multitarget Tracking using The PHD Filter for Microscopic Video Data," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 22, no.5, pp. 702-713, 2012.

[23] N. Ikoma,"Multiple pedestrians tracking with composite sensor of laser range finder and omni-directional camera by SMC implementation of PHD filter," In Proc. *Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on. IEEE*, 2012.