

Probability Hypothesis Density Approach For Multi-Camera Multi-Object Tracking

Nam Trung Pham^{1,2}, Weimin Huang¹, S. H. Ong²

¹ Institute for Infocomm Research, Singapore

² Department of Electrical and Computer Engineering,
National University of Singapore

Abstract. Object tracking with multiple cameras is more efficient than tracking with one camera. In this paper, we propose a multiple-camera multiple-object tracking system that can track 3D object locations even when objects are occluded at cameras. Our system tracks objects and fuses data from multiple cameras by using the probability hypothesis density filter. This method avoids data association between observations and states of objects, and tracks multiple objects in single-object state space. Hence, it has lower computation than methods using joint state space. Moreover, our system can track varying number of objects. The results demonstrate that our method has a high reliability when tracking 3D locations of objects.

1 Introduction

Tracking moving objects is an important part of many applications. Some people proposed methods to track objects by using one camera [1]. However, when persons might be occluded by other persons in the scene, using one camera to track these persons is difficult. This is because information of these persons from one camera is not enough to solve the occlusion problem. An idea to solve this problem is to use multiple cameras to recover information that might be missing from a particular camera. Furthermore, multiple cameras can be used to recover the 3D information of objects.

There are some approaches for tracking with multiple cameras. Most of them have two stages. They are single-view stage and multiple-view data fusion stage. In the single-view stage, they extract observations, estimations. Then in the second stage, these data are fused to obtain the final results. Some methods are proposed to track one object using multiple cameras [2], [3]. These methods track an object and switch to another camera when the system predicts that the current camera no longer has a good view of the object. However, these methods need to consider data association when extending from tracking one object to multiple objects. Some other methods can track multiple objects [4], [5], [6], [7]. Among them, methods match objects between different camera views [4], [5] or incorporate classification methods [6] to do the data association between observations and objects in multiple views. These methods can collaborate multiple cameras for multiple-object tracking. However, when the appearances of objects

are similar or occlusions occur, these methods might not be suitable. This is because some wrong matches may occur. The other idea is to find 3D observations that correspond with observations from different views [7]. However, the association of observations from different views can increase computational cost in 3D observation searching.

Recently, there has been increasing research interest on using random set theory to solve multiple-object tracking. Here, the states of objects and measurements are represented as random finite sets (RFS). Mahler [8] presented a probability hypothesis density (PHD) filter that operates on a single-object state space. Vo [9], [10] proposed implementations of the PHD filter. Especially, the implementation in [10] is a closed-form of the PHD filter. It is called Gaussian mixture probability hypothesis density (GMPHD) filter.

In this paper, we extend the GMPHD filter from single sensor to multiple sensors to track several people using multiple cameras in a room. It is assumed that we have projection matrices from 3D space to cameras. Our method can recover the 3D object locations and handle the occlusion at each camera. We assume that color models are available. Then, the proposed tracking method can be efficiently applied to track a varying number of objects. Further, because the fusion stage of multiple cameras to obtain 3D object locations is based on the GMPHD filter, it reduces the amount of computation compared with other methods such as search based method or the particle filter.

2 PHD filter approach

In multiple-object tracking, it is difficult to obtain the posterior density function when the number of objects increases. Fortunately, this density function can be approximately recovered from a probability hypothesis density (PHD) [8]. To obtain the PHD at each time step, the PHD filter [8] can be applied. Now, we review an implementation of the PHD filter. It is the GMPHD filter [10]. The GMPHD filter is a close-form of the PHD filter with assumptions on linear Gaussian system. These assumptions are as follows. Each object follows a linear Gaussian model, i.e.,

$$f_{k|k-1}(x|\zeta) = \mathcal{N}(x; F_{k-1}\zeta, Q_{k-1}), \quad (1)$$

$$g_k(z|x) = \mathcal{N}(z; H_k x, R_k), \quad (2)$$

where $\mathcal{N}(\cdot; m, P)$ denotes a Gaussian density with mean m and covariance P , F_{k-1} is the state transition matrix, Q_{k-1} is the process noise covariance, H_k is the observation matrix, and R_k is the observation noise covariance. The survival and detection probabilities are $p_{S,k}$ and $p_{D,k}$, respectively. The intensity of the spontaneous birth RFS is

$$\gamma_k(x) = \sum_{i=1}^{J_{\gamma,k}} w_{\gamma,k}^{(i)} \mathcal{N}(x; w_{\gamma,k}^{(i)}, P_{\gamma,k}^{(i)}) \quad (3)$$

where $J_{\gamma,k}$ is the number of birth Gaussian components. It is assumed that the posterior intensity at time $k-1$ is a Gaussian mixture of the form

$$v_{k-1}(x) = \sum_{i=1}^{J_{k-1}} w_{k-1}^{(i)} \mathcal{N}(x; w_{k-1}^{(i)}, P_{k-1}^{(i)}) \quad (4)$$

where J_{k-1} is the number of Gaussian components of $v_{k-1}(x)$.

Under these assumptions, the predicted intensity to time k is given by

$$v_{k|k-1}(x) = v_{S,k|k-1}(x) + \gamma_k(x) \quad (5)$$

where

$$\begin{aligned} v_{S,k|k-1}(x) &= p_{S,k} \sum_{j=1}^{J_{k-1}} w_{k-1}^{(j)} \mathcal{N}(x; m_{S,k|k-1}^{(j)}, P_{S,k|k-1}^{(j)}), \\ m_{S,k|k-1}^{(j)} &= F_{k-1} m_{k-1}^{(j)}, \\ P_{S,k|k-1}^{(j)} &= Q_{k-1} + F_{k-1} P_{k-1}^{(j)} F_{k-1}^T. \end{aligned}$$

Because $v_{S,k|k-1}(x)$ and $\gamma_k(x)$ are Gaussian mixtures, $v_{k|k-1}(x)$ can be expressed as a Gaussian mixture of the form

$$v_{k|k-1}(x) = \sum_{i=1}^{J_{k|k-1}} w_{k|k-1}^{(i)} \mathcal{N}(x; m_{k|k-1}^{(i)}, P_{k|k-1}^{(i)}) \quad (6)$$

Then, the posterior intensity at time k is also a Gaussian mixture, and is given by

$$v_k(x) = (1 - p_{D,k}) v_{k|k-1}(x) + \sum_{z \in Z_k} v_{D,k}(x; z) \quad (7)$$

where

$$\begin{aligned} v_{D,k}(x; z) &= \sum_{j=1}^{J_{k|k-1}} w_k^{(j)}(z) \mathcal{N}(x; m_{k|k}^{(j)}, P_{k|k}^{(j)}), \\ w_k^{(j)}(z) &= \frac{p_{D,k} w_{k|k-1}^{(j)} q_k^{(j)}(z)}{\kappa_k(z) + p_{D,k} \sum_{l=1}^{J_{k|k-1}} w_{k|k-1}^{(l)} q_k^{(l)}(z)}, \\ q_k^{(j)}(z) &= \mathcal{N}(z; H_k m_{k|k-1}^{(j)}, R_k + H_k P_{k|k-1}^{(j)} H_k^T), \\ m_{k|k}^{(j)}(z) &= m_{k|k-1}^{(j)} + K_k^{(j)}(z - H_k m_{k|k-1}^{(j)}), \\ P_{k|k}^{(j)} &= [I - K_k^{(j)} H_k] P_{k|k-1}^{(j)}, \\ K_k^{(j)} &= P_{k|k-1}^{(j)} H_k^T (H_k P_{k|k-1}^{(j)} H_k^T + R_k)^{-1}. \end{aligned}$$

3 System overview

We propose a method to track 3D locations of heads of people using multiple cameras with assumptions that the cameras are calibrated and the field of views of cameras overlap. The proposed method, as shown in Fig. 1, consists of two major components: single-view tracking and multiple-camera fusion. In the first component, at each camera at time k , we find color observations and then use the GMPHD filter to estimate the 2D locations of objects. Let $Y_k^i = \{y_{1,k}^i, \dots, y_{m,k}^i\}$ be the set of 2D estimations of objects at time k , view i . We have n single views, so the set of 2D estimations of objects at time k can be defined by

$$Y_k = [Y_k^1; Y_k^2; \dots; Y_k^n] \quad (8)$$

More details on the first step will be shown in Section 5.

In the second component, we consider the set of 2D estimations of objects Y_k as observations for a data fusion step to estimate the 3D locations of objects by the GMPHD filter. This method can avoid the data association between observations and states of objects. More details of the second step will be shown in Section 6.

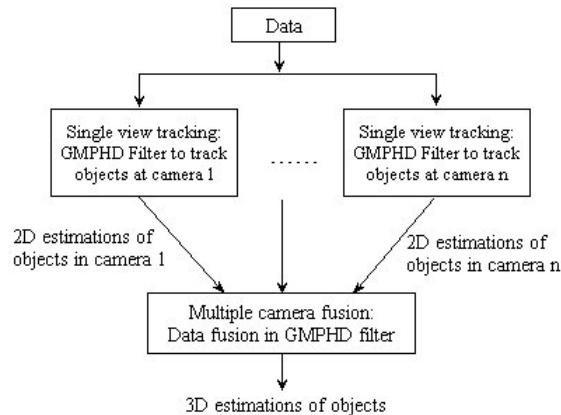


Fig. 1. The sketch of our system for multiple object tracking using multiple cameras

4 Color likelihood

The state of single object is described by $x = \{x_c, y_c, H_x, H_y\}$. This is a rectangle with center and size defined by $\{x_c, y_c\}$ and $\{H_x, H_y\}$, respectively. Let the color histogram of object be denoted as $p(u)$, the color histogram of template as $q(u)$.

The similarity function between an object and a template is measured by the Bhattacharyya distance [11].

$$D = \sqrt{1 - \int \sqrt{p(u)q(u)}du} \quad (9)$$

In multiple-object tracking, we can have many color models of templates, and let these models be as $\{q_1(u), q_2(u), \dots, q_n(u)\}$. The similarity function between an object and templates is modified by

$$D = \min_i \left(\sqrt{1 - \int \sqrt{p(u)q_i(u)}du} \right) \quad (10)$$

The color likelihood function is defined as in [1]

$$l_z(x) = \mathcal{N}(D; 0, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{D^2}{2\sigma^2}\right\} \quad (11)$$

where z is the current image, x is the state of object and σ^2 is the variance of noise.

5 Single-view tracking

At each single view, we assume that the object state does not change much between frames and each object in multiple-object tracking is evolved from a dynamic moving equation

$$x_k = x_{k-1} + w_k \quad (12)$$

where the state of an object in a single view $x_k = \{x_c, y_c, H_x, H_y\}$, and w_k is the process noise.

Single-view tracking consists of two parts: obtaining the color measurement random set, and using these color measurements to obtain the PHD. Now, we consider the i th camera. Let $v_k^i(x)$ be the PHD of the i th camera at time k and $v_{k|k-1}^i(x)$ be the predicted PHD of the i th camera at time k . From [12], we have

$$v_k^i(x) \propto \tilde{v}_k^i(x) = l_z(x)v_{k|k-1}^i(x) \quad (13)$$

where $l_z(x)$ is the color likelihood that is defined in Section 4. Hence, peaks of $\tilde{v}_k^i(x)$ are also peaks of $v_k^i(x)$. We apply the method in [12] to collect peaks in $\tilde{v}_k^i(x)$. The set of these peaks is considered as the color measurement random set.

Secondly, we use the color measurement random set to update the PHD by the updating step in the GMPHD filter (Equation (7)). After updating predicted PHD $v_{k|k-1}^i(x)$ with the color measurement random set, we obtain PHD $v_k^i(x)$. From PHD $v_k^i(x)$, we find Gaussian components whose weights are larger than a threshold (0.5). The set of means of these Gaussian components are 2D estimations of objects at the i th camera. They are denoted as $Y_k^i = \{y_{1,k}^i, \dots, y_{m,k}^i\}$. (See [12] for more details of single-view tracking.)

6 Multiple-camera fusion

We assume that the dynamic moving equation for 3D tracking is

$$x_k = x_{k-1} + w_k \quad (14)$$

where the state of an object $x_k = \{x_{1,k}, x_{2,k}, x_{3,k}\}$ is a 3D coordinate, w_k is the process noise.

The observations are 2D estimations from multiple cameras. So, the measurement equation at the i th camera is described by

$$\begin{aligned} \begin{pmatrix} l_{1,k} \\ l_{2,k} \\ l_{3,k} \end{pmatrix} &= \begin{pmatrix} a_{11}^i & a_{12}^i & a_{13}^i & a_{14}^i \\ a_{21}^i & a_{22}^i & a_{23}^i & a_{24}^i \\ a_{31}^i & a_{32}^i & a_{33}^i & a_{34}^i \end{pmatrix} \begin{pmatrix} x_{1,k} \\ x_{2,k} \\ x_{3,k} \\ 1 \end{pmatrix} \\ \begin{pmatrix} y_{1,k}^i \\ y_{2,k}^i \end{pmatrix} &= \begin{pmatrix} l_{1,k}/l_{3,k} \\ l_{2,k}/l_{3,k} \end{pmatrix} + u_k \end{aligned} \quad (15)$$

where u_k is the measurement noise, and a_{mn}^i are projection parameters from 3D coordinate to the i th camera plane. Assuming that cameras are calibrated, we have projection parameters a_{mn}^i .

The idea of fusing data from multiple cameras is to use the GMPHD filter sequentially at each camera. There are some related work that used sequential sensor updating method in the PHD approach [8]. Let $V_k(x)$ be the PHD for multiple-camera tracking at time step k . We propose the fusion stage as follows

- Step 1: Assuming that we have the PHDs of previous time step $k-1$ of multiple-camera fusion stage $V_{k-1}(x)$ and single-view tracking stage $v_{k-1}^1(x)$ at camera 1, we employ the method in Section 5 to obtain the set of 2D estimations of objects, Y_k^1 , and PHD $v_k^1(x)$. Then, from $V_{k-1}(x)$, we use dynamic moving equation (14) and measurement equation (15) to predict $V_{k|k-1}^1(x)$ at camera 1 by Equation (5). Because measurement equation (15) is not linear, we have to use unscented transform in the prediction step (more details is in [10]). Then, the set of 2D estimations of objects at the camera 1, Y_k^1 , is used to update the $V_{k|k-1}^1(x)$ to $V_k^1(x)$ by the updating step in the GMPHD filter (Equation (7)). From assumptions on the GMPHD filter, $V_{k-1}(x)$ is a Gaussian mixture, so $V_k^1(x)$ is also a Gaussian mixture.
- Step 2: Set $i = 2$
- Step 3: At the camera i , set $V_{k|k-1}^i(x) = V_k^{i-1}(x)$. Assuming that we have the PHD of previous time step $k-1$ of single-view tracking stage at camera i , $v_{k-1}^i(x)$, the method described in Section 5 is performed to obtain the set of 2D estimations of objects at camera i , Y_k^i , and PHD $v_k^i(x)$. Because $V_{k|k-1}^i(x)$ is a Gaussian mixture, we can use the updating step of the GMPHD filter to update $V_{k|k-1}^i(x)$ with observations in Y_k^i . This means

$$V_k^i(x) = (1 - p_{D,k})V_k^{i-1}(x) + \sum_{y \in Y_k^i} V_{D,k}(x; y) \quad (16)$$

- Then, we obtain the $V_k^i(x)$.
- Step 4: Set $i = i + 1$. If $i \leq n$ then we repeat the step 3. Otherwise, we have $V_k^n(x)$. The PHD of the system is $V_k(x) = V_k^n(x)$. For estimating the 3D object locations, we investigate the PHD of the system $V_k(x)$ and choose Gaussian components whose weights are larger than a threshold (0.5) to obtain the 3D estimations of objects.

We note that the GMPHD filter in [10] do not include the track labels of objects. For label tracking, our method is described as follows. Each Gaussian component is associated with a label. For birth Gaussian components, we assign them a special label (for example -1). After the updating step in the first camera, Gaussian components with labels become the predicted Gaussian components for the second camera and then they are used to update the PHD in the second camera. At the last camera, for each label, we choose the Gaussian component that has the largest weight. The estimations of object locations are from the means of these largest Gaussian components. If a Gaussian component has a special label and its weight is large enough, we assign it a new label. This means a new person occurs. Hence, the identifications of people are defined in the tracking. This track label method is extended from the work in [13] from single sensor to multiple sensors and then applied in multiple-camera multiple-object tracking.

7 Experimental results

We test the performance of our method with data from the first and second cameras in scenarios seq24-2p-0111, seq35-2p-1111, and seq44-3p-1111 in test database [14]. There are about 4500 time steps (9000 image frames). The errors of 3D estimations are measured by the Wasserstein distance [9] and are shown in Table 1. For visualization, we show the results from test case 'seq44-3p-1111'.

Table 1. Error of 3D estimation

Scenarios	Mean error (m)
seq24-2p-0111	0.06
seq35-2p-1111	0.05
seq44-3p-1111	0.07

In this scenario, there are three persons. They appear and disappear at different times. This scenario is challenging because occlusions occur between persons when they cross together. Moreover, in this scenario, the lighting of the room changes through the tracking, so it is difficult to apply segmentation methods. In addition, because the color models of heads are different between views, it is sometimes difficult to apply methods such as Stereo Matching to find the correspondences. Hence, the 3D reconstructions from correspondences are not

reliable in this data. However, our method successfully track 3D object locations in this scenario.

At each camera, we used 400 samples to detect peaks of PHD. The maximum of Gaussian mixture components are 30. We assume that persons enter the tracking areas from two entrances. Hence, the birth intensity is the mixture of Gaussian components whose means are locations at these entrances. The clutter density in the multiple-view camera fusion is an uniform distribution on the tracking area $3\text{m} \times 2\text{m} \times 2\text{m}$ and the clutter density in the single-view tracking stage is an uniform distribution of the size of image (it is the projection from tracking area to cameras) and the range of radius H_x and H_y ([5,15]). The probability of survival is $p_S = 0.99$ and the probability of detection is $p_D = 0.98$. These parameters are set by experiments.

Figure 2 shows the performance of 3D people tracking. The dots are ground-truth and the lines are estimations from our methods. The results indicate that tracks of people are maintained. The x and y components are reliable while the z component has some errors, for example at steps 600 to 700. This is because at steps 600 to 700, the color of the background near the person’s location at camera 2 is similar to color of templates. However, these errors are quite small. In this sequence, when a person moves out of the view and then moves back, we will assign it a new label, which is treated as correct detection. Figure 3 shows the results when we project 3D locations to the camera plane. Each cell in the figure has two images. The left image is from camera 1 and the right image is from camera 2. In this figure, we can see that at time $k = 99, 144, 247$, the first, second, third persons appear in the overlapped region sequentially. They are detected and tracked automatically. At time $k = 264, 295$, the occlusions between the second and third persons occur in camera 1 and 2. However, the tracks are maintained after the occlusions. At time $k = 809$, the occlusion between the first and third persons occurs at camera 1 and the occlusion between the first and second persons occurs at camera 2. We can see in the figure that our method can handle these cases. This is because the PHD from camera 1 is a good prediction for the PHD at camera 2. Information from two cameras is fused to obtain the reliable 3D estimations without using data association methods.

8 Conclusions

The paper described a method of using the GMPHD filter to track 3D locations of objects. The method can track a varying number of objects. Moreover, it can solve some occlusion problems for which single-camera system has difficulty. The fusion stage using the GMPHD filter reduced a lot of computations compared with other methods that search whole space or the particle filter with multiple objects. Experimental results have shown that the proposed approach is promising.

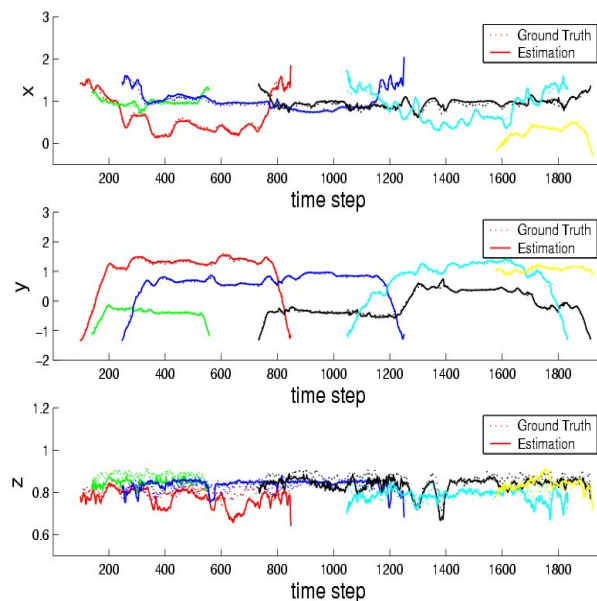


Fig. 2. 3D results of tracking multiple people using PHD filter

9 Acknowledgements

The authors would like to thank Prof. Ba Ngu Vo at Melbourne University for his helps and fruitful discussions. This work is partially supported by EU project ASTRALS (FP6-IST-0028097).

References

1. Czyz, J., Ristic, B., Macq, B.: A color-based particle filter for joint detection and tracking of multiple objects. In: ICASSP. (2005)
2. Cai, Q., Aggarwal, J.K.: Automatic tracking of human motion in indoor scenes across multiple synchronized video streams. In: ICCV, Bombay, India (1998)
3. Nummiaro, K., Koller-Meier, E., Svoboda, T., Roth, D., Gool, L.V.: Color-based object tracking in multi-camera environments. In: 25th Pattern Recognition Symposium, DAGM. (2003)
4. Chang, T., Gong, S.: Tracking multiple people with a multi-camera system. In: IEEE Workshop on Multi-Object Tracking. (2001)
5. Mittal, A., Davis, L.S.: M2Tracker: a multi-view approach to segmenting and tracking people in a cluttered scene. *International Journal of Computer Vision* **51**(3) (2003) 189–203
6. Kim, K., Davis, L.S.: Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering. In: ECCV, Graz, Austria (2006)

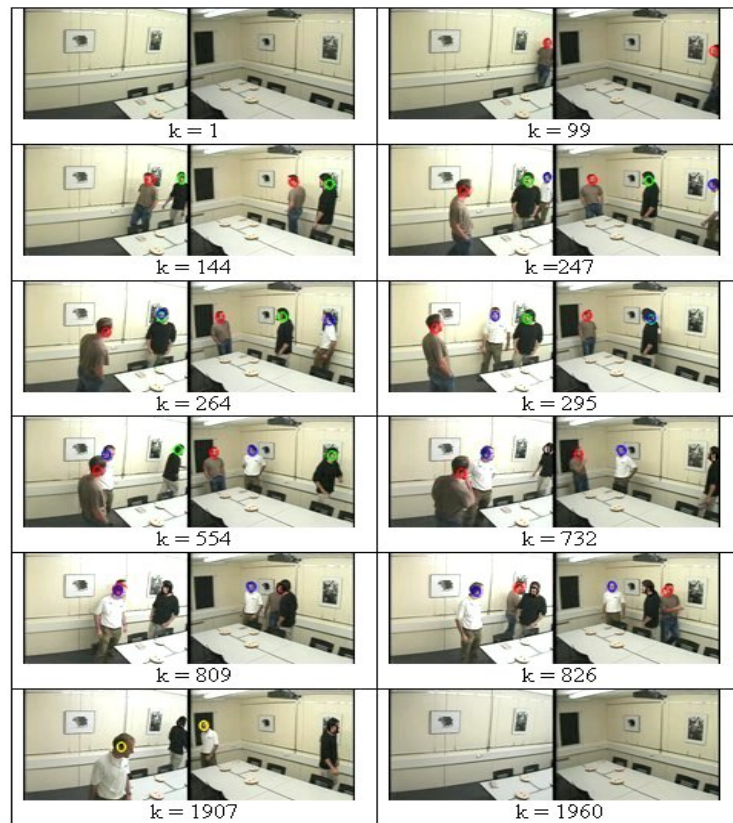


Fig. 3. Projection 3D estimations to two camera planes

7. Dockstader, S., Tekalp, A.M.: Multiple camera tracking of interacting and occluded human motion. *Proceedings of the IEEE* **89**(10) (2001)
8. Mahler, R.: Multi-target Bayes filtering via first-order multi-target moments. *IEEE Trans. on Aerospace and Electronic Systems* **39**(4) (2003) 1152–1178
9. Vo, B.N., Singh, S., Doucet, A.: Sequential Monte Carlo methods for Bayesian multi-target filtering with random finite sets. *IEEE Trans. Aerospace and Electronic Systems* **41**(4) (2005) 1224–1245
10. Vo, B.N., Ma, W.K.: The Gaussian mixture probability hypothesis density filter. *IEEE Transaction Signal Processing* **54**(11) (2006) 4091–4104
11. Comaniciu, D., Meer, P.: Mean shift analysis and applications. In: *ICCV*. (1999)
12. Pham, N.T., Huang, W.M., Ong, S.H.: Tracking multiple objects using probability hypothesis density filter and color measurements. In: *ICME*. (2007)
13. Clark, D., Panta, K., Vo, B.: The GM-PHD filter multiple target tracker. In: *Proceedings of FUSION 2006*, Florence (2006)
14. Lathoud, G., Odobez, J., Perez, D.: Av16.3: an audio-visual corpus for speaker localization and tracking. In: *Proceedings of the MLMI'04 Workshop*. (2004)