BAYESIAN INTEGRATION OF AUDIO AND VISUAL INFORMATION FOR MULTI-TARGET TRACKING USING A CB-MEMBER FILTER

Reza Hoseinnezhad* *Ba-Ngu Vo*† *Ba-Tuong Vo*† *David Suter*‡

* RMIT University, Victoria, Australia
† The University of Western Australia, WA, Australia
‡ The University of Adelaide, SA, Australia

ABSTRACT

A new method is presented for integration of audio and visual information in multiple target tracking applications. The proposed approach uses a Bayesian filtering formulation and exploits multi-Bernoulli random finite set approximations. The work presented in this paper is the first principled Bayesian estimation approach to solve the sensor fusion problems that involve intermittent sensory data (e.g. audio data for a person who occasionally speaks.) We have examined our method with case studies from the SPEVI database. The results show nearly perfect tracking of people not only when they are silent but also when they are not visible to the camera (but speaking).

Index Terms— audio-visual tracking, Bayesian filtering, random finite sets, finite set statistics, sensor fusion.

1. INTRODUCTION

Audio-visual multi-target tracking is an essential component of various applications such as monitoring people behaviour, traffic monitoring and smart rooms. In a Bayesian estimation framework, the states of the targets are predicted (based on a stochastic motion model for the targets) and updated (using the measurements) in each iteration. Sensor fusion naturally takes place in the update step where instead of the raw measurements, their likelihoods are combined. The main challenge here is that a target can be silent in periods of time and not detectable through the audio measurements, or it can be hidden from the camera while emitting sounds (detectable by the audio measurements but not appearing in the image).[1] It is important to note that in both cases we might still have clutter measurements for the undetectable target.

To integrate the sensory data, we need to efficiently combine the likelihoods in the update process of a Bayesian filter, in such a way that the intermittent nature of the sensory data are considered. Several solutions have appeared in the literature. The simplest one is to multiply the two likelihoods based on independence assumptions. But this will only work to track the targets which are visible and emit sounds at the same time (e.g. the *speaker* tracking application in [1]).

This paper focuses on applications where multiple *active* speakers are to be tracked (e.g. active participants in a round table discussion or multiple speakers lecturing to a silent audience). A straightforward solution is to multiply the audio and visual likelihoods but setting the likelihood to 1 if the modality is unavailable [2]. However, the availability itself needs to be determined and can be erroneous in presence of clutter measurements or measurements corresponding to other speaking and visible targets. The most common solution is to linearly combine the measurement likelihoods of the visual and audio observations where the weights of the combination are adjusted dynamically according to an acoustic confidence measure [3] or using separate confidence measures for the audio and video channels [4]. However, linear combination of the two likelihoods is mainly heuristic and not mathematically accurate.

In this paper, we present a principled approach to combine audio and video data in a Bayesian estimation framework. Our tracking method is formulated based on treating the states of multiple targets as a single random finite set (RFS) and using the finite set statistics (FISST) to formulate the prediction and update steps. The basic difference with other approaches is that an RFS formulation allows an elegant and rigorous modelling of targets birth and death as well as false measurements and missed detections. Our solution involves applying consecutive update steps, each time using a single source of sensory information (audio or visual). The major point of novelty lies in our implementation of the consecutive updates in an RFS framework based on modifying the Cardinality-Balanced MeMBer (CB-MeMBer) filter [5] and modelling the intermittency of the sensory information in terms of the detection probabilities. A sequential Monte Carlo implementation of the multi-Bernoulli approximation to the Bayesian filter is explained and examined in three challenging case studies from the SPEVI database.

This work was supported by the Australian Research Council through the ARC Discovery Project grants DP0880553 and DP0989007.

[1] All trackable targets are assumed to be never silent and invisible at the same time.

2. THE CB-MEMBER FILTER

Mahler's Finite Set Statistics (FISST) [6] has been recently recognised by the tracking community as an appropriate framework to formulate multi-target tracking solutions. In this framework, the multi-target state is modelled as a finite set. This modelling based on finite sets admits a mathematically consistent notion of estimation error since distance between sets is a well understood concept. In addition, since in a set the elements are not ordered, the filtering scheme works without the need for the data association problem to be explicitly solved. FISST provides practical mathematical tools for dealing with RFSs, based on a notion of integration and density that is consistent with point process theory. Using these tools, the prediction and update steps of Bayesian estimation of the posterior density of an RFS have been properly formulated [6, 7]. These steps constitute what is commonly known as the Bayes multi-target of multi-object filtering.

Among various RFS models used for implementation of a multi-object Bayesian filter, we employ a special type of RFS called *multi-Bernoulli* RFS, which is defined as the union of M independent Bernoulli RFSs $X^{(i)}$. Each Bernoulli RFS is either empty or a singleton with probabilities $1 - r^{(i)}$ and $r^{(i)}$, respectively. In case $X^{(i)}$ is a singleton, its only element is distributed according to a probability density $p^{(i)}(\cdot)$. Mahler [6] has shown that the parameter M, existence probabilities r_i and the distributions $p_i(\cdot)$ all together form a complete characterisation of the multi-Bernoulli RFS denoted by $X \sim \{(r^{(i)}, p^{(i)}(\cdot))\}_{i=1}^{M}$. With multi-Bernoulli assumptions, Mahler [6] derived the prediction and update steps of a particular implementation of the Bayesian filter, called the MeMBer filter. Vo et al. [5] later derived a modified version which involved unbiased cardinality and called it the Cardinality-Balanced MeMBer (CB-MeMBer) filter. Since the Bayes recursion is generally intractable, a sequential Monte Carlo implementation of the multi-Bernoulli filter is presented.

Suppose that at time $k-1$, the posterior density $\{(r_{k-1}^{(i)}, p_{k-1}^{(i)}(\cdot))\}_{i=1}^{M_{k-1}}$ is given. In the prediction step of the filter, the random finite set of targets evolves to a new multi-Bernoulli RFS including two ensemble of tracks associated with surviving and new born targets. Existence probabilities and distributions of these predicted targets are computed using a target death process modelled by a death probability, a target birth process modelled by a multi-Bernoulli RFS, and a target survival process modelled by a survival probability.[2] Let us denote the predicted multi-Bernoulli distribution by $\{(r_{k|k-1}^{(i)}, p_{k|k-1}^{(i)}(\cdot))\}_{i=1}^{M_{k|k-1}}$ where each predicted Bernoulli component density $p_{k|k-1}^{(i)}(\cdot)$ comprises $L_{k|k-1}^{(i)}$ particles, i.e. $p_{k|k-1}^{(i)}(x) = \sum_{j=1}^{L_{k|k-1}^{(i)}} w_{k|k-1}^{(i,j)} \delta_{x_{k|k-1}^{(i,j)}}(x)$.

We consider a measurement model in the form of the likelihood function $g_k(z|x)$ where z is a point measurement corresponding to a single target with the state x. The measurement model also includes a probability of detection denoted by $p_{D,k}(x)$ and the clutter measurements modelled as a Poisson RFS (with Poisson distributed cardinality with its mean denoted by $\kappa_k(z)$). The updated RFS comprises the union of two multi-Bernoulli sets: $\pi_k = \{(r_{L,k}^{(i)}, p_{L,k}^{(i)}(\cdot))\}_{i=1}^{M_{k|k-1}} \cup \{(r_{U,k}(z), p_{U,k}(\cdot;z))\}_{z \in Z_k}$. The first set, called *legacy tracks*, includes the parameters of the undetected targets. The second set is called *measurement-corrected tracks* and includes the parameters of detected targets modified according to the measurements. The parameters of the two tracks are given by [6, 5]:

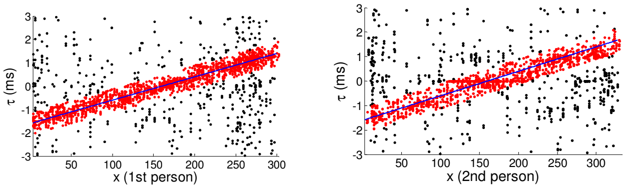
$$
\begin{aligned}
r_{L,k}^{(i)} &= r_{k|k-1}^{(i)}(1 - \varrho_{L,k}^{(i)})/(1 - r_{k|k-1}^{(i)}\varrho_{L,k}^{(i)}) \\
p_{L,k}^{(i)}(x) &= \textstyle\sum_{j=1}^{L_{k|k-1}^{(i)}} \tilde{w}_{L,k}^{(i,j)} \delta_{x_{k|k-1}^{(i,j)}}(x) \\
r_{U,k}(z) &= \frac{\sum_{i=1}^{M_{k|k-1}} \frac{r_{k|k-1}^{(i)}(1-r_{k|k-1}^{(i)})\varrho_{U,k}^{(i)}(z)}{(1-r_{k|k-1}^{(i)}\varrho_{L,k}^{(i)})^2}}{\kappa_k(z)+\sum_{i=1}^{M_{k|k-1}} \frac{r_{k|k-1}^{(i)}\varrho_{U,k}^{(i)}(z)}{1-r_{k|k-1}^{(i)}\varrho_{L,k}^{(i)}}} \\
p_{U,k}(x;z) &= \textstyle\sum_{i=1}^{M_{k|k-1}} \sum_{j=1}^{L_{k|k-1}^{(i)}} \tilde{w}_{U,k}^{(i,j)}(z)\delta_{x_{k|k-1}^{(i,j)}}(x) \\
\varrho_{L,k}^{(i)} &= \textstyle\sum_{j=1}^{L_{k|k-1}^{(i)}} w_{k|k-1}^{(i,j)} p_{D,k}(x_{k|k-1}^{(i,j)}) \\
\tilde{w}_{L,k}^{(i,j)} &= w_{L,k}^{(i,j)} \Big/ \textstyle\sum_{j'=1}^{L_{k|k-1}^{(i)}} w_{L,k}^{(i,j')} \\
w_{L,k}^{(i,j)} &= w_{k|k-1}^{(i,j)}(1 - p_{D,k}(x_{k|k-1}^{(i,j)})) \\
\varrho_{U,k}^{(i)}(z) &= \textstyle\sum_{j=1}^{L_{k|k-1}^{(i)}} w_{k|k-1}^{(i,j)} p_{D,k}(x_{k|k-1}^{(i,j)}) g_k(z|x_{k|k-1}^{(i,j)}) \\
\tilde{w}_{U,k}^{(i,j)}(z) &= w_{U,k}^{(i,j)}(z) \Big/ \textstyle\sum_{i'=1}^{M_{k|k-1}} \sum_{j'=1}^{L_{k|k-1}^{(i')}} w_{U,k}^{(i',j')}(z) \\
w_{U,k}^{(i,j)}(z) &= \frac{w_{k|k-1}^{(i,j)} r_{k|k-1}^{(i)} p_{D,k}(x_{k|k-1}^{(i,j)}) g_k(z|x_{k|k-1}^{(i,j)})}{1-r_{k|k-1}^{(i)}}.
\end{aligned}
$$

In order to avoid numerical explosion, after the update step, the Bernoulli targets (tracks) with very small probabilities of existence are removed. The CB-MeMBer filtering algorithm also includes resampling the particles for each track and merging the tracks that are very close to each other. See [5] for details.

3. AUDIO-VISUAL TRACKING

In our implementation of the CB-MeMBer filter, we use the constant-velocity model as motion model [5] and the state of each target in the image includes the location and dimensions of a rectangular blob containing the target as well as the location derivatives, i.e. $\mathbf{x} = [x_{\text{im.}}\ y_{\text{im.}}\ \dot{x}_{\text{im.}}\ \dot{y}_{\text{im.}}\ w_{\text{im.}}\ h_{\text{im.}}]^\top$. The audio and visual signals need to be processed so as to extract information pertaining to the target states. For video signals, we use the kernel-based background subtraction method [9] followed by a number of morphological image operations. The result is a set of rectangular blobs in each frame characterised by their image locations and dimensions

[2] Due to space limits, details of prediction step are not presented here. Interested readers can refer to [6, 5, 8].



(a)

(b)

Fig. 1. Working out the linear relationship between the target x locations and TDOA measurements. There are two targets in the scene and the line equations are consistently the same for both persons as shown in (a) and (b). The red points are the inliers segmented by the HBM robust estimator and the rest (black points) are outliers.

in pixels. The set of visual point measurements is denoted by $Z_v = \{z_{v_i}\}$ where each point measurement is formed as $z_v = [x_{\text{im.}}\ y_{\text{im.}}\ w_{\text{im.}}\ h_{\text{im.}}]^\top$ and with Gaussian noise assumptions, the visual measurement likelihood function is given by $g_v(z_v|x) = \mathcal{N}(z_v; C_v x, \sigma_v^2)$ where σ_v is the scale of noise and $\mathcal{N}(z; \mu, \sigma^2) \triangleq \exp(-(z-\mu)^2/(2\sigma^2))/(\sqrt{2\pi}\sigma)$ and $C_v = \text{diag}(1, 1, 0, 0, 1, 1)$.

The audio signals from the microphones on two sides of the camera are processed to compute the time difference of arrival (TDOA). The processing involves computation of cross-correlation between the signals using the Generalised Cross Correlation function - Phase Transform (GCC-PHAT). Due to reverberation effects, there are usually several peaks in the GCC-PHAT curve plotted versus time difference. In our experiments which involve tracking of up to two people, we pick at most five largest peaks of the GCC-PHAT values and consider them as TDOA measurements in each frame. Some of these are clutter measurements. Since the distance of the targets from the microphones is large compared to the distance between the two microphones, there is an approximately linear relationship between the $x_{\text{im.}}$ location of a target and its corresponding TDOA [2]. To estimate the parameters of this linearity (calibration of the audio sensors), we have plotted all the TDOA measurements versus the ground truth x-coordinates of the targets in the image (see Fig. 1). The results are plotted for one of the three case studies from the SPEVI database involving two targets.[3]

It is important to note that many of the data points plotted in Fig. 1 are not relevant to the target state. Indeed, only one of the maximum five TDOA's measured in each frame correspond to a target location, the rest are either irrelevant peaks (due to reverberation effects) or correspond to the other target. To calibrate the audio sensor, we need to detect such points (outliers) and remove them before estimating the line parameters. For this purpose, we have used a high-breakdown robust estimator called HBM [10]. Figures 1(a) and (b) show that the lines estimated for each of the targets are almost identical which demonstrates the accuracy of calibration. If $z_a = \alpha x_{\text{im.}} + \beta$ is the estimated line equation, then audio measurement likelihood function is given by $g_a(z_a|x) = \mathcal{N}(z_a; C_a x + \beta, \sigma_a^2)$ where σ_a is the scale of noise and $C_a = [\alpha\ 0\ 0\ 0\ 0\ 0]$.

3.1. Sensor fusion

In order to integrate the information provided by audio and visual sensors in a Bayesian estimation framework, the update step of the CB-MeMBer filter is run twice, first using the visual measurements then audio measurements. The important point to note here is that detection probability for each sensor is determined based on our definition of "active speaker". For instance if an active speaker is considered to be a person who is expected to be visible to the camera in no less than 95% of the time and to be speaking in at least 40% of the time, then we set $p_{D_v} = 0.95$ and $p_{D_a} = 0.40$.

When the detection probability is close to one, most of legacy tracks are assigned very small existence probabilities – see the update equations. Thus, in the first round of update (using visual cues), most of the legacy tracks almost *die* and few of them are passed to get updated using the audio cues along with the measurement-corrected tracks. In this round of update, they evolve to a new set of legacy and measurement-corrected tracks. Since the audio detection probability is not very close to one, some legacy tracks can have large existence probabilities, representing the silent targets. More precisely, the targets which are visible to camera but are occasionally silent will be tracked by this method. On the other hand, the targets that are occasionally invisible to camera will be tracked as long as they speak. This is because their corresponding tracks will be among the few legacy tracks that survive through the first round of update. The existence probabilities of such tracks will be increased in the second round of update, because they will be associated to audio cues in the sensory data.

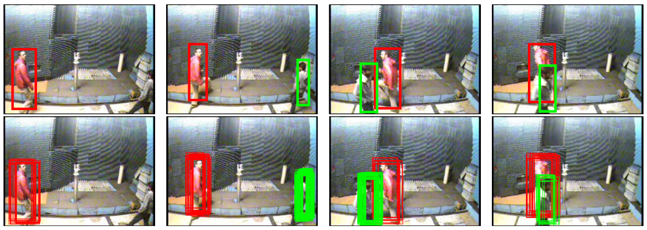
4. SIMULATION RESULTS

We have examined the ability of our method to track speakers in three audio-visual sequences from the SPEVI database. Figures 2–3 show snapshots of the tracking results in two of the sequences.[4] For the first sequence shown in in Fig. 2, we have shown the particle blobs as well as the final estimates.

The results show nearly perfect tracking performance. Indeed, in 98.5% of all the frames, the existing targets are all detected, correctly labeled and tracked. Labels are never switched after or during occlusions, and an invisible target is successfully tracked using the audio cues. The superior tracking performance of our method is due to the principled approach to the modelling of the intermittency of sensory information in terms of detection probabilities using random

[3] `http://www.eecs.qmul.ac.uk/~andrea/spevi.html`

[4] Videos showing the tracking results in three cases can be downloaded from: `www.ee.uwa.edu.au/~bnvo/icassp11`.



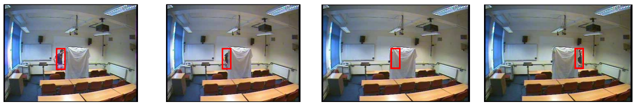
Frame 115 Frame 127 Frame 296 Frame 301

Fig. 2. Tracking results for sequence 1.

Table 1. Quantitative comparison of tracking performance. The label switching rate is not applicable (n/a) to sequence 2 where a single target is tracked.

	Without Audio			With Audio		
	FNR	FAR	LSR	FNR	FAR	LSR
Seq. 1	9%	2%	4%	3%	0%	0%
Seq. 2	32%	3%	n/a	5%	0%	n/a
Seq. 3	11%	2%	3%	2%	0%	0%

finite set theory as well as efficient solution to the multi-object filtering problem.



Frame 237 Frame 252 Frame 264 Frame 312

Fig. 3. Tracking results for sequence 2.

To show the effect of fusion of visual and audio information, we have quantified the detection and tracking errors via computing three quantities in our experiments, once with fusion of audio and visual information and once without the audio information. The quantities include ratio of missed targets (called false negative rate or FNR for short), the ratio of wrong detections (called false alarm rate or FAR for short) and the ratio of label switching events (called label switching rate or LSR for short) over all frames. The results are listed in Table 1 and present substantial improvement in detecting and tracking the targets when audio and visual information are integrated.

5. CONCLUSIONS

A new method for audio-visual tracking of multiple targets was proposed. The method is formulated in a random finite set framework based on multi-Bernoulli approximations, and implemented using sequential Monte Carlo techniques. Audio and visual cues are integrated by multiple updates. The random finite set formulation allows a natural and principled way to model the intermittent nature of sensory data (mainly audio).

Simulation results show that the proposed method almost perfectly tracks multiple interacting targets, not only when they are silent, but also in times when they are invisible to the camera.

6. REFERENCES

[1] Andrew Rae, Alaa Khamis, Otman Basir, and Mohamed Kamel, “Particle filtering for bearing-only audio-visual speaker detection and tracking,” in *Proc. Int. Conf. Signals, Circuits and Systems (SCS’09)*, Medenine, Tunisia, 2009.

[2] Matteo Bregonzio, Murtaza Taj, and Andrea Cavallaro, “Multi-modal particle filtering tracking using appearance, motion and audio likelihoods,” in *ICIP’07*, San Antonio, TX, United states, 2006, vol. 5, pp. V33–V36.

[3] R. Stiefelhagen, K. Bernardin, H. K. Ekenel, J. McDonough, K. Nickel, M. Voit, and M. Wölfel, “Audio-visual perception of a lecturer in a smart seminar room,” *Signal Processing*, vol. 86, no. 12, pp. 3518–3533, 2006.

[4] M. Taj and A. Cavallaro, “Audio-assisted trajectory estimation in non-overlapping multi-camera networks,” in *ICASSP 2009*, Taipei, Taiwan, 2009, pp. 3517–20.

[5] Ba-Tuong Vo, Ba-Ngu Vo, and Antonio Cantoni, “The cardinality balanced multi-target multi-Bernoulli filter and its implementations,” *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 409–423, 2009.

[6] R.P.S. Mahler, *Statistical multisource-multitarget information fusion*, Artech House, Norwood, MA, USA, 2007.

[7] B.-N. Vo, S. Singh, and A. Doucet, “Sequential Monte Carlo methods for multi-target filtering with random finite sets,” *IEEE Tran. AES*, vol. 41, no. 4, pp. 1224–1245, 2005.

[8] B.-N. Vo, B.-T. Vo, N.-T. Pham, and D. Suter, “Joint detection and estimation of multiple objects from image observations,” *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5129–5141, 2010.

[9] Ahmed Elgammal, Ramani Duraiswami, David Harwood, and Larry S. Davis, “Background and foreground modeling using nonparametric kernel density estimation for visual surveillance,” *Proceedings of the IEEE*, vol. 90, no. 7, pp. 1151 – 1162, 2002.

[10] Reza Hoseinnezhad and Alireza Bab-Hadiashar, “A novel high breakdown m-estimator for visual data segmentation,” in *IEEE International Conference on Computer Vision (ICCV’2007)*, Rio de Janeiro, Brazil, October 2007, Digital Object Identifier: 10.1109/ICCV.2007.4408971.