IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING (ICASSP), TAIPEI, TW, 19-24 APRIL 2009

AUDIO-ASSISTED TRAJECTORY ESTIMATION IN NON-OVERLAPPING MULTI-CAMERA NETWORKS

Murtaza Taj and Andrea Cavallaro

Multimedia and Vision Group, Queen Mary, University of London,
Mile End Road, London E1 4NS, UK

ABSTRACT

We present an algorithm to improve trajectory estimation in networks of non-overlapping cameras using audio measurements. The algorithm fuses audiovisual cues in each camera's field of view and recovers trajectories in unobserved regions using microphones only. Audio source localization is performed using Stereo Audio and Cycloptic Vision (STAC) sensor by estimating the time difference of arrival (TDOA) between microphone pair and then by computing the cross correlation. Audio estimates are then smoothed using Kalman filtering. The audio-visual fusion is performed using a dynamic weighting strategy. We show that using a multi-modal sensor with combined visual (narrow) and audio (wider) field of view can enable extended target tracking in non-overlapping camera settings. In particular, the weighting scheme improves performance in the overlapping regions. The algorithm is evaluated in several multi-sensor configurations using synthetic data and compared with state of the art algorithm.

Index Terms— TDOA, fusion, tracking.

1. INTRODUCTION

The use of multiple cameras for trajectory estimation in wide areas is of great interest for many applications, such as surveillance and sports analysis. In many scenarios the environment to be monitored cannot be covered completely by a single sensor, hence multiple cameras are used to observe the behaviour of the targets [1]. However, in many cases even multiple cameras cannot cover the whole environment, thus reducing the number of target observations estimated in the scene. The missing information can be estimated either by a prediction based on the targets state in the cameras' fields of view and their motion dynamics [1, 2] or by using sensors with a wider field of observation, such as the sound field of microphones, as discussed in this paper. Audio sensors overcome some of the limitations of visual sensors, such as bad lighting and visual occlusion due to vegetation or dust. This makes a network of heterogeneous sensors consisting of both cameras and microphones a desirable solution for wider-area coverage. Each sensor in such a network can be a simple Stereo Audio and Cycloptic Vision (STAC) sensor [3] consisting of a camera mounted between a pair of microphones (Fig. 2).

This paper is organized as follows. An overview of the related work is given in Section 2. The proposed trajectory estimation using audiovisual fusion is discussed in Section 3. In Section 4 we show some experimental results and there evaluation. Finally, in Section 5 we draw conclusions.

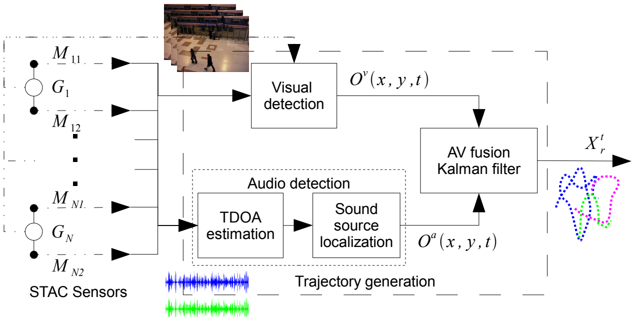

Fig. 1. Flowchart of the proposed audiovisual tracking algorithm (Key. STAC: Stereo Audio Cycloptic Vision; TDOA: time difference of arrival; M_{i1} and M_{i2}: pair of microphones; G_i: camera; AV: audiovisual).

2. RELATED WORK

Video-based tracking faces several challenges due to factors such as local and global illumination change and visual occlusions. To overcome these issues sound source tracking has been studied [4]. The detection of a sound source is performed using either time difference of arrival (TDOA) or beamforming. The former performs localization by estimating delays between pair of microphones whereas, later maximizes steered response of a beamformer for localization. Audio modality also suffers from environmental factors such as background noise, reverberation and reflections. It has been shown that information from heterogeneous sensors such as cameras and microphones can be fused in a unified manner both at sensor level [5] or at feature level [6]. This fusion of modalities can compensate for the failure of each other and is used in target localization and tracking in indoor and outdoor scenarios using a network of audiovisual sensors.

In indoor scenarios, active speaker localization and tracking in meeting rooms can be done by using audiovisual fusion where audio localization is done using beamforming and visual target is tracked using Kalman filter [7] or by using TDOA with Particle filter [8]. Tracking using audio modality is strongly effected by reverberation. In [3] the audiovisual tracking algorithm [8] is improved by using Weighted Probabilistic Data Association filter (WPDA) which takes into account the weighted probability of detections and enhance the performance in reverberation scenarios. Multiple moving target tracking in surveillance scenarios can also benefit from fusion of modalities by applying fusion using iterative decoding algorithm based on the theory of turbo codes and factor graphs. The

The authors acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC), under grant EP/D033772/1

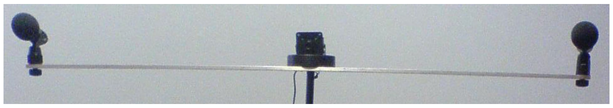

Fig. 2. Stereo Audio and Cycloptic Vision (STAC) sensor.

turbo codes uses two simple codes working in parallel to achieve higher performance. In case of multimodal tracking, multiple cues serve as two simple codes and turbo code is used to improve performance. In outdoor scenarios, the use of audio can help in case of visual occlusions due to dust or vegetation. In [9] TDOA is used to have an estimate of the target location which is then refined using visual cues under a particle filtering framework for tracking of tanks in the presence of dust and clutter. One of the challenges in using multiple modalities is synchronization. A joint audiovisual filter is used to address the synchronization issues where a sliding window of direction-of-arrival forms the audio observations whereas, adaptive appearance model is used for visual observations and synchronization is handled using a delay variable in the state.

Most of the multi-modal tracking algorithms uses audio in the camera's field of view only for compensating failures in video. In this work we exploit the audio modality to estimate trajectories in regions *outside* the camera's field of view (FOV). The missing visual observations are compensated by audio using multiple microphones having wider field of observation. The trajectories are generated using video only or joint audiovisual data over wide-area. In case of joint audiovisual data we also propose a dynamic weighting strategy based on the arrival angle and the target-sensor distance. The flow diagram of the framework is shown in Fig. 1.

3. TRAJECTORY ESTIMATION AND FUSION

Let a wide-area be monitored by a set $G = \{G_1, \dots, G_N\}$ of N cameras with non-overlapping fields of view (FOV). Let each camera be equipped with a microphone pair, with $M = \{M_1, \dots, M_N\}$ being the set of N microphone pairs, where $M_i = (M_{i1}, M_{i2})$. We assume that the microphones' sound field is wider than the corresponding cameras' field of view and that the sound field of multiple microphone pairs M_i overlap each other (Fig. 3). Let each target generate a sound which is received at the microphones after a certain attenuation and delay. Let $\mathbf{y}(t)$ be a sound wave generated by the source containing f_s/n_v samples, where f_s is the sampling frequency and n_v is the number of video frames per second. This signal reaches the Stereo Audio Cycloptic Vision (STAC) sensor (Fig. 2), consisting of a camera mounted between two microphones, at a certain arrival angle θ. Let the audio signals received at two microphones be defined as

$$\hat{\mathbf{y}}_1(t) = \Gamma_1 \mathbf{y}(t+n) + \mathcal{N}_1, \tag{1a}$$

$$\hat{\mathbf{y}}_2(t) = \Gamma_2 \mathbf{y}(t+n+\tau) + \mathcal{N}_2, \tag{1b}$$

where Γ_1 and Γ_2 are the attenuation factors; n is the delay, in samples, occurred for the signal to reach the first microphone M_{i1}; τ is the extra delay, in samples, for the signal to reach the second microphone M_{i2}; and $\mathcal{N}_1$ and $\mathcal{N}_2$ are the process noise added to the signal which is assumed to be zero mean Gaussian with unit variance. The attenuation Γ is calculated using the Beer-Lambert law [1] as

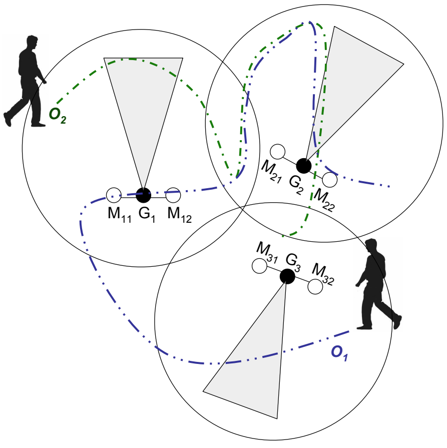

Fig. 3. Sample network of multi-modal sensors. (Key. O_i targets; M_{i1} and M_{i2}: microphones; G_i: camera; circles: sound field; triangles: field of view).

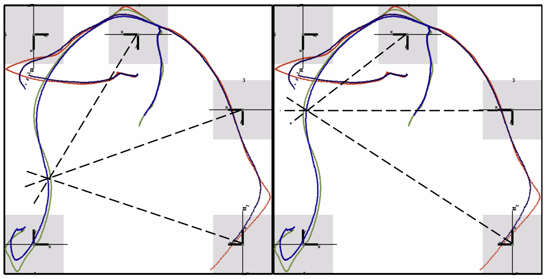

Fig. 4. Target localization using TDOA with multiple STAC sensors. Red and green lines: ground truth; blue and black line: estimated trajectories. Grey squares: overlapping regions; black dashed lines: audio source localization using arrival angles with 3 STAC sensors.

$$\Gamma_i = \Gamma_0 \exp(-\alpha L_{M_{ij}}). \tag{2}$$

where Γ_i is the attenuation for i^{th} microphone, Γ_0 is the initial sound intensity and $L_{M_{ij}}$ is the path length between the i^{th} microphone and j^{th} object.

In case of synthetic data the observation generated in camera's FOV gives video target location information. The audio signals received at each microphone couple (M_{i1}, M_{i2}) at each time step are used to compute the time difference of arrival (TDOA) τ of the audio signal for estimation of arrival angle θ for target localization in regions covered by the cameras' FOVs as well as in the uncovered areas (Fig. 4). The TDOA is estimated by computing the cross correlation of the two audio signals ($\hat{\mathbf{y}}_{i1}(t)$ and $\hat{\mathbf{y}}_{i2}(t)$) as

$$\hat{R}_{\hat{\mathbf{y}}_{i1}\hat{\mathbf{y}}_{i2}}(f) = \mathcal{F}(\hat{\mathbf{y}}_{i1})(f)^* \times \mathcal{F}(\hat{\mathbf{y}}_{i2})(f), \tag{3}$$

where $\mathcal{F}$ indicates the discrete Fourier transform and '$*$' indicates complex conjugation. Then the arrival angle, θ_i is estimated as

[1] http://elchem.kaist.ac.kr/vt/chem-ed/spec/beerslaw.htm Last accessed: 29 Sep, 2009

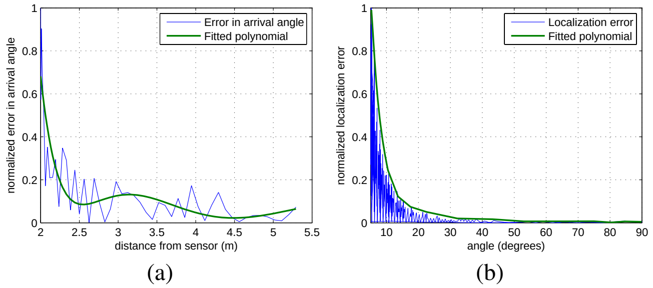

Fig. 5. Arrival angle and localization error analysis. Blue line: error; green line: fitted polynomial. (a) Example of increase in the error in the arrival angle when the target moves closer than 5m from the sensor. This error is due to the violation of the parallel line propagation assumption. (b) Example of increase in the localization error with the decreasing of the angle between the intersecting lines from two STAC sensors.

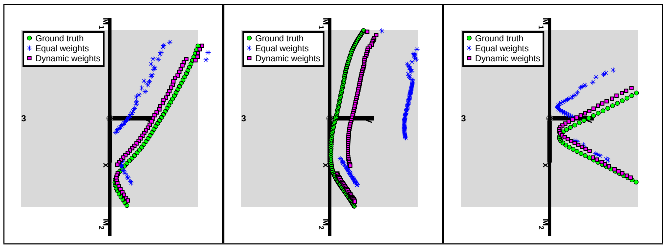

Fig. 6. Examples of audiovisual fusion. Grey square: field of view of a camera; white background: regions unobserved by a camera; green circles: ground truth; blue asterisk: audiovisual fusion with equal weights (i.e., $\gamma = 0$ in Eq. 4); Magenta squares: audiovisual fusion with dynamic weights computed using Eq. (4).

$\theta_i = \arccos(v_c\tau/L_{M_i})$, where v_c is the speed of sound in air and L_{M_i} is the distance between the two microphones. After the estimation of the arrival angle θ_i, we apply *audio-audio fusion* to estimate the targets position. A line is projected from the mid-point of the two microphones in the direction θ from each STAC sensor and the intersection of these lines from multiple STAC sensors gives the target position. However, this localization can be erroneous and the error increases as $\rho \longrightarrow 0$ (Fig. 5(b)) or as $\rho \longrightarrow 180$, where ρ is the angle between the two intersecting lines. The minimum localization error is achieved at $\rho = 90$. The estimation done by the pair of STAC sensors for $147^o < \rho < 33^o$ is ignored and the information from other STAC pairs is used. The audio performance also decreases as the target moves closer than 5m from the sensor as the assumption of parallel sound waves in TDOA estimation will no more be valid. In case no STAC sensor is able to provide the localization information, we apply trajectory estimation using the first order motion model as $\mathbf{x}(t+1) = \mathbf{x}(t) + U\nu(t) + \mathcal{N}(\mu, \Sigma)$ where $\nu(t) = (0, \nu_x, 0, \nu_y)$. The *audiovisual fusion* is then performed within Kalman filtering by taking a weighted sum of the two observations as $\gamma o_i^\nu + (1-\gamma)o_i^a$ where γ is computed as

$$\gamma = \begin{cases} 1 & \text{video only} \\ 0 & \text{audio only} \\ 0.5 + 0.25\left(\psi(d_e) + \psi(\rho)\right) & \text{otherwise} \end{cases}, \qquad (4)$$

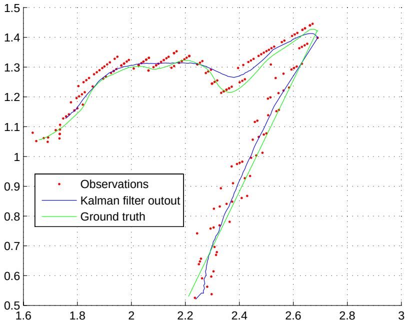

Fig. 7. Examples of audio only trajectory estimation using TDOA followed by correlation and fusion within Kalman filter. Red dots: estimated target position using audio; green line: ground truth; blue line: Kalman filter output.

where $\psi(d_e)$ is a 25^{th} order polynomial fitted over the normalized error in the estimation of the arrival angle θ with respect to the Euclidean distance d_e between the target and the microphone pairs (Fig. 5(a)) and $\psi(\rho)$ is a 9^{th} order polynomial fitted over the normalized error in localization based on ρ (Fig. 5(b)). This weighting mechanism will only penalize audio detections in overlapping regions and will give a weight of at least 0.5 to the video modality, if available.

This weighting has contributed to a 13.17% error reduction. Moreover, the error standard deviation has also decreased by approximately 1 decimal place (when evaluated on 50 randomly generated trajectories, each consisting of 1500 points and a total of 2928 points in the visible region of a single sensor in a network of 3 STAC sensors). Figure 6 shows examples of the obtained improvement using this dynamic weighting technique compared to using equal weights for both modalities. The trajectories estimated from the audio are further smoothed using Kalman filter (Fig. 7).

4. EXPERIMENTAL RESULTS

To analyze the benefits of audiovisual tracking, we performed the comparison with (i) audio only tracking (TDOA), (ii) video only tracking (CLUTE [2]), (iii) audiovisual fusion using dynamic weighting (AV), and (iv) audiovisual fusion using dynamic weighting with trajectory smoothing using Kalman filter (AVKF). To evaluate the robustness of each algorithm we further performed the test with missing audio observations.

The trajectory estimation was performed with 4 different sensor configuration consisting of 2, 3, 4 and 5 STAC sensors, respectively. The evaluation is performed on 181 trajectories containing approximately 2200 points each (see Fig. 8). All the trajectories pass through the FOV of each STAC sensor to have fair comparison with [2]. These trajectories are generated using visual and audio signals. The audio data are generated by transmitting an audio signal from the position of the target and then recording it at the sensor location after applying environmental constraints (see Section 3). Synthetic video data are generated using a first-order motion model defined as

$$\mathbf{x}(t+1) = U\mathbf{x}(t) + \mathcal{N}(\mu, \Sigma), \qquad (5)$$

where U is the observation model which ensures smooth transformation of the target state at time t to the next state at time $t+1$ and

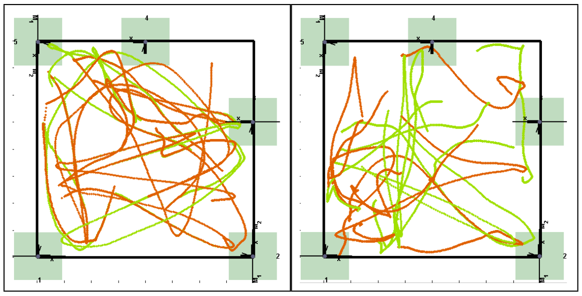

Fig. 8. Sample synthetic trajectories for 2 targets in a network of 5 STAC sensors.

is defined as

$$U = \begin{bmatrix} 1 & 0.35 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.35 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad (6)$$

where 0.35 is chosen to maintain slow target speed. $\mathcal{N}(\mu, \Sigma)$ is a zero-mean Gaussian noise and serves as a process noise to introduce small variation in the motion. The covariance Σ of this process noise is defined as

$$\Sigma = diag[10^{-10}, 10^{-6}, 10^{-10}, 10^{-6}]. \quad (7)$$

Table 1 shows the mean (μ) and standard deviation (σ) of the trajectory estimation error for TDOA, AV, AVKF and CLUTE. The estimation error is the Euclidean distance between the estimated and the true target position. The algorithm is also evaluated with approximately 50% randomly missing audio observation as in real data the targets will not be producing continuous audio signal. The error after applying smoothing using Kalman filtering is increased compared to TDOA and AV Fusion only. This is mainly because Kalman filtering estimation deteriorates when the target exhibits sharp turns. The error in trajectory estimation with audiovisual data is due to the approximation in computing delay in samples. The delay is estimated in seconds since the delay can only be by a discrete number of samples. Hence this rounding off introduces a quantization effect (Fig. 7) and creates an error of maximum $0.2233°$ (considering a rounding off error of 0.5 samples at 44.1kHz) in the arrival angle estimation. This error can be reduced by increasing the sampling frequency. Table 1 also shows that with missing audio observation there is an increase in error by 0.017 units for Kalman filtering on AV Fusion for 5 STACs, whereas the increase in mean error for AV Fusion for 5 STACs is 0.018. This increase of a small value indicates that audio estimation can also be used for complete path estimation in case of non-continuous audio observations. Note that the error for CLUTE remains the same as it does not depends on audio observations.

5. CONCLUSIONS

We have shown that audio can be used in a multiple non-overlapping camera setting to estimate track information in regions unobserved by visual sensor, and presented a trajectory estimation algorithm for wide-area surveillance. It is also demonstrated that using sensors with a large coverage area together with cameras enables extended target tracking. Future work include evaluation on real data and the extension of the proposed approach for improving target hand-over across multiple cameras and event detection.

Table 1. Accuracy comparison for trajectory estimation using TDOA, AV, AVKF and CLUTE (see text for definitions) using 2,3,4 and 5 STAC sensors without(W)/with missing(M) audio observations

		number of sensors							
		5		4		3		2	
		W	M	W	M	W	M	W	M
TDOA	μ	0.0373	0.0579	0.0414	0.0623	0.0560	0.0771	0.4607	0.4833
	σ	0.0662	0.0946	0.0780	0.1048	0.1115	0.1400	0.5242	0.5437
AV	μ	0.0278	0.0461	0.0319	0.0507	0.0434	0.0623	0.4078	0.4833
	σ	0.0459	0.0657	0.0562	0.0761	0.0762	0.0968	0.4812	0.5437
AVKF	μ	0.5181	0.5347	0.5183	0.5351	0.5191	0.5362	0.7681	0.4833
	σ	0.2925	0.3019	0.2922	0.3019	0.2921	0.3021	0.5319	0.5437
CLUTE	μ	4.4561	4.4561	5.3760	5.3760	6.1361	6.1361	6.4242	6.4242
	σ	4.0786	4.0786	5.5655	5.5655	6.4181	6.4181	9.7167	9.7167

6. REFERENCES

[1] A. Rahimi, B. Dunagan, and T. Darrell, “Simultaneous calibration and tracking with a network of non-overlapping sensors,” in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Washington, DC, USA, June 2004, vol. 1, pp. 187–194.

[2] N. Anjum, M. Taj, and A. Cavallaro, “Relative position estimation of non-overlapping cameras,” in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Honolulu, USA, April 2007.

[3] H. Zhou, M. Taj, and A. Cavallaro, “Target detection and tracking with heterogeneous sensors,” *IEEE Journal Of Selected Topics In Signal Processing (J-STSP)*, vol. 2, no. 4, 2008.

[4] M. Wing-Kin, V. Ba-Ngu, S.S. Singh, and A. Baddeley, “Tracking an unknown time-varying number of speakers using tdoa measurements: a random finite set approach,” *IEEE Trans. on Signal Processing*, vol. 54, no. 9, pp. 3291–3304, September 2006.

[5] A. O’Donovan, R. Duraiswami, and J. Neumann, “Microphone arrays as generalized cameras for integrated audio visual processing,” *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pp. 1–8, June 2007.

[6] N. Checka, K.W. Wilson, and M.R. Siracusa T. Darrell, “Multiple person and speaker activity tracking with a particle filter,” in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Cambridge, MA, USA, May 2004, vol. 5.

[7] A. Abad et al., “UPC audio, video and multimodal person tracking systems in the CLEAR evaluation campaign,” in *CLEAR, Springer LNCS 4122*, Southampton, UK, April 2006.

[8] H. Zhou, M. Taj, and A. Cavallaro, “Audiovisual tracking using STAC sensors,” in *ACM/IEEE Int. Conf. on Distributed Smart Cameras*, Vienna, Austria, September 25-28 2007.

[9] R. Chellappa, G. Qian, and Q. Zheng, “Vehicle detection and tracking using acoustic and video sensors,” in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, College Park, MD, USA, May 2004, vol. 3.