IDENTITY ASSOCIATION USING PHD FILTERS IN MULTIPLE HEAD TRACKING WITH DEPTH SENSORS

Qingju Liu, Teofilo E. de Campos, Wenwu Wang, Adrian Hilton

CVSSP, University of Surrey, Guildford GU2 7XH, UK

ABSTRACT

The work on 3D human pose estimation has been through a significant amount of progress in recent years, particularly due to the widespread availability of commodity depth sensors. However, most pose estimation methods follow a tracking-as-detection approach which does not explicitly handle occlusions, thus introducing outliers and identity association issues when multiple targets are involved. To address these issues, we propose a new method based on Probability Hypothesis Density (PHD) filter. In this method, the PHD filter with a novel clutter intensity model is used to remove outliers in the 3D head detection results, followed by an identity association scheme with occlusion detection for the targets. Experimental results show that our proposed method greatly mitigates the outliers, and correctly associates identities to individual detections with low computational cost.

Index Terms— 3D tracking, PHD filter, identity association, outlier detection

1. INTRODUCTION

Person tracking has been extensively studied in the field of computer vision [2, 3, 4, 5, 6, 7, 8], with various applications ranging from surveillance, video retrieval, teleconferencing to human-computer interaction and games. Of particular interest in multiple person tracking is when a person occludes another, which causes naïve algorithms to swap the identities of targets or get multiple trackers to follow the same person.

Several approaches have been proposed for person re-identification [9], most of them are appearance-based [10, 11]. The intended application in this paper is to track users in home theaters for spatial audio applications [12]. In this scenario, illumination tends to be quite poor, as dimmed lights are commonly used or the only light source comes from the TV. Therefore, appearance-based methods cannot be applied.

We are grateful for the help of Mark Barnard on building the dataset used in this paper. We would like to acknowledge the support of the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership. Data underlying the findings are fully available without restriction, details are available from [1].

A number of methods have been proposed to detect and track people in depth images [13, 14], particularly those generated using sensors based on structured light projection, such as the commercial system Kinect for Xbox 360. Kinect is designed to work on living rooms, the range of distance where it operates is optimal for our application, whereas other implementations available off-the-shelf have been optimized to be used on webcam scenarios, with a much smaller working distance range. To address the ID association problem for Kinect sensors, the method in [15] combines face recognition, clothing color tracking and height estimation exploiting both RGB and depth streams. Yet, it still depends on the illumination condition. Barbosa et al. [16] extract skeleton-based and surface-based features from the range data to re-identify individuals. However, its computational cost is high.

In this paper, we use a second generation Kinect sensor [17] (dubbed Kinect2 in this paper), which is a time-of-flight depth sensor. For spatial audio applications, only head positions are required. Kinect for Windows SDK 2.0 offers tools to detect up to six people and estimate their poses based on a skeleton model with 25 joints, from which the 3D head positions can be detected straightforwardly. However, this skeleton detector follows a tracking-as-detection strategy, which does not filter out results and a number of outliers are generated. More crucially, identities of targets tend to get swapped or re-assigned to new values when occlusions happen.

To address these issues, a Probability Hypothesis Density (PHD) filter [18] is applied to 3D head detection results to remove outliers, followed by a proposed ID association scheme to correct these swapped identities. A novel clutter intensity model is used in the PHD filter, which is measurement-driven and depends on the depth sensor's Field Of View (FOV). Processing of the 3D head detection results, rather than RGB images or depth maps, requires very little data bandwidth, enabling real-time implementation in a separate computer, not interfering with the machine that detects skeletons.

The remainder of the paper is organized as follows. Section 2 introduces the PHD filter with the proposed clutter intensity model. Section 3 presents the ID association scheme based on occlusion detection. Experimental results are shown and analyzed in Section 4. Conclusions and insights for future research directions are raised in Section 5.

2. PHD FILTER WITH DOMAIN-SPECIFIC CLUTTER FUNCTION

We propose a modification of the Probability Hypothesis Density (PHD) filter [18] to take clutter potential into account. Here, clutter is used to denominate failure cases such as outlying position detections and detection failure.

The PHD filter was introduced by Mahler, which addresses the problem of computational intractability of Bayes filters. Suppose the measurement set at the k-th frame contains m_k observations $\mathbf{Z}_k = \{\mathbf{z}_1, ..., \mathbf{z}_{m_k}\}$, and we aim to find the hidden multi-target state $\mathbf{X}_k$ from the accumulated measurement sets up to time k. For our specific application, the measurement set $\mathbf{Z}$ contains the 3D head positions $\mathbf{z} = [x, y, z]^\top$ from Kinect2 skeletal tracking. We aim to find the 3D head position as well as the velocity for each target $\mathbf{x} = [x, y, z, \dot{x}, \dot{y}, \dot{z}]^\top$, while filtering out clutters.

The Sequential Monte Carlo (SMC) implementation of the PHD filter as proposed in [19] is used in this paper. There is a parameter $\kappa(\mathbf{z})$, i.e. the PHD of clutter or clutter intensity, which plays an important role in SMC-PHD. Instead of uniform distribution as used in [19, 20], we propose a novel clutter intensity model $\kappa(\mathbf{z}|\mathbf{Z})$, which is based on occlusion detection from the measurement $\mathbf{Z}$ as well depth sensor's range and FOV constraints:

$$\kappa(\mathbf{z}|\mathbf{Z}) = \kappa + \kappa_1(\mathbf{z}) + \kappa_2(x, z) + \kappa_3(\mathbf{z}|\mathbf{Z}). \quad (1)$$

In the above equation, κ is the clutter intensity for visible targets in the light shaded area in Fig. 1. κ_1 and κ_2 are the clutter intensity increments for measurements that are out of the sensor's range and FOV respectively. The FOV in the horizontal plane, i.e. the x-z plane is considered. κ_3 is the clutter intensity increment caused by occlusions. Occlusions are illustrated in Fig. 1, where an object close to the depth sensor might occlude an object behind it. When the overall clutter intensity is large, the weights of all particles that represent the targets (newborn or not) decreases. When it is small, the weights of newborn target near any new measurement dramatically increases. Details about the parametrization of $\kappa(\mathbf{z}|\mathbf{Z})$ are given in Section 4.2.

3. IDENTITY ASSOCIATION

After applying the PHD filter at each time instance, the estimated state is less noisy, but person identification problems can persist. We propose to re-assign identities in two steps: short- and long-term analysis. To facilitate the analysis, a lookup table IDstatus is created in the format below, which tracks the status of existing IDs for up to e.g. 5s:

$$\left\{ \begin{array}{c} [\mathsf{id}_1, \mathsf{time}_1, \mathsf{pos}_1 = (\mathsf{pos}_{1,x}, \mathsf{pos}_{1,y}, \mathsf{pos}_{1,z})] \\ [\mathsf{id}_2, \mathsf{time}_2, (\mathsf{pos}_{2,x}, \mathsf{pos}_{2,y}, \mathsf{pos}_{2,z})] \\ \cdots \\ [\mathsf{id}_N, \mathsf{time}_N, (\mathsf{pos}_{N,x}, \mathsf{pos}_{N,y}, \mathsf{pos}_{N,z})] \end{array} \right\},$$

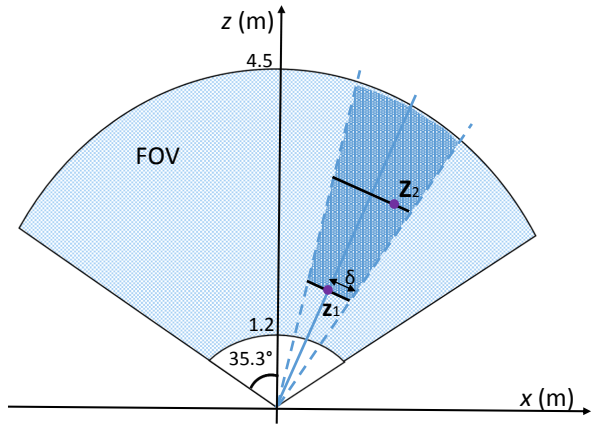


Fig. 1: The Kinect2 person detector's range goes from 1.2m and 4.5m of distance and its FOV is of $70.6°$ in the horizontal plane. A target at position $\mathbf{z}_1$ with radius δ occludes the area illustrated by the shadow confined by the two dashed lines. A target at $\mathbf{z}_2$ is likely to be occluded and results in a high clutter intensity $\kappa_3(\mathbf{z}_2|\mathbf{Z})$.

where id last appears at time at the position pos.

The short-term analysis aims to keep the consistency within a small time interval (e.g. 0.2s). The distance between the target state $\mathbf{x}$ and the states from the lookup table whose last appearance time is within the short-term threshold is calculated. If the minimum distance to sample j is smaller than a threshold (*distance upper bound*), and the distances to other samples are larger by more than a *distance difference lower bound*, then $\mathbf{x}$ can be assigned to the id_j. The associated time in the lookup table will be updated with $\mathsf{time}_j = 0$.

After this step, if there is still any target $\mathbf{x}$ without any assigned ID, two assumptions are imposed to address this problem. First, it might have appeared before, and its ID is saved in the lookup table IDstatus. However, none of the existing IDs are in the near neighborhood of the target since it might have been occluded for a while. Second, it might be a newborn target with a new ID.

Thus the long-term analysis is performed considering occlusions. The analysis for this step is done on the horizontal (x-z) plane. All the existing IDs with subject to $\mathsf{time}_j \neq 0, j = 1, \cdots, N$ might be the current target $\mathbf{x}$ that got occluded for time_j frames. Assuming the occluded target walks through a straight line between pos_j and $\mathbf{x}$, we can evenly divide the line into several segments, with each one lasting about one second. The center of the i-th segment is denoted as $\mathsf{p}_{j,i} = (\mathsf{p}_{j,i,x}, \mathsf{p}_{j,i,y}, \mathsf{p}_{j,i,z})$. If all of these segments are detected as occluded, then $\mathbf{x}$ is likely to be originated from id_j. Otherwise, a new random ID will be assigned to the target. If more than one existing IDs are likely to originate $\mathbf{x}$, then the nearest ID will be assigned to it.

For the particular segment center $\mathsf{p}_{j,i}$, occlusion detection is performed by searching for any detected targets in the i-th

time segment, whose occluded area covers the segment center position $\mathsf{p}_{j,i}$. Therefore, the log of all detected targets at the i-th time segment is analyzed. However, the calculation of occluded area of each of these targets is time consuming and there are more factors affecting the occlusion such as the height and width of each target. A computationally efficient approximation is proposed by first drawing the bearing line between the origin and $\mathsf{p}_{j,i}$. The distance between the above targets to this bearing line is then calculated. If any target is closer to the origin and its distance to the bearing line is smaller than a certain threshold, then $\mathsf{p}_{j,i}$ is detected as occluded.

Note that the lookup table IDstatus is updated after ID association is performed at each frame. New IDs with the related positions and time $= 1$ will be added; for existing IDs, their positions and last appearance time will be updated.

4. EXPERIMENTS

4.1. Data

Two sequences of data were recorded for our experiments. The first sequence, which lasts about 2.5 minutes, was recorded in a living room, with a setup scenario, as shown in Fig. 2. Four people were involved in this session. Person 1 walks along the L-shaped blue line back and forth, while Person 2 walks along the red line. Person 3 is a manikin who stands still at the center of the room. Person 4 leaves the room in the beginning of the sequence, and re-enters it at towards end. Person 1, 2, 4 all walk asynchronously, at a pace of their preference.

Sequence 2 lasts about 30 seconds. It was recorded in an office space. Two people were involved in this session. We intentionally designed a much more challenging scenario, where these two subjects walk/run around each other freely, often getting very close to each other. Moreover, one of the subjects runs about and jumps from time to time to increase challenges in person detection.

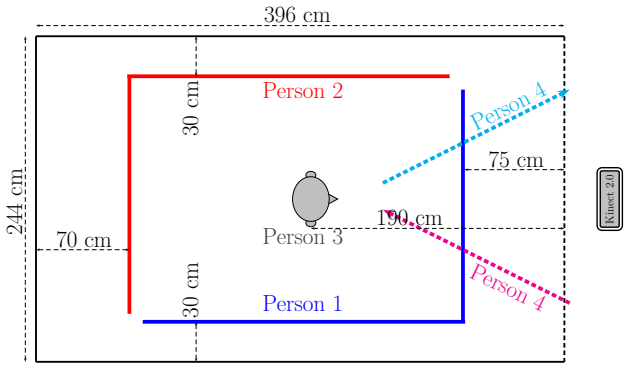


Fig. 2: Recording setup of Sequence 1. Person 1 to 3 are in the FOV all the time. Person 4 leaves and enters the FOV in the beginning and end of the sequence.

4.2. Parameter setup

Using the same notation as in [19], the parameter setup for the SMC-PHD filter is determined based on a validation subset of Sequence 1. The survival and detection probability were set to $P_S = 0.98$ and $P_D = 0.9$. Measurement-centered newborn targets were drawn with a Gaussian distribution whose Σ equals to the identity matrix times 0.02. The birth intensity $\nu = 0.1$ and the particle number per persistent or newborn target is $M_p = M_b = 400$. The measurement likelihood follows $g(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\|\mathbf{z} - \mathbf{x}\|_2|0, 0.02)$.

For the proposed intensity in Eq. (1), $\kappa = 0.2$. $\kappa_1(z)$ linearly increases from 0 to 2 when $\|\mathbf{z}\|_2$ increases from 4.5m to 5.5m or decreases from 1.2m to 0m. $\kappa_2(x, z)$ is set based on the FOV angle: when $|x| > \tan(35.3°)|z|$, $\kappa_2(x, z) = 1$; otherwise, it is 0. We set the radius of a target as $\delta = 0.2$, as illustrated in Fig. 1. Assuming $\mathbf{z}_i = [x_i, y_i, z_i]^\top$ is a target detected in front of $\mathbf{z} = [x, y, z]^\top$, then its mapped radius δ_i at the position of $\mathbf{z}$ roughly equals[1] $\delta_i = \delta \frac{\|\mathbf{z}\|_2}{\|\mathbf{z_i}\|_2}$. The occlusion-based clutter function is then defined as

$$\kappa_3(\mathbf{z}|\mathbf{Z}) = \sum_{\mathbf{z}_i \in \mathbf{Z}, z_i < z} 0.3 \exp\left(\frac{-d_i^2}{2\delta_i^2}\right),$$

where d_i is the distance from $\mathbf{z}$ to $\mathbf{z}_i$'s bearing line.

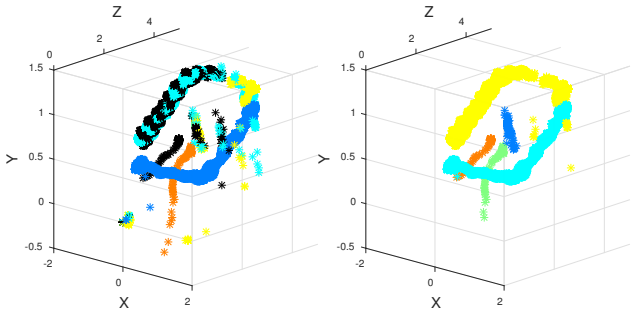
In the first step of the ID association scheme, short-term windows of 0.2s are used. The distance upper bound is 0.3m, and the distance difference lower bound is 0.2m. In the second step, the long-term log list contains a 5s buffer. Occlusion detection is carried out at each segment lasting 1s, and the distance threshold of 0.3m is used.

4.3. Results and analysis

The original noisy Kinect2 head detection results on Sequence 1 are shown in Fig. 3. Each color represents one ID. Person 1's ID is represented by the blue L-shaped curve. Person 2 is represented by the tangled curve containing black, yellow and cyan. Person 3 is around the center of the plots and its label (color) changes a lot during this session. Person 4 is represented by the black and orange from the edge to the center. A high number of outliers occur, especially in the occluded region on the right side. The oscillations are head movements resulting from walking motion. After applying our proposed method, the tracking results show a significant reduction in the number of outliers and corrected IDs are obtained, as shown in Fig. 3 (b).

In Sequence 2, since the walking trajectories for both participants are very complex, only the first 14 seconds are shown to improve understanding. The full sequence is available from [1]. As shown in top left subplot in Fig. 4, a high degree of occlusions occur within in this short period of time.

[1]This is only an approximation. The accurate result requires the calculation of the intersection point from $\mathbf{z}$ to the bearing line. Since the depth sensor is located at the origin of the coordinate system, $\|\mathbf{z}\|_2$ and $\|\mathbf{z_i}\|_2$ are the Euclidean distances from the targets to the sensor.



(a) Original (b) Processed

Fig. 3: Sequence 1 before (a) and after (b) applying the proposed method. A large number of outliers is present in (a) and the IDs of Person 2 and Person 3 change frequently, especially due to occlusions. The proposed method has successfully addressed this problem.

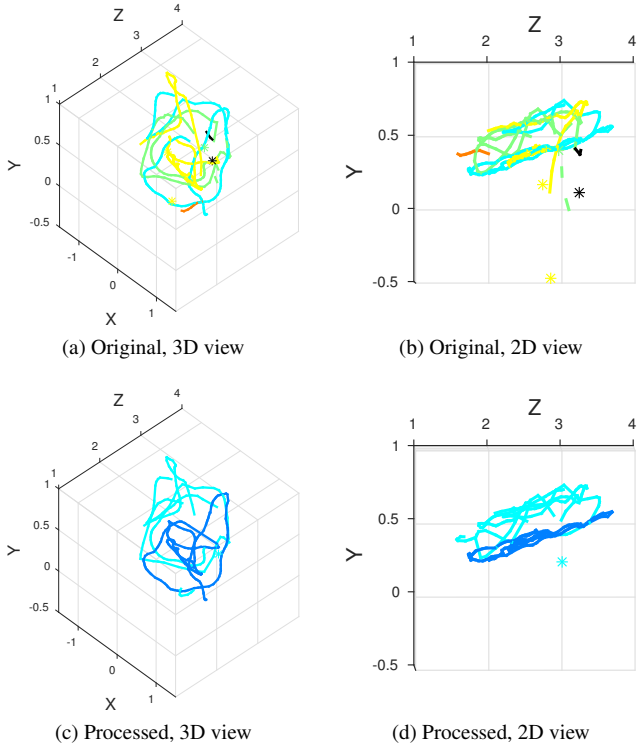
The two participants have different height, which can be seen in the vertical plane, i.e. the y-z plane as presented in the 2D view. We notice that after processing the detection results with our proposed method, most of the swapped or inaccurate IDs have been rectified, and only two colors are present, as shown in the bottom row of Fig. 4. Both detected trajectories can be fitted into one plane associated with one subject, as shown in the bottom left subplot. As expected, the subject shown in cyan was jumping up and down and the subject shown in dark blue kept the head at a roughly constant height.

In order to present quantitative analysis of our results, we created ground truth by manually labeling the IDs of each person. Table 1 shows the number of times when the ID of each target person has erroneously changed.

For Sequence 1, our proposed method successfully addressed the ID association problem of Person 2 and 3. Person 4 is still assigned with two IDs since the duration between its absence and presence is too long ($>$ 1 minute). Moreover, most of the outliers were removed. In the original tracking results of Sequence 1, 407 out of a total of 12455 detections are outliers, i.e. mis-detected targets take 3.3% of the samples. After processing, only 24 out of 11243 detections are outliers, i.e., 0.2%. However, this is done at the cost of missing a small number of correct detections. This is because the PHD filter needs several frames to converge to the target when it disappears and comes back again.

Dataset	Sequence 1				Sequence 2	
Subject index	P1	P2	P3	P4	P1	P2
Original	0	3	3	1	7	6
Processed	0	0	0	1	2	0

Table 1: Number of incorrect label changes.



(a) Original, 3D view (b) Original, 2D view

(c) Processed, 3D view (d) Processed, 2D view

Fig. 4: Sequence 2 before (top) and after (bottom) applying the proposed method for the first 14 seconds. Five IDs are assigned to the two participants in the original tracking results and their IDs often get swapped. After processing, only two IDs are associated to the two participants, and no swapping happens, as can be observed from subplot (d) as the two trajectories fit to two lines.

For Sequence 2, Person 2 is assigned with a consistent ID. However, Person 1's ID still changes twice. One reason is that Person 1 walks/runs very fast, with some free poses such as waving and bending over, which adds difficulties to both the head detection and ID association.

5. CONCLUSION

An efficient ID association method for multiple person tracking using depth sensors was proposed. The PHD filter with a novel clutter intensity model is used to remove outliers, and an ID association scheme with occlusion detection has been integrated. Experimental results show that our method successfully addresses the person re-identification problem without requiring appearance measurements. Future work will take into account long-term trajectory estimation, and individual attributes such as motion profile.

6. REFERENCES

[1] T. de Campos, Q. Liu, and M. Barnard, "S3A speaker tracking with Kinect2," Online, available from `cvssp.org/data/s3a`, February 2015, DOI 10.15126/surreydata.00807708.

[2] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer vision and image understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[3] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: an experimental survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1442–1468, 2014.

[4] J. Kwon and K. M. Lee, "Visual tracking decomposition," in *Proc of the IEEE Conf on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1269–1276.

[5] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–575, May 2003.

[6] K. Kim and L. S. Davis, "Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering," in *Proc 9th European Conf on Computer Vision, Graz, Austria, May 7-13*, 2006, pp. 98–109.

[7] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267–282, Feb 2008.

[8] Qingju Liu, Teofilo de Campos, Wenwu Wang, Philip Jackson, and Adrian Hilton, "Person tracking using audio and depth cues," in *ICCV Workshop on 3D Reconstruction and Understanding with Video and Sound*, December 2015.

[9] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person Re-Identification*, vol. 1, Springer, 2014.

[10] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc of the IEEE Conf on Computer Vision and Pattern Recognition*. 2011, pp. 649–656, IEEE Computer Society.

[11] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, July 2012.

[12] M. F. Simón Gálvez, D. Menzies, F. M. Fazi, T. E. de-Campos, and A. Hilton, "A listener position adaptive stereo system for object-based reproduction," in *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[13] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall, "A survey on human motion analysis from depth data," in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, vol. 8200 of *LNCS*, pp. 149–187. Springer, 2013.

[14] C. Redondo-Cabrera, R. Lopez-Sastre, and T. Tuytelaars, "All together now: Simultaneous detection and continuous pose estimation using a hough forest with probabilistic locally enhanced voting," in *Proc 25th British Machine Vision Conf, Nottingham, Sept 1-5*. 2014, BMVA Press.

[15] T. Leyvand, C. Meekhof, Y.-C. Wei, J. Sun, and B. Guo, "Kinect identity: Technology and experience," *Computer*, vol. 44, no. 4, pp. 94–96, Apr. 2011.

[16] I. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, "Re-identification with RGB-D sensors," in *Proceedings of the ECCV Workshops and Demonstrations*, 2012, vol. 7583, pp. 433–442.

[17] Microsoft, "Kinect for Windows," Online, Retrieved in September 2015, `https://dev.windows.com/en-us/kinect/`.

[18] R. P. S. Mahler, "Multitarget Bayes filtering via first-order multitarget moments," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1152–1178, Oct 2003.

[19] B. Ristic, D. Clark, B.-N. Vo, and B.-T. Vo, "Adaptive target birth intensity for PHD and CPHD filters," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 2, pp. 1656–1668, 2012.

[20] B.-N. Vo and W.-K. Ma, "The gaussian mixture probability hypothesis density filter," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4091–4104, Nov 2006.