AUDIO CONSTRAINED PARTICLE FILTER BASED VISUAL TRACKING

Volkan Kılıç, Mark Barnard, Wenwu Wang, and Josef Kittler

Centre for Vision, Speech and Signal Processing, University of Surrey, UK
Emails:{v.kilic, mark.barnard, w.wang, j.kittler}@surrey.ac.uk

ABSTRACT

We present a robust and efficient audio-visual (AV) approach to speaker tracking in a room environment. A challenging problem with visual tracking is to deal with occlusions (caused by the limited field of view of cameras or by other speakers). Another challenge is associated with the particle filtering (PF) algorithm, commonly used for visual tracking, which requires a large number of particles to ensure the distribution is well modelled. In this paper, we propose a new method of fusing audio into the PF based visual tracking. We use the direction of arrival angles (DOAs) of the audio sources to reshape the typical Gaussian noise distribution of particles in the propagation step and to weight the observation model in the measurement step. Experiments on AV16.3 datasets show the advantage of our proposed method over the baseline PF method for tracking occluded speakers with a significantly reduced number of particles.

Index Terms— Particle filter, visual tracking, DOAs

1. INTRODUCTION

Tracking of multiple moving speakers in indoor environments has received much interest in the fields of computer vision and signal processing in the past decades. Speaker tracking may be achieved in a single modality domain through video or audio. Video tracking is generally accurate, but it suffers from a limited field of view, occlusions, and changes in appearance and illumination. On the other hand, audio tracking is not restricted by these limitations, but it is prone to the errors caused by acoustic noise, room reverberations and the intermittency between utterance and silence. As shown in some already published works (see Section 5), fusing both audio and visual modalities can provide more robust tracking in comparison to the use of only a single modality, as is the focus here.

We propose a new algorithm for joint audio-visual (AV) tracking based on particle filtering (PF). In this algorithm, the direction of arrival angles (DOAs) of the sources, estimated from microphone recordings are used to reshape the distribution of the particles in the propagation step and to weight the observation model in the measurement step of the visual tracker. We show in our experiments that incorporating audio information in this way, not only addresses the occlusion problem, a challenging scenario in visual tracking, but also significantly reduces the number of particles required in visual tracking to robustly model the distribution of the particles and the estimation of the state vector.

The following section introduces PF based visual tracking. Our proposed algorithm is given in Section 3, and experimental results are presented in Section 4. Related work is discussed in Section 5, followed by the conclusions.

2. PARTICLE FILTER BASED VISUAL TRACKING

The sampling importance resampling (SIR) PF is used to track the face of the speakers in visual tracking and it has five steps. First, the particles are initialized by $\mathbf{x}_0^{(n)} \sim p(\mathbf{x}_0)$, $w_0^{(n)} = \frac{1}{N}$ for $n = 1, ..., N$. Here N is denoted as the number of particles and $w_0^{(n)}$ is the initial weights of the particles. The PF utilizes a state vector $\mathbf{x} = \begin{bmatrix} x_1 & \dot{x}_1 & x_2 & \dot{x}_2 & s \end{bmatrix}^T$, where x_1 and x_2 are the horizontal and vertical position of the rectangle centred around the face, $\dot{x}_1$ is the horizontal velocity, $\dot{x}_2$ is the vertical velocity and s is the scale of the rectangle centred around (x_1, x_2). In the second step, the particles are propagated by a dynamic model,

$$\mathbf{x}_k^{(n)} = \mathbf{F}\mathbf{x}_{k-1}^{(n)} + \mathbf{q}_k^{(n)} \tag{1}$$

where $\mathbf{x}_k^{(n)}$ is the state of n^{th} particle at time frame $k = 1, ..., K$ and $\mathbf{q}_k^{(n)}$ is the zero-mean Gaussian noise with covariance $\mathbf{Q}$, $\mathbf{q}_k^{(n)} \sim \mathcal{N}(0, \mathbf{Q})$ for each particle. $\mathbf{F}$ is the linear motion model,

$$\mathbf{F} = \begin{bmatrix} 1 & T & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & T & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

T is the period between two adjacent frames. The particles are weighted in the third step by the observation model,

$$w_k^{(n)} = p(\mathbf{y}_k^{(n)} | \mathbf{x}_k^{(n)}) = e^{-\lambda (D^{(n)})^2} \tag{3}$$

where $\mathbf{y}_k^{(n)}$ is the observation. The weights are then normalized to ensure that $\sum_{n=1}^{N} w_k^{(n)} = 1$. The observation $\mathbf{y}_k^{(n)}$ is

obtained for each state estimate $\mathbf{x}_k^{(n)}$ by the design parameter λ and $D^{(n)}$ which is the Bhattacharyya distance,

$$D^{(n)} = \sqrt{1 - \sum_{u=1}^{U} \sqrt{r(u)q^{(n)}(u)}} \tag{4}$$

where, U is the number of bins used by the histogram, $r(u)$ is the Hue histogram of the reference image, which is determined by the user in the initialisation step, and $q^{(n)}(u)$ is the Hue histogram extracted from the rectangle centred on the position of the n^{th} particle. In the literature the RGB or HSV colour model is commonly used. In our case, HSV is used as it is observed to be more robust to varying illumination. In the fourth step, the position of the speaker is estimated by:

$$\tilde{\mathbf{x}}_k = \sum_{n=1}^{N} w_k^{(n)} \mathbf{x}_k^{(n)} \tag{5}$$

Lastly the resampling procedure is performed to remove the particles with very small weights and duplicating the particles with large weights and generate new particle set from $\left\{\mathbf{x}_k^{(n)}, w_k^{(n)}\right\}_{n=1}^{N}$. Then we return to the second step and continue recursively.

3. PROPOSED ROBUST AND EFFICIENT AUDIO-VISUAL TRACKER

The visual tracker presented in Section 2 is likely to fail, for example, when the speaker moves out of the view of the camera. In this scenario, the particles $\mathbf{x}_k^{(n)}$ will converge to the region on the image frame that is most similar to the face in color measured by $D^{(n)}$. Such a region can be anywhere in the image, e.g. the background of the visual scene. To prevent this, we introduce the DOAs information estimated from audio measurements (Section 3.2) to constrain the propagation of particles and the weights in the observation model (Section 3.1) of the visual tracker.

3.1. Audio Constrained Visual Tracker

We use the visual PF approach described above to approximately localise the position of the face and the particles in the image frame, and also use the audio tracker discussed in Section 3.2 to locate the approximate position of the speaker by estimating the DOAs from the microphone measurements.

We then draw the DOAs line from the centre of the microphone array to a point (a_k, b_k) in the image frame k. This point is projected from the three dimensional point (A, B_k, C) where A is the distance from the centre of the microphone array to the wall in metres (which is 1.75 meters in our experiments), C is estimated as the height of the speaker, typically chosen as 1.80 metres in our experiment, and B_k is calculated as

$$B_k = \tan(\theta_k \times \frac{\pi}{180}) \cdot A \tag{6}$$

where θ_k is the DOA (azimuth) angle (in degrees) of the speaker estimated from the audio frame that is synchronised with image frame k. The Euclidean distances $\mathbf{d}_k = \begin{bmatrix} d_k^{(1)} & \dots & d_k^{(N)} \end{bmatrix}^T$ of the particles to this line are then calculated, and used to derive the movement distances $\hat{\mathbf{d}}_k$ which guide how much distance the particles should be moved towards the DOA line,

$$\hat{\mathbf{d}}_k = \frac{\mathbf{d}_k \odot \mathbf{d}_k}{\|\mathbf{d}_k\|_1} \tag{7}$$

where $\hat{\mathbf{d}}_k = \begin{bmatrix} \hat{d}_k^{(1)} & \dots & \hat{d}_k^{(N)} \end{bmatrix}^T$ and $\odot$ is the dot (element-wise) product and $\|.\|_1$ is the ℓ_1 norm. This information is then used to reshape the particle distribution during the propagation step in (8).

The reliability and accuracy of DOAs, however, can be affected by the noise within the audio measurements. To counter these effects, we adjust the contribution of audio to the calculation of particle propagation and importance weighting by using a weighting parameter ξ_k. To this end, we choose the image patch $q(u)$ centred on the estimated position and calculate ξ_k as the distance between $q(u)$ and the reference image patch $r(u)$, by substituting $q(u)$ for $q^{(n)}(u)$ in (4). The dynamic model given in (1) is then revised to

$$\hat{\mathbf{x}}_k^{(n)} = \mathbf{x}_k^{(n)} \oplus \hat{d}_k^{(n)} tan(\theta_k) \xi_k \tag{8}$$

where $\oplus$ is the element-wise addition. The importance weights are also adapted using $\hat{d}_k^{(n)}$ and ξ_k as follows:

$$\hat{w}_k^{(n)} = (e^{-\lambda (D^{(n)})^2}) \frac{\|\mathbf{d}_k\|_1}{d_k^{(n)}} \xi_k \tag{9}$$

It is then normalized to ensure that $\sum_{n=1}^{N} \hat{w}_k^{(n)} = 1$. Position estimation follows the weighting step and it is calculated using (5). Then the resampling step is performed to generate the new particles $\mathbf{x}_k^{(n)}$ from the set $\left\{\hat{\mathbf{x}}_k^{(n)}, \hat{w}_k^{(n)}\right\}_{n=1}^{N}$. The pseudo code of proposed algorithm is given in Table 1.

With the proposed modifications in (8) and (9), the tracking algorithm preserves the position of the face even if the visual tracker is lost, due to the use of the DOAs from audio. Moreover, the weighting parameter ξ_k mitigates the potential influence of acoustic noise on the tracking system.

3.2. Audio Detection and Localization

We discuss in this section the estimation of the DOAs and the enhancement of the estimates using a smoothing process based on the Auto-Regressive (AR) model.

To estimate DOAs, we use a two-step method proposed in [1]. The first step consists of a sector based combined detection and localization. In this step the space around a circular microphone array is divided into a number of sectors. At each time frame for each sector an activeness measure is taken using the SAM-SPARSE-MEAN approach [2]. This measure of activeness is then compared to a threshold in order to determine whether there is an active source in that sector. In

Table 1. Proposed Algorithm

Initialize: N, Q, U, T, $\mathbf{F}$, λ, $r(u)$, $\mathbf{x}_0^{(n)}$, $w_0^{(n)}$, k

while $k < K$ **do**

- Propagate particles: $\mathbf{x}_k^{(n)} = \mathbf{F}\mathbf{x}_{k-1}^{(n)} + \mathbf{q}_k^{(n)}$
- Calculate $D^{(n)}$ using equation (4), for $n = 1...N$
- Weighting: $w_k^{(n)} = e^{-\lambda(D^{(n)})^2}$, for $n = 1...N$
- Estimate target position using equation (5)
- Calculate ξ_k using equation (4)
- Get corresponding DOA angle θ_k
- Calculate distances $\mathbf{d}_k = \left[d_k^{(1)} \quad ... \quad d_k^{(N)}\right]^T$
- Find movement distances: $\hat{\mathbf{d}}_k = \frac{\mathbf{d}_k \odot \mathbf{d}_k}{\|\mathbf{d}_k\|_1}$
- Re-propagate particles: $\hat{\mathbf{x}}_k^{(n)} = \mathbf{x}_k^{(n)} \oplus \hat{d}_k^{(n)} tan(\theta_k)\xi_k$
- Re-weighting: $\hat{w}_k^{(n)} = (e^{-\lambda(D^{(n)})^2})\frac{\|\mathbf{d}_k\|_1}{d_k^{(n)}}\xi_k$
- Re-estimate target position using equation (5)
- Resampling: Generate $\mathbf{x}_k^{(n)}$ from the set $\left\{\hat{\mathbf{x}}_k^{(n)}, \hat{w}_k^{(n)}\right\}_{n=1}^N$
- $k = k + 1$

end

the second step a point based search is conducted in each of the sectors labelled as having at least one active source. The localization uses a parametric approach [1], the location parameters are optimized with respect to a cost function such as SRP-PHAT [3].

Then, we perform a third order AR model to reduce the noise in the estimate of the azimuth.

$$\theta_k = \sum_{i=1}^{3} \varphi_i \theta_{k-i} + \varepsilon_k \tag{10}$$

where φ_1,...,φ_i are the parameters of the model and ε_k is white noise.

4. EXPERIMENTS

4.1. Setup

The proposed algorithm was tested using the $AV16.3$ corpus developed by IDIAP Research Institute [4]. The corpus consists of subjects moving and speaking at the same time whilst being recorded by three calibrated video cameras and two circular eight-element microphone arrays. The audio was recorded at 16 kHz and video was recorded at 25 Hz. They were synchronized before being used in our system. Each video frame is a colour image of 288x360 pixels. The corpus is annotated for speaker position which allows us to measure the accuracy of each tracker and compare the performance of the algorithms. To do this, the tracking error between the estimation and the ground truth is calculated as the Euclidean distance in pixels. The error for frame k is the average of the errors from frame 1 to k. In the sequences, the speakers wear a ball for annotation but in our application this ball is never used. We used two single speaker sequences (sequence 11, cameras #1 and #3) which have 817 and 769 frames and one multiple speaker sequence (sequence 24, camera #1) which has 1201 frames to test our proposed algorithm. We are only able to show annotated audio with ground truth part of the sequences in the result.

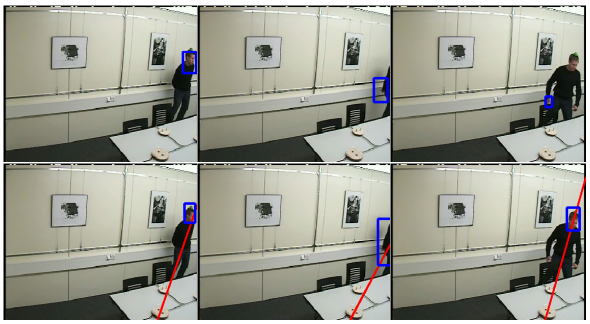


Fig. 1. Sequence 11 (camera #1): Speaker disappears for a while and re-enters to the scene.

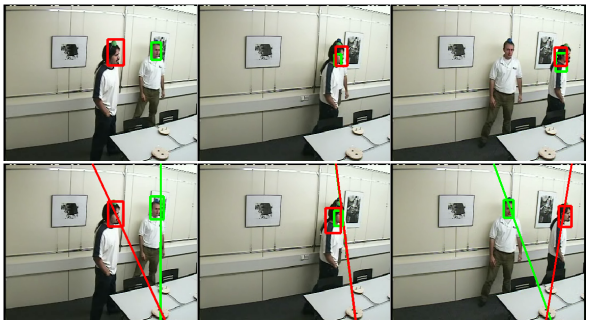


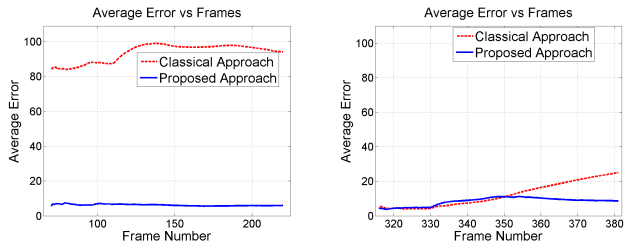
Fig. 2. Sequence 24: Multiple speakers with occlusion.

In all simulations the number of particles, N, is selected to be 10. Covariance matrix Q is a diagonal matrix with $\sigma^2 = 50$. This is used as the standard deviation for both the position and velocity. T is the period between frames, which equals 0.04 seconds and λ in (3) is chosen as 150. The number of bins used for the Hue histogram is 8.

4.2. Results and Analysis

We first demonstrate our approach on two challenging scenarios: a single speaker (Sequence 11, camera #1) moving in and out of camera view, and multiple speakers (Sequence 24, camera #1) occluding each other.

The tracking results for the single speaker sequence are given in Figure 1. In the first row, the classical PF approach results are given and as seen from the frames, when the speaker comes back after disappearing for a while, the tracker fails to track the face, but locks onto the hand of the speaker. In contrast, our proposed algorithm resumes tracking after the speaker reappears as shown in the second row. The plot in Figure 3 (a) shows the tracking error for this sequence. Since the ground truth data does not start from the first frame, the average error is calculated starting from frame 71 where the



(a) (b)

Fig. 3. In (a) and (b) performance of algorithms are given for sequence 11 (camera #1) and sequence 24, respectively.

classical approach has already lost tracking, but our proposed algorithm continued tracking.

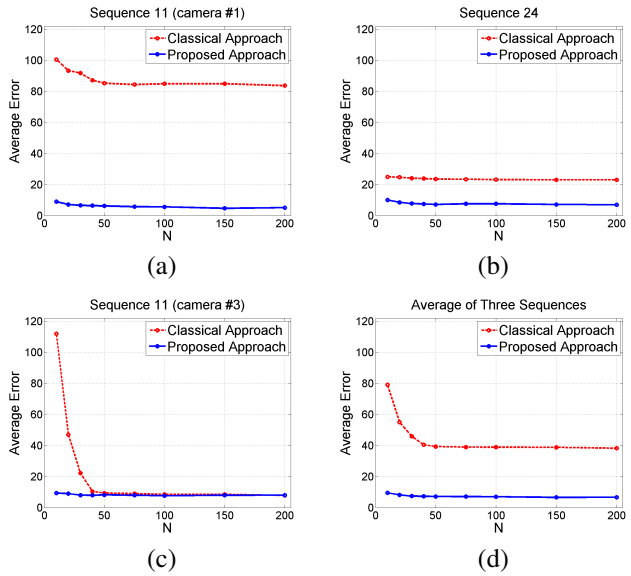
The results for sequence 24 are given in Figure 2, where one speaker is occluded by the other. Our algorithm (in the second row) resumes tracking after the occlusion finishes. It can be seen that, with our proposed algorithm, the speaker's identity is preserved after the occlusion. In Figure 3 (b), the average error for this sequence is shown and after the 350^{th} frame the classical approach fails, but our proposed algorithm continues tracking with a small errors.

We also performed experiments by changing the number of particles: 10, 20, 30, 40, 50, 75, 100, 150 and 200 for all the three sequences. The results for sequence 11 (camera #1), sequence 24, and sequence 11 (camera #3) are shown in the subplots (a), (b) and (c) of Figure 4 respectively. The average error of all three sequences is shown in Figure 4 (d).

It can be seen that even for high numbers of particles the classical approach fails for sequence 11 (camera #1) and sequence 24. Sequence 11 (camera #3) features a person making a variety of rapid movements, despite the fact that no occlusion is involved. From Figure 4 (c), we can observe that given a large number of particles, the visual tracker perform almost equally well as our proposed method. However, when the number of particles is reduced significantly, e.g. when $N = 10$, using the classical PF method, the tracking errors increase dramatically, while our proposed algorithm continues to show excellent performance.

5. RELATED WORKS

The problem of tracking and localization of speakers in enclosed spaces using AV information has recently received much attention. Many approaches have been proposed by researchers and one of the most popular is the PF. PF became widely used in tracking after being proposed by Isard and Blake [5]. Much success has been achieved in visual tracking [5], [6], [7] and audio tracking [8], [9]. It has been shown in many studies [10], [11], [12], [13] using AV data in tracking gives more reliable results than using each individually. However, it is crucial to effectively fuse AV data. To solve this problem data association and fusion algorithms have been developed. They have been implemented with a PF in [7] as



(a) (b) (c) (d)

Fig. 4. Performance comparison between the proposed and baseline methods versus the number of particles N.

a good example of data association and in [10], [11], [13] as examples of fusion algorithms. In this study, our proposed tracker is mainly based on visual data, but audio data is also used to enhance the visual tracker in order to improve its robustness in challenging cases such as occlusion but also reduces considerably the number of particles used in tracking. To the best of our knowledge, the algorithm presented in this paper is the first to constrain the number of particles with audio information in visual speaker tracking within a PF framework.

6. CONCLUSION

In this study, we presented a new audio-visual tracking algorithm in which audio information is used to modify particle propagation and the weights assigned to the particles. Our proposed algorithm was tested on both single and multiple speaker sequences. It showed significantly improved tracking performance over the classical approach for the scenarios where the speaker is either occluded by other speakers or out of the range of camera view. We demonstrate that by using audio information we can significantly reduce the number of particles, whilst maintaining good tracking performance. This approach has the potential for handling weight degeneracy and particle impoverishment problems due to the significant reduction in the number of particles being used in tracking, which we will study in the future.

7. ACKNOWLEDGEMENT

This research was supported by the Engineering and Physical Sciences Research Council of the UK (grant no. EP/H050000/1).

8. REFERENCES

[1] G. Lathoud, J. Bourgeois, and J. Freudenberger, "Sector-based detection for hands-free speech enhancement in cars," *EURASIP Journal on Applied Signal Processing, Special Issue on Advances in Multimicrophone Speech Processing*, 2006.

[2] G. Lathoud and M. Magimai-Doss, "A sector-based, frequency-domain approach to detection and localization of multiple speakers," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05)*, march 2005, vol. 3, pp. iii/265 – iii/268 Vol. 3.

[3] J. DiBiase, "A high-accuracy, low-latency technique for talker localisation in reverberant environments," in *Ph.D. dissertation, Brown University, Providence, RI, USA*, 2000.

[4] G. Lathoud, J. M. Odobez, and D. Gatica-perez, "Av16.3: an audio-visual corpus for speaker localization and tracking," in *S. Bengio and H. Bourlard Eds Proceedings of the 2004 MLMI Workshop*. 2005, Springer Verlag.

[5] M. Isard and A. Blake, "Condensation - conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, pp. 5–28, 1998.

[6] M.S.A. Ramli, H. Zamzuri, and M.S.Z. Abidin, "Tracking human movement in office environment using video processing," in *2011 4th International Conference on Modeling, Simulation and Applied Optimization (ICMSAO)*, 19-21 April 2011, pp. 1–6.

[7] C.R. del Blanco, F. Jaureguizar, and N. Garcia, "Visual tracking of multiple interacting objects through rao-blackwellized data association particle filtering," in *17th IEEE International Conference on Image Processing (ICIP)*, 26-29 Sept. 2010, pp. 821–824.

[8] D.B. Ward, E.A. Lehmann, and R.C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 836–836, Nov. 2003.

[9] X. Zhong and A. B. Premkumar, "Particle filtering approaches for multiple acoustic source detection and 2-d direction of arrival estimation using a single acoustic vector sensor," *IEEE Transactions on Signal Processing*, vol. 60, pp. 4719–4733, Sept. 2012.

[10] M. Heuer, A. Al-Hamadi, B. Michaelis, and A. Wendemuth, "Multi-modal fusion with particle filter for speaker localization and tracking," in *International Conference on Multimedia Technology (ICMT)*, 26-28 July 2011, pp. 6450–6453.

[11] S.T. Shivappa, B.D. Rao, and M.M. Trivedi, "Audio-visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 882–894, Oct. 2010.

[12] F. Talantzis, A. Pnevmatikakis, and A.G. Constantinides, "Audiovisual active speaker tracking in cluttered indoors environments," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, pp. 799–807, June 2008.

[13] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 601–616, Feb. 2007.