Person tracking using audio and depth cues

Qingju Liu, Teofilo de Campos, Wenwu Wang, Philip Jackson and Adrian Hilton
CVSSP, University of Surrey, Guildford GU2 7XH, UK
{q.liu, t.decampos, w.wang, p.jackson, a.hilton}@surrey.ac.uk

Abstract

In this paper, a novel probabilistic Bayesian tracking scheme is proposed and applied to bimodal measurements consisting of tracking results from the depth sensor and audio recordings collected using binaural microphones. We use random finite sets to cope with varying number of tracking targets. A measurement-driven birth process is integrated to quickly localize any emerging person. A new bimodal fusion method that prioritizes the most confident modality is employed. The approach was tested on real room recordings and experimental results show that the proposed combination of audio and depth outperforms individual modalities, particularly when there are multiple people talking simultaneously and when occlusions are frequent.

1. Introduction

Person tracking has been extensively studied in the field of computer vision, with various applications ranging from surveillance, video retrieval and teleconferencing to human-computer interactive activities such as video games. Person tracking can be applied to different modalities, e.g. RGB images [3, 11, 8], acoustic recordings [24, 14, 15, 6], depth sensors [13, 22, 17], GPS and thermal sensors. There is a consensus that different modalities are complementary to each other, which has motivated an increasing interest in cross-modal tracking in the last decade. Most of these works are done in the audio-visual domain [25, 9, 10]. Combination of other modalities has recently started to become more popular. For instance, [16, 23] tracks person from both laser range and camera data; the work in [18] fuses RGB, depth and thermal features; [27] reconstructs 3D scenes involving transparent objects, using a depth camera and an ultrasonic sensor. Yet, there are some essential limitations associated with the existing mono- or cross-modal person tracking methods. The mono-modal tracking is not robust enough, while the cross-modal methods often require a high hardware load.

To address the above limitations, we implemented a bimodal person tracking algorithm that combines depth and audio cues. A time-of-flight depth sensor, i.e. Kinect for Windows v2.0 [17], as well as a pair of binaural microphones, i.e. Cortex Manikin MK2 binaural head and torso simulator, are used for person tracking, which are respectively denoted as Kinect2 and Cortex MK2, as shown on the top right and top left of Fig. 1. Individually, both modalities have issues. Audio measurements from Cortex MK2 suffer from heavy background noise and the non-stationary nature of speech. Moreover, they are unable to disambiguate between front and rear sound sources. On the other hand, depth cues from Kinect2 are affected by occlusions. By exploiting the complementary between these two modalities, our tracking method becomes more robust. Based on the random finite set (RFS) theory, we propose a full-probabilistic model for multi-person tracking. Particle filters are implemented based on Bayesian filtering [5, 2].

There are several contributions in our method. Firstly, this is a seminal work in the fusion of audio and depth cues. Secondly, the proposed method balances the bimodal difference in their structures and robustness, which evaluates the validity of both streams and prioritizes the most confident modality. Thirdly, a measurement driven birth model is used to quickly localize any emerging person.

The remainder of the paper is organized as follows. Section 2 briefly introduces the RFS particle filters in person tracking. Section 3 presents the overall frame work of our proposed algorithm, and describes in detail the fusion of depth and audio streams. Experimental results are shown and analyzed in Section 4. Finally, conclusions and insights for future research directions are raised in Section 5.

2. RFS-based particle filters

For single target tracking, the hidden state at time frame k, e.g. the position and velocity of the target, is often represented via a vector $\mathbf{x}_k$. To generalize this problem to the multi-object tracking problem, the hidden state is a finite-set-valued variable $X_k = \{\mathbf{x}_{k,1}, \cdots, \mathbf{x}_{k,N_k}\}$ that contains N_k targets, with each $\mathbf{x}_{k,i}$ being the state vector associated with the i-th target. When $N_k = 0$, $X_k = \emptyset$ denotes no target being detected.

X_k can be estimated from a sequence of measure-

ments $[Z_1, Z_2, \cdots, Z_k]$ collected/extracted from the sensors, where $Z_k = \{\mathbf{z}_{k,1}, \cdots, \mathbf{z}_{k,M_k}\}$ is also a finite-set-valued variable. Note that M_k does not necessarily equal N_k, and $\mathbf{x}_{k,i}$ is not necessarily associated with $\mathbf{z}_{k,i}$. Some measurements are clutter (false alarms) and some targets may fail to generate any measurement.

Bayesian filtering [5, 2] is often applied in target tracking, which propagates the posterior density over time with a recursive prediction and update process. It exploits the temporal involvement as well as the relationship between the underlying positions and the measurements, i.e. the state-space approach. However, this problem might be intractable if the state-space model does not satisfy certain restrictions. Sequential Monte Carlo (SMC) [5] methods can be devoted to its approximations, resulting the so-called particle filters or bootstrap filters [1]. In multi-target tracking, the random finite set (RFS) approach can be used, which takes into account the association uncertainty as well as spurious measurements. More details on RFS-based particle filters are available in [15].

3. Proposed method

Particle filters are applied to a sequence of measurements for target tracking. These measurements are often features extracted from the sensors, which are related to the underlying target state. In this paper, we exploit the complementary relationship between the audio and depth streams, which are collected by Cortex MK2 and Kinect2 respectively. A novel bimodal person tracking scheme is proposed, whose main flow is shown in Figure 1.

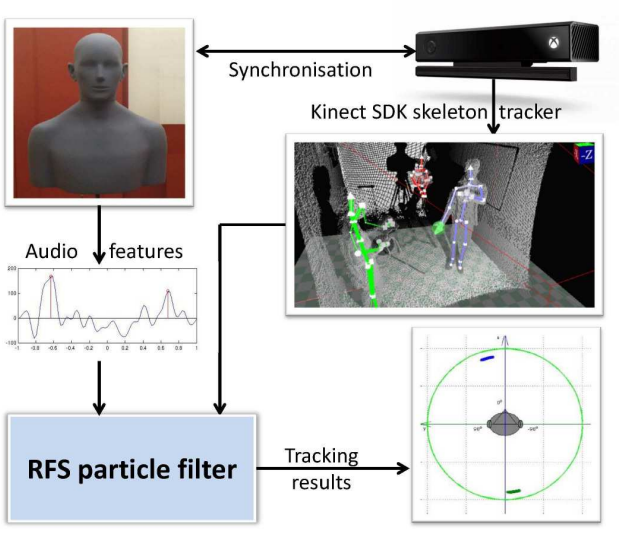

Figure 1. Flow of the proposed audio-depth person tracking method. Synchronized audio and depth measurements are collected from Cortex MK2 and Kinect2 respectively. A RFS particle filter is then employed to these synchronized measurements for person tracking.

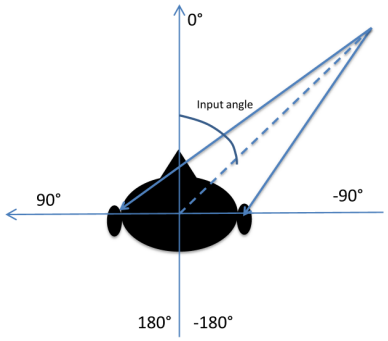

Figure 2. The input angle (azimuth) of a target source in the horizontal plane. A sound source arrives at the two ears via different paths, resulting an inter-aural time difference. The input angle increases from 0° from the nose anti-clockwisely.

We aim to find the relative angle of any person to Cortex MK2 in the horizontal plane, i.e. the azimuth direction, as shown in Figure 2. Being a one-dimensional measurement, the azimuth direction is not as informative as 3D position, but it is of great importance to attention switching in machine audition (e.g. for source separation and objectification) or as an auxiliary measurement to handle occlusions in computer vision. Azimuth estimation can be challenging and in this paper we demonstrate the advantages of bimodal tracking over mono-modalities.

In what follows, we will provide details about the measurements methods and how they are fused together.

3.1. Audio-based likelihood function

The time delay of arrival (TDOA) cues are used as audio measurements in our method. The phase transform (PHAT)-GCC [12] method is applied to Cortex MK2 binaural recordings. Suppose $L_k(\omega)$ and $R_k(\omega)$ are the short time Fourier transforms (STFT) of the two audio segments at time k. The PHAT-GCC function can be applied as:

$$C(\tau) = \int_{-\infty}^{\infty} \frac{L_k(\omega) R_k^*(\omega)}{|L_k(\omega) R_k^*(\omega)|} e^{j\omega\tau}\, d\omega, \qquad (1)$$

where the superscript $*$ denotes the conjugate operator and $|\cdot|$ is a modulus operator. By finding peak positions in PHAT-GCC, M_k^{a} TDOAs $Z_k^{\mathrm{a}} = \{\tau_{k,1}, \cdots, \tau_{k,M_k^{\mathrm{a}}}\}$ can be obtained as the audio measurements[1].

Different positions (azimuths) yield different TDOAs. We need to model the relationship between the audio measurement with the azimuth, i.e. the audio likelihood function, which is complex due to reflections and diffraction of the head. From off-line training, we notice there exists a nonlinear relationship between the resultant TDOA τ and the azimuth α, as shown in Figure 3. Firstly, the curve is

[1]The superscript a indicates audio. Similarly, the superscript d stands for depth, and ad denotes audio-depth.

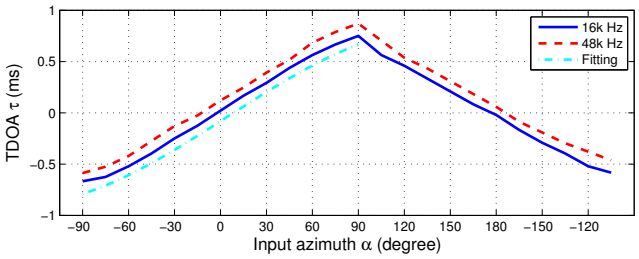

Figure 3. Illustration of the relationship between the resultant TDOA τ with the azimuth α. This was trained from off-line recordings at 16 kHz and 48 kHz. The third-order polynomial curve fitting is applied to model the audio likelihood function. We lifted the curve at 48 kHz, and lowered the fitted curve via a shift of 0.1 either way to improve visualization.

symmetric through the axis of $90°$ or $-90°$. This is quite understandable as TDOAs are ambiguous between front and back. Secondly, TDOA from the front can be linearly fitted with the input azimuth (from $-90°$ to $90°$) using the polynomial fitting:

$$\tau = f(\alpha) = p_1\alpha + p_3\alpha^3 \,, \tag{2}$$

and $p_1 = 2.405\times10^{-6}$ and $p_3 = 1.807\times10^{-2}$ are obtained in the off-line training process.

For an azimuth from the back, a mapping function $map(\cdot)$ can be applied to get its mirror reflection:

$$map(\alpha) = \begin{cases} \alpha, & \text{if } |\alpha| \leq 90°, \\ sign(\alpha)(180° - |\alpha|) & \text{otherwise.} \end{cases} \tag{3}$$

Considering zero-mean additive Gaussian noise with variance $\delta^{\mathrm{a},2}$, we can model the audio likelihood as:

$$g(\tau|\alpha) = \mathcal{N}(\tau - f(map(\alpha))|0, \delta^{\mathrm{a},2}), \tag{4}$$

where $\mathcal{N}(\cdot)$ denotes the Gaussian distribution. This noise term also relaxes the non-perfect fitting in Equation (2).

3.2. Depth-based likelihood function

As mentioned before, to get depth measurements, we used the Kinect for Windows v2.0 time of flight sensor, dubbed as Kinect2 in this paper. This sensor emits near infra-red pulses and a fast infrared camera estimates depth based on phase difference. The SDK provided by Microsoft [17] includes a method that detects up to six people and estimates their pose based on a skeleton model with 25 joints. It is based on a method that classifies each point in a point cloud into a body part (hand, arm, elbow, forearm, etc.) or as background, where their features do not match a body part. This is done using simple depth comparison features and a random decision forest (RDF). This RDF is trained in with millions of samples of humans, combining real and synthetic images, at a wide range of poses, with ground truth labels annotated for each body part. The resulting labeled point cloud is spatially filtered and a post-processing method fits up to six plausible human skeleton models to the scene and generates the 3D position and orientation for each body joint.

The center right sub-plot in Figure 1 shows detected skeletons in a sample point cloud. Since our goal is to objectify speakers, we are interested in the location of their mouth, which is close to the center of their head. We thus use the position of the heads detected by Kinect2 SDK as the 3D position of the sound sources.

A number of methods have been proposed to detect and track people in depth images [26, 19], particularly those generated using sensors based on structured light projection, such as the first version of Kinect. Although the full pipeline implemented in Kinect2 SDK has not been disclosed, we have performed a set of preliminary experiments comparing this method with other state of the art implementations available for 3D head tracking from depth measurements, such as the method of Fanelli et al. [7] and RGB methods, such as that of Saragih et al. [21]. Our qualitative observations indicate that the method implemented in Kinect2 SDK robustly achieves state-of-the-art accuracy in head position estimation. Since it has been designed to work on living rooms, the range of distances where it operates is optimal for our application, whereas other implementations available off-the-shelf have been optimized to be used on web-cam scenarios, with a much smaller working distance range.

Despite its robustness, this method has some drawbacks. Since it is based on a *tracking-as-detection* framework, it does not incorporate a mechanism to handle occlusions based on inference from tracking results. Occlusions cause this implementation to lose measurements or to generate noise outliers and to swap the identity of people being tracked, as shown in Figure 4. It can also fail due to the limits of its operational range: its field of view is of $70.6°$ and the working distance between the sensor and targets ranges from 1.2 meters to approximately 4.5 meters [17]. It also fails in cluttered scenes or when people are close to each other.

Further to detecting people, the Kinect2 skeleton detector also locates the binaural head and torso simulator (Cortex MK2) automatically as a static person sitting at the center of the room. By having the prior knowledge that the Cortex MK2 is the audio recording device and that it remains static, we can easily detect it by analyzing a sequence of recordings in a pre-processing step. This enables us to label it as a dummy and distinguish it from moving people. It also enables us to project the 3D position of detected humans to the polar coordinate system centered at Cortex MK2. Since the head position is estimated in 3D from Kinect2's viewpoint, there is no front/back ambigu-

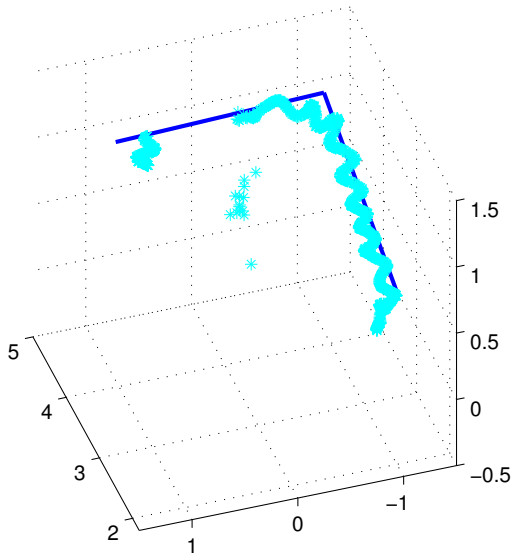

Figure 4. Kinect2 tracking results for a person (Actor B) following an L-shaped trajectory in the room of Figure 5. The target path that this person followed is highlighted in blue lines. The cyan stars show detected positions, which wiggles because the head actually swings from side to side as this person walked. There is a cluster of misdetected positions, i.e. this subject's head was detected around the dummy head position when he was occluded by the dummy. Shortly after that, there was a period of consecutive frames where the target is not detected because of this occlusion.

ity w.r.t. MK2 and the mapping of Equation 3 is not necessary for depth-based cues. The obtained azimuth angle measurements are used as depth-based observations, denoted as $Z_k^{\mathrm{d}} = \{\theta_{k,1}, \cdots, \theta_{k,M_k^{\mathrm{d}}}\}$, i.e., in the remainder of this paper, we assume that the pipeline that maps from depth images to azimuth angles relative to Cortex MK2 is part of the measurement process.

As mentioned earlier, the head tracker is usually reliable, but occlusions introduce severe noise or missing depth-based measurements, which we approximate using the zero-mean additive Gaussian noise with variance, defined as $\delta^{\mathrm{d},2}$. Therefore, the likelihood of the associated Kinect detection given an input angle follows

$$g(\theta|\alpha) = \mathcal{N}(\theta - \alpha|0, \delta^{\mathrm{d},2}) \,. \tag{5}$$

3.3. Audio-depth fusion

As introduced in Section 2, RFS particle filters are applied to the audio and depth measurements. Their state space model contains two essential parts: the dynamic model and the measurement model.

3.3.1 Dynamic model

The dynamic model describes the temporal evolution of target states. For multi-targets, each state vector $\mathbf{x}_k \in \mathcal{X}_k$ at frame k can either survive with probability P_s or die with probability $1 - P_s$ at the next frame. Let $\mathbf{x}_k$ contain the input angle α and the angular velocity $\dot{\alpha}$; the Largeiven model can be utilized to model the relationship between a survived target $\mathbf{x}_{k+1}$ and its previous state:

$$\mathbf{x}_{k+1} = \begin{bmatrix} 1 & T \\ 0 & e^{-\beta T} \end{bmatrix} \mathbf{x}_k + \begin{bmatrix} 0 \\ \nu\sqrt{1 - e^{-2\beta T}}\mathcal{N}(\cdot|0,1) \end{bmatrix}, \tag{6}$$

where T is the time duration between two consecutive frames; β and ν parametrize the motion model.

Moreover, a new target may be born in the searching field with probability P_b. To quickly localize any appearing target, we propose a measurements-driven target birth model as follows.

The current measurements $\mathcal{Z}_k$ can be mapped to a group of azimuths. We assume the birth model as is mixture of Gaussian kernels, whose mean and standard deviation are these mapped azimuths and 0.1 m. The velocity of newborn targets is zero. Following that distribution, newborn targets are enforced to those potential positions yielding the current measurements. The proposed method can therefore quickly localize any emerging target. A similar idea of adaptive target birth intensity is used in [20].

3.3.2 Measurement model

The measurement or observation model describes the relationship between the target state and the measurement. From sections 3.1 and 3.2, we know the relationship between a single observed mono-modal feature and its associated single-target state. However, for cross-modal multi-person tracking, we need $g(Z_k|X_k)$, where both the bi-modal feature Z_k and the multiple-target state X_k are set-valued variables. From empirical study, we notice the azimuth estimates based on depth data alone has far fewer outliers as compared to the audio stream. As a result, a depth-dominant fusion scheme is proposed.

We assume there are up-to-two people in the searching field. As a result, the hidden target state X_k can either be $\emptyset$, $\{\mathbf{x}_{k,1}\}$ or $\{\mathbf{x}_{k,1}, \mathbf{x}_{k,2}\}$.

When there is no detected target, i.e. $X_k = \emptyset$,

$$g(Z_k|\emptyset) = \left(\frac{P_c^{\mathrm{a}}}{2\tau_{max}}\right)^{|Z_k^{\mathrm{a}}|_0} \left(\frac{P_c^{\mathrm{d}}}{360}\right)^{|Z_k^{\mathrm{d}}|_0}, \tag{7}$$

where P_c^{a} and P_c^{d} are the expected number of false alarms at each frame, and $|\cdot|_0$ computes the cardinality of a set.

When there is one detected target, i.e. $X_k = \{\mathbf{x}_{k,1}\}$,

$$g(Z_k|\{\mathbf{x}_{k,1}\}) = p(Z_k|\emptyset)((1 - P_d) + P_d g^{\mathrm{ad}}(\mathbf{x}_{k,1})), \tag{8}$$

where $g^{\mathrm{ad}}(\mathbf{x}_{k,1}) = \max(g^{\mathrm{a}}(\mathbf{x}_{k,1}), g^{\mathrm{d}}(\mathbf{x}_{k,1}))$ with $g^{\mathrm{a}}(\mathbf{x}_{k,1}) = \max_{\mathbf{z}\in Z_k^{\mathrm{a}}} \frac{g(\mathbf{z}|\mathbf{x}_{k,1})2\tau_{max}}{P_c^{\mathrm{a}}}$ using Equation (4), and $g^{\mathrm{d}}(\mathbf{x}_{k,1})$ similarly using Equation (5). P_d is the chance a target being detected.

When there are two detected targets,

$$\begin{aligned} g(Z_k|\{\mathbf{x}_{k,1}, \mathbf{x}_{k,2}\}) = p(Z_k|\emptyset)((1-P_d)^2 \\ +P_d(1-P_d)g^{\mathrm{ad}}(\mathbf{x}_{k,1}) \\ +P_d(1-P_d)g^{\mathrm{ad}}(\mathbf{x}_{k,2}) \\ +P_d^2 g^{\mathrm{ad}}(\mathbf{x}_{k,1})g^{\mathrm{ad}}(\mathbf{x}_{k,2})). \end{aligned} \tag{9}$$

Computational complexity of the above full-probabilistic model becomes much greater with an increasing number of targets. For efficiency reasons, we therefore constrain our implementation to track up to two moving targets. To prioritize the depth stream, we make P_c^{a} greater than P_c^{d}. We also evaluate the validity of the audio stream in each frame via straightforward energy thresholding. If the audio frame is invalid, i.e., the speech energy is very low, our model is dynamically pruned by keeping only the depth term in Equations (7-9).

4. Experiments

4.1. Recording setup

Our testbed is a TV/film studio set built following professional media production standards, with furniture and features of a relatively large hallway whose dimensions are very similar to those of a typical living room: $244 \times 396 \times 242$ cm. As with typical TV/film production sets, its ceiling and one of the walls are missing, though this set was assembled inside a larger room. The reverberation time of this room is about 430 ms. In the recordings for our experiments, the binaural microphone (Cortex MK2) stood in the center of the room with ear height of 165 cm. The depth sensor was placed around the center at the height of 170 cm, 329 cm away from the depth sensor, as shown in Figure 5. The sampling rate for audio is $Fs^{\mathrm{a}} = 44.1$ kHz. The depth-based head tracker has a sampling rate of $Fs^{\mathrm{d}} = 27.43$ Hz. We used hand clapping at the beginning and end of each recording session to synchronize these two streams. The hand claps can be detected from the audio stream via energy thresholding, and arm pose detection using skeletal tracker from the depth stream.

Three sequences were recorded in total about 7.5 minutes, involving two actors: Actor A is a male, with height of 1.82 m and Actor B is a female, 1.58 m. In the first sequence, Actor A started at Position 1 (labeled with a yellow circle in Figure 5), facing the center, walking slowly along the gray circular trajectory anti-clock-wisely, reading randomly-selected sentences from the TIMIT database. He walked back clock-wisely along the gray circle when reaching Position 24. Actress B repeated this process with a higher speed, and this was recorded in the second sequence. In the third sequence, Actor A started at Position A, walking along the path $A \rightarrow B \rightarrow A \rightarrow D \rightarrow A$, facing forward. At the same time Actress B started at Position C, walking along the path $C \rightarrow D \rightarrow C \rightarrow B \rightarrow C$, facing forward. Therefore, both actors followed L-shaped paths (symmetric to each other, relative to the room), moving independently from each other, each walking at his/her preferred pace while reading the material mentioned earlier. This dataset is available from [4].

4.2. Implementation details

To obtain audio measurements, 8192-point (approximately 186 ms) Hamming windowed STFT with 0.75 overlap is applied. The time length between two neighboring audio frames is therefore $T = 139$ ms. The candidate τ is linearly sampled in the range of -1 ms to 1 ms ($\tau_{max} = 1$ ms) with the resolution of $1/Fs^{\mathrm{a}}$. At each time frame, at most two TDOAs are extracted as audio measurements.

Figure 6 shows the extracted audio measurements from Sequences 1 and 3. Sequence 1 has only one speaker facing the binaural microphone, while Sequence 3 has two speakers and they do not face the microphone most of the time, In addition, Sequence 3 contains heavy background noise.

To implement RFS-particle filters, the following parameters are used. The target survival chance is $P_s = 0.99$, and a target birth chance is $P_b = 0.02$. Parameters in the Langevin model are set as $\beta = 10, \nu = 10$. The target detection probability is $P_d = 0.75$, and the false alarm expectations are $P_c^{\mathrm{a}} = 0.5$ and $P_c^{\mathrm{d}} = 0.1$. The mono-modal likelihood functions in Equations 4 and 5 have the standard variance of $\delta^{\mathrm{a}} = 1/16$ ms and $\delta^{\mathrm{d}} = 5^\circ$.

4.3. Results and analysis

4.3.1 Single person, audio-only features

Firstly, we tested the proposed algorithm on the first two sequences, using only audio features. In Sequences 1 and 2, only less than 10 seconds occlusions is observed. In addition, when there is no occlusion, very accurate depth-based tracking results are obtained except for only a few frames of outlier. As a result, we manually corrected these outlier and labeled the misdetected frames from the depth images when occlusions happened to obtain the ground-truth, which was down-sampled to the be synchronized with the audio measurements on a frame basis.

Note that, the TDOA audio cues cannot distinguish a signal from front or back. For instance, the signal from 45° and 135° yields the same TDOA features. To address this ambiguity, an audio range assumption of $[-90^\circ, 90^\circ]$ was imposed. In other words, we assumed the signal comes in

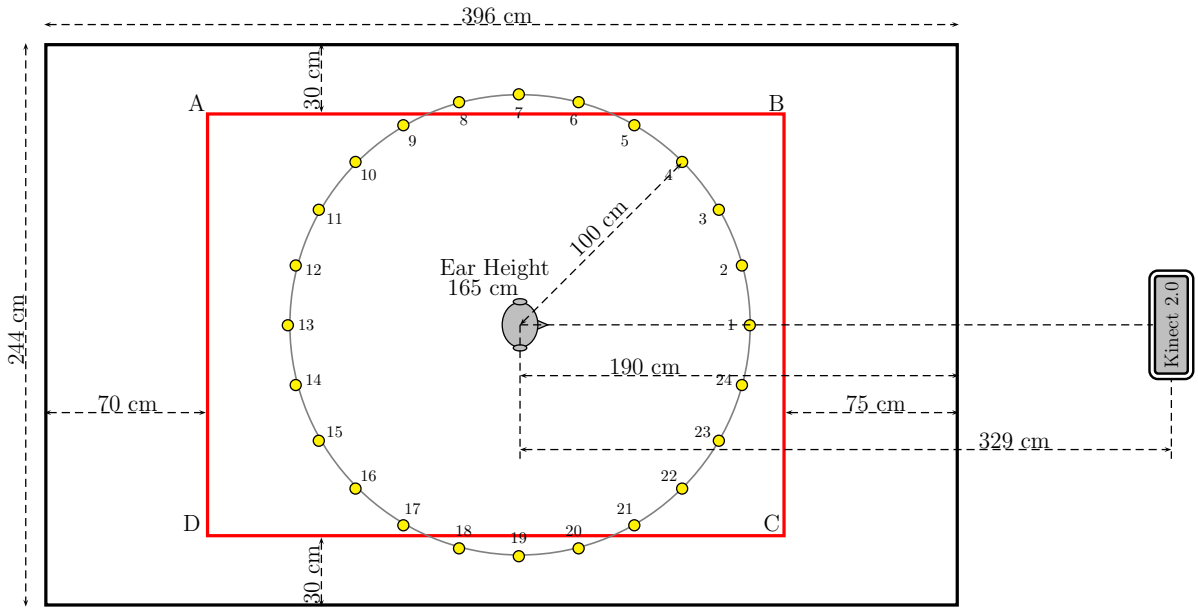

Figure 5. Setup for data recordings. The 24 highlighted dots in a circle labels the positions used to model the relationship between different input angles and the exhibited audio features.

front of the dummy head. The tracking results for Sequence 1 are shown in Figure. 7.

We then quantitatively evaluated the proposed method for single speakers using only audio cues. Sequence 1 has 1282 audio frames in total, and the proposed method results show that 1273 frames have one speaker, and 9 frames have no speaker, caused by periods of silence. The standard deviation from the ground truth angle is $7.8°$. Sequence 2 has 599 frames in total, and the proposed method detected one speaker in 573 frames and no speaker in 26 frames. In 109 frames out of 573, the angle estimation error was greater than $30°$. These frames occurred in the beginning and end of the recording session, when the target person was not silent, and an interfering speaker outside the recording field was talking. The standard deviation from the ground truth angle in the remaining 464 frames is $10.6°$.

4.3.2 Single person, audio and depth features

Secondly, we tested the proposed algorithm on the first two sequences, using both audio and depth features. The results for Sequence 1 are shown in Figure 8. It can be observed that the detected trajectory is of high quality as it almost overlaps with the ground-truth.

We then did some quantitative evaluations. In Sequence 1, in all of the 1282 frames, one person was successfully detected, with the deviation of $2.4°$. In Sequence 2, in all but 2 frames one person was detected, with the deviation of $3.8°$. This can be observed in the sub-rectangle in Figure 8, where we zoomed in a short segment of tracking results. The converged results are very close to the ground-truth, which proves the robustness of our proposed audio-depth fusion scheme. Compared with the results using audio-only features, the combination of audio and depth greatly reduced the error.

4.3.3 Two people, audio and depth features

Finally, we tested our algorithm on the two people scenario. Using audio-only features, the proposed method did not converge since the audio measurements are too noisy. Using depth-only features, the outliers were removed, but occlusions caused tracking loss. Using both audio and depth features, we successfully tracked both speakers, as shown for Sequence 3 in Figure 9. However, note that the identity of the speakers got swapped occasionally. This problem can be solved by applying a simple filter in space-time, e.g. by calculating the distance between detected person in two consecutive frames. Our depth-audio results on Sequence 2 were also consistent with the walking trajectory described earlier, demonstrating success with the fusion of depth and audio cues.

5. Conclusions

We presented a method for multimodal tracking using audio and depth features. TDOA features are extracted from binaural recording (Cortex MK2); 3D positions from a depth sensor (Kinect2) are mapped into 1D azimuth relative to Cortex MK2 as the depth features. The measurements from both modalities are fused in a particle filtering framework that enables birth and death of multiple tracks

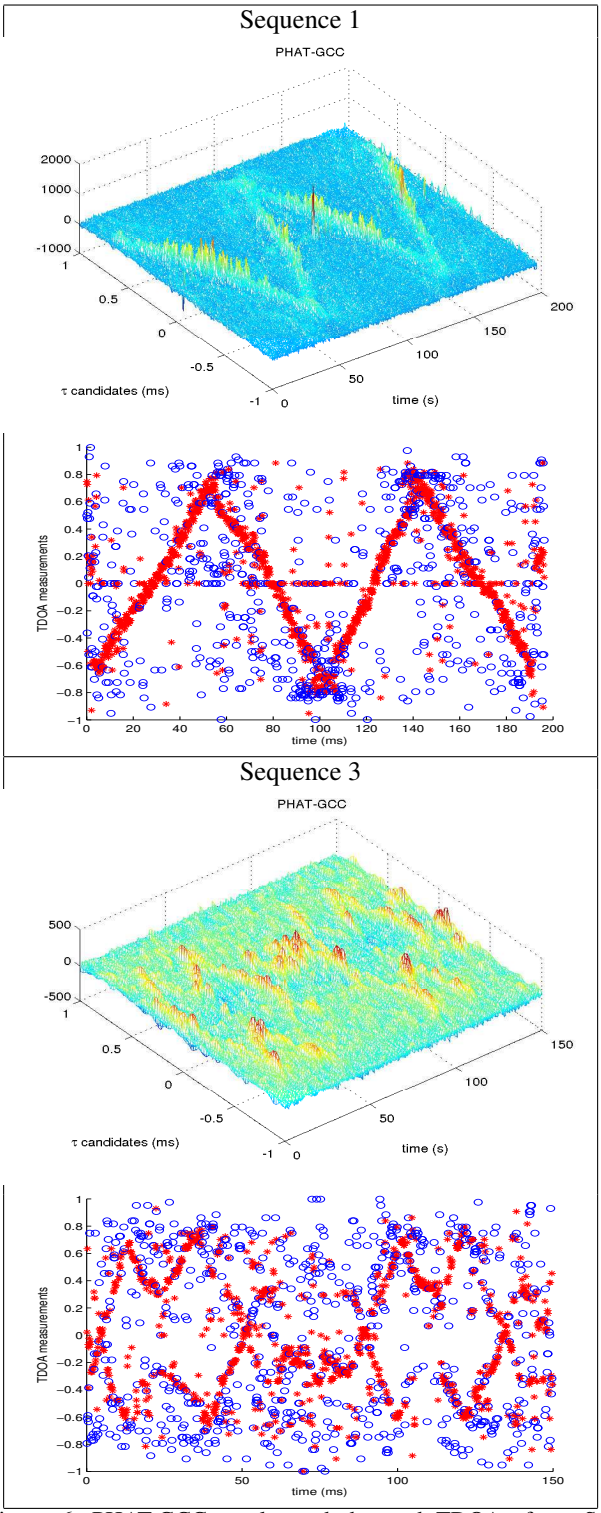

Figure 6. PHAT-GCC results and detected TDOAs from Sequence 1 and Sequence 3. The red-starred points denote the first peak-related TDOA while the blue circles represent the second one. The peaks in Sequence 1 are very smooth, which clearly exhibits the speaker's trajectory. However, despite some peaks being related to the real positions in Sequence 3, many more false alarms are obtained.

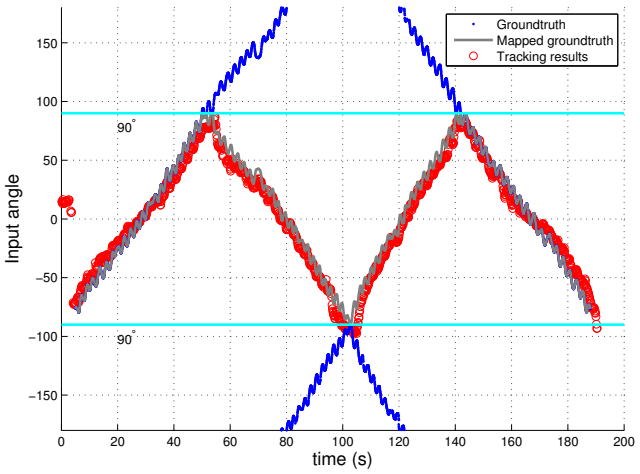

Figure 7. Azimuth angles relative to the dummy head estimated by the proposed method for Sequence 1, using only the audio features. Since the audio features have the front and back confusion, we imposed the input angle range of $[-90, 90]$. The blue dots represent the ground-truth input angle. We symmetrically mapped the angles at the back of the dummy head, i.e. $[-180, -90)$ and $(90, 180]$, to the front. The mapped ground-truth is the gray curve. The tracking results are represented via the red circles. Comparison between the tracked results and the mapped ground-truth demonstrates the practicability of the proposed method, and the adequacy of the previously-set parameters.

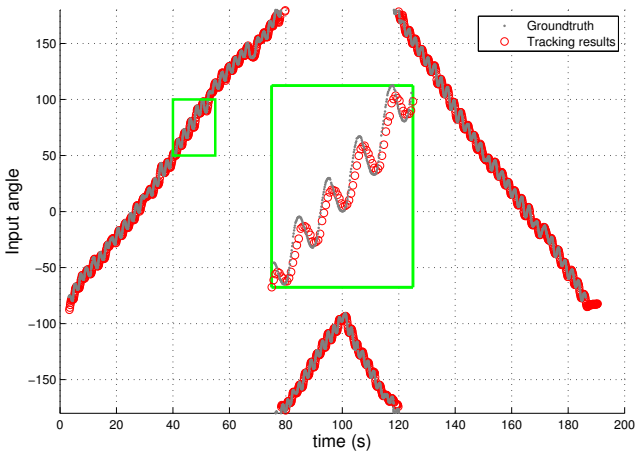

Figure 8. Application of the proposed method to Sequence 1 using both the audio and depth features. The gray dots represent the ground-truth input angle. The tracking results are represented via the red circles. We noticed the tracking results almost overlapped with the ground-truth. We have zoomed in a short segment highlighted in the rectangle.

using Random Finite Sets (RFS). These two modalities are obviously very different and have very different levels of confidence. We showed how to take that into account and how they can complement each other. Our results show that this combination clearly outperforms individual modalities, particularly when there are multiple people talking simulta-

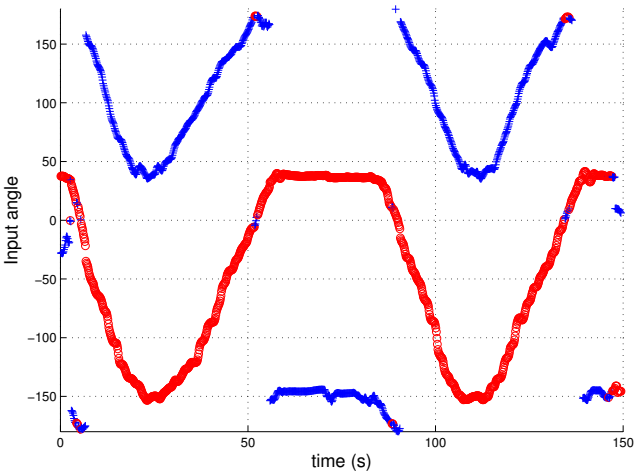

Figure 9. Application of the proposed method to Sequence 3 using both the audio and depth features. The blue crosses represent Actor B, and the red circles represent Actress A.

neously and when there is a significant amount of occlusion.

As future work, we plan to perform experiments on more datasets, aiming to highlight the method's potential to handle birth and death of targets. We also intend to compare our results against other tracking and fusion methods. The RFS tracking framework is a principled way to simultaneously track a varying number of targets, but its complexity grows as the number of targets increase. We suggest that depth-based tracking results, including the detected targets identities, should help us to design a modified version of RFS, with lower complexity w.r.t. the number of targets. We also plan to use the most confident modality to provide strong priors on the birth and death of tracking targets.

Acknowledgements

We acknowledge Luca Remaggi and Phil Coleman, who provided inpulse response measurements used in our experiments. Thanks to Mark Barnard for the help with the acquisition of our dataset. Data underlying the findings are fully available without restriction, available from [4].

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.

References

[1] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, Feburary 2002.

[2] Z. Chen. Bayesian filtering: From Kalman filters to particle filters, and beyond. Technical report, McMaster University, 2003.

[3] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–575, May 2003.

[4] T. de Campos, Q. Liu, and M. Barnard. S3A speaker tracking with Kinect2. Dataset, DOI 10.15126/surreydata.00807708, February 2015. Available from cvssp.org/data/s3a.

[5] A. Doucet, N. de Freitas, and N. Gordon. An introduction to sequential monte carlo methods. In *Sequential Monte Carlo Methods in Practice*, Statistics for Engineering and Information Science, pages 3–14. Springer New York, 2001.

[6] M. Fallon and S. Godsill. Acoustic source localization and tracking of a time-varying number of speakers. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1409–1415, May 2012.

[7] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool. Random forests for real time 3d face analysis. *Int. J. Comput. Vision*, 101(3):437–458, February 2013.

[8] 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*, 30(2):267–282, Feb 2008.

[9] D. Gatica-Perez, G. Lathoud, J. Odobez, and I. McCowan. Audiovisual probabilistic tracking of multiple speakers in meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):601–616, Feb 2007.

[10] V. Kilic, M. Barnard, W. Wang, and J. Kittler. Audio assisted robust visual tracking with adaptive particle filtering. *IEEE Transactions on Multimedia*, 17(2):186–200, Feb 2015.

[11] K. Kim and L. S. Davis. Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering. In *Proceedings of the 9th European Conference on Computer Vision*, pages 98–109, 2006.

[12] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(4):320–327, Aug 1976.

[13] S. J. Krotosky and M. M. Trivedi. Mutual information based registration of multimodal stereo videos for person tracking. *Computer Vision and Image Understanding*, 106(23):270–287, 2007.

[14] E. A. Lehmann and R. C. Williamson. Particle filter design using importance sampling for acoustic source localisation and tracking in reverberant environments. *EURASIP Journal on Advances in Signal Processing*, 2006(1):1–9, 2006.

[15] W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley. Tracking an unknown time-varying number of speakers using TDOA measurements: a random finite set approach. *IEEE Transactions on Signal Processing*, 54(9):3291–3304, 2006.

[16] M. P. Michalowski and R. Simmons. Multimodal person tracking and attention classification. In *Proceedings of the 1st Annual Conference on Human-Robot Interaction*, pages 347–348. ACM, March 2006.

[17] Microsoft. Meet Kinect for Windows. Online, retrieved in October 2015. dev.windows.com/en-us/kinect/.

[18] A. Mogelmose, C. Bahnsen, T. Moeslund, A. Clapes, and S. Escalera. Tri-modal person re-identification with RGB, depth and thermal features. In *Proceedings of CVPR Workshops*, pages 301–307, June 2013.

[19] 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*. BMVA Press, 2014.

[20] 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*, 48(2):1656–1668, 2012.

[21] J. M. Saragih, S. Lucey, and J. F. Cohn. Face alignment through subspace constrained mean-shifts. In *12th International Conference on Computer Vision*, pages 1034–1041. IEEE, 2009.

[22] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proc of the IEEE Conf on Computer Vision and Pattern Recognition*, 2011.

[23] L. Spinello, R. Triebel, and R. Siegwart. Multimodal people detection and tracking in crowded scenes. In *Proceedings of the 23rd National Conference on Artificial Intelligence*, pages 1409–1414, 2008.

[24] J. Vermaak and A. Blake. Nonlinear filtering for speaker tracking in noisy and reverberant environments. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 3021–3024, 2001.

[25] J. Vermaak, M. Gangnet, A. Blake, and P. Perez. Sequential monte carlo fusion of sound and vision for speaker tracking. In *IEEE International Conference on Computer Vision*, volume 1, pages 741–746. IEEE, 2001.

[26] 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*, volume 8200 of *LNCS*, pages 149–187. Springer, 2013.

[27] M. Ye, Y. Zhang, R. Yang, and D. Manocha. 3D reconstruction in the presence of glasses by acoustic and stereo fusion. In *Proc of the IEEE Conf on Computer Vision and Pattern Recognition*, 2015.