ACCURATE TARGET ANNOTATION IN 3D FROM MULTIMODAL STREAMS

Oswald Lanz[1], *Alessio Brutti*[1], *Alessio Xompero*[2], *Xinyuan Qian*[2], *Maurizio Omologo*[1], *Andrea Cavallaro*[2]

[1]ICT-irst, Fondazione Bruno Kessler, Trento, Italy
[2]Centre for Intelligent Sensing, Queen Mary University of London, UK

ABSTRACT

Accurate annotation is fundamental to quantify the performance of multi-sensor and multi-modal object detectors and trackers. However, invasive or expensive instrumentation is needed to automatically generate these annotations. To mitigate this problem, we present a multi-modal approach that leverages annotations from reference streams (e.g. individual camera views) and measurements from unannotated additional streams (e.g. audio) to infer 3D trajectories through an optimization. The core of our approach is a multi-modal extension of Bundle Adjustment with a cross-modal correspondence detection that selectively uses measurements in the optimization. We apply the proposed approach to fully annotate a new multi-modal and multi-view dataset for multi-speaker 3D tracking.

Index Terms**—** Multi-modal annotation; Multi-view; Audio-visual speaker tracking.

1. INTRODUCTION

The annotation of multi-modal and multi-sensor datasets is cumbersome in the absence of an accurate positioning system, which is expensive or invasive [1, 2, 3, 4, 5, 6]. Furthermore, annotation errors may be introduced that significantly bias any further evaluation procedure when shortcuts to speed-up the process are taken, such as interpolating between two manually annotated frames [7].

Annotations are easy to produce on the image plane and therefore most audio-visual datasets focus on image-plane tracking only [8, 9, 10, 11]. However, measuring the accuracy of the results of a 3D tracker requires annotations in the target domain (i.e. the 3D space). Moreover, the aggregation of annotations of the same target from multiple visual streams requires a selection or fusion step. Finally, expensive instrumentation and human intervention are necessary to calibrate all the sensors used in the data collection and to annotate non-visual streams, such as data from microphones.

In this paper, we present MM-BA, a Multi-Modal extension of Bundle Adjustment (BA) [12] that infers the 3D trajectories of freely moving speakers from multi-modal streams captured by multiple sensors (cameras and microphones). MM-BA selects sensor measurements only when supporting 3D estimates with a low re-projection error in the respective stream within a cross-modal correspondence selection mechanism. The proposed approach minimizes in each stream a re-projection error that is a function of the 3D location and calibration parameters. Given an initialization of the positions of the targets from multiple views, such as an algorithmic estimation of mouth locations on the image plane or a manual annotation, MM-BA infers their 3D trajectories through an optimization that involves measurements from non-annotated additional streams, such as cross-correlation features from paired audio streams, as well as the calibration of the sensors for each stream (see Fig. 1). We apply the proposed approach on a new audio-visual dataset with multiple speakers, CAV3D, which we distribute to the research community together with its full annotation[1].

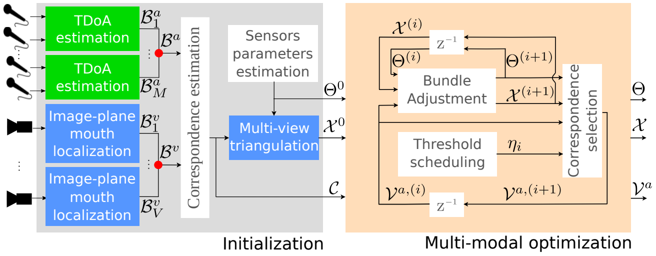


Fig. 1. Our multi-modal approach to define trajectories in 3D, $\mathcal{X}$, and in the audio-visual streams, $(\mathcal{B}^a, \mathcal{B}^v)$, as well as the calibration parameters, Θ. A cross-modal correspondence selection updates the flags $\mathcal{V}^a \subset \mathcal{B}^a$ of the speech presence for the audio observations $\mathcal{B}^a$. The multi-modal optimization iterates until convergence over Bundle Adjustment and the correspondence selection (TDoA: Time Difference of Arrival; • union operation).

2. PROBLEM FORMULATION

Let $x_{n,t} \in \mathbb{R}^3$ be the position in the 3D space of target n at time t and $\mathcal{X}_n$ its trajectory (with duration $T_n + 1$) defined as

$$\mathcal{X}_n = \{x_{n,t} : t = t_{n,0}, \dots, t_{n,T_n}\}. \quad (1)$$

Let an audio-visual recording system with M pairs of microphones[2] and V cameras generate for the same duration synchronized streams of audio and video frames.

The trajectory of a target n is observed as an annotation, $b_{n,t,s}$, within the time interval of the stream captured by each sensor $s = 1, \dots, M + V$, whose sensing model is $f(\cdot)$:

$$b_{n,t,s} = f(x_{n,t}, \theta_s) + \epsilon_{n,t}, \quad (2)$$

where θ_s contains the calibration parameters of the sensor and $\epsilon_{n,t}$ models the noise (e.g. interferences, reverberations, or background noise for the audio and clutter and noise for the video).

The annotation of the trajectory of target n in the stream of a visual sensor $k = 1, \dots, V$ (e.g. the bounding boxes of a face or a point indicating the position of a mouth on the image plane) is

$$B_{n,k} = \{b_{n,t,k}, \nu_{n,t,k} : t = t_{n,0}, \dots, t_{T_n}\}, \quad (3)$$

[1] `https://ict.fbk.eu/units/speechtek/CAV3D/`

[2] We model each microphone pair as a single acoustic sensor.

where $\nu_{n,t,k} \in \{0,1\}$ denotes the visibility of $b_{n,t,k}$ at time t: when $b_{n,t,k}$ is undefined because the target is unobservable due to an occlusion or because it is outside the field of view, then $\nu_{n,t,k} = 0$. The set of all visual annotations is $\mathcal{B}^v = \{B_{n,k}\}$. The sensing model for camera k is a perspective camera whose calibration parameters are the 6D camera pose and 3 intrinsics (focal length and principal point), $\theta_k \in \mathbb{R}^9$ [13]. The set of all camera calibration parameters is $\Theta^v = \{\theta_k\}$.

The annotation of the trajectory of target n in the stream of an acoustic sensor $m = 1, \ldots, M$ is

$$B_{n,m} = \{b_{n,t,m}, \nu_{n,t,m} : t = t_{n,0}, \ldots, t_{T_n}\}, \qquad (4)$$

where $\nu_{n,t,m} \in \{0,1\}$ denotes the presence or absence of speech[3] at time t. This annotation is a set of pair-wise TDoAs associated to the acoustic signal generated by the speaker at each position $x_{n,t}$ and received by the microphone pair m. The set of all audio annotations is $\mathcal{B}^a = \{B_{n,m}\}$. The sensing model for the microphone pair m is the TDoA function whose calibration parameters are the 3D positions of the pair of microphones, $\theta_m = (\mu_{m,1}, \mu_{m,2}) \in \mathbb{R}^6$ with $\mu_{m,q} \in \mathbb{R}^3$ and $q \in \{1,2\}$. The set of all microphone calibration parameters is $\Theta^a = \{\theta_m\}$.

To conclude, $\Theta = \Theta^a \cup \Theta^v$ denotes the calibration parameters of all the sensors in the system, $\mathcal{B} = \mathcal{B}^a \cup \mathcal{B}^v$ is the set of all annotations of the trajectories of all targets in all sensors, and $\mathcal{X} = \cup_{n=1}^{N} \mathcal{X}_n$ is the set of all the 3D trajectories, and N is the number of targets. Given a partial annotation of Θ and $\mathcal{B}^v$, the problem is to produce a complete annotation of the system that comprises

- $\mathcal{X}$: the 3D trajectories;
- Θ: the calibration parameters of all the sensors; and
- $\mathcal{B}$: the location of all targets' observations in all the streams.

3. MULTI-MODAL BUNDLE ADJUSTMENT

We now demonstrate how to extend BA to account for other modalities, such as audio, and complement the optimization with a cross-modal correspondence selection. We term this approach Multi-modal Bundle Adjustment (MM-BA).

3.1. Bundle Adjustment extension

The problem of simultaneously estimating 3D speaker trajectories and sensor parameters is closely related to the image-based 3D reconstruction problem, which is usually solved with feature-based Structure-from-Motion (SfM) [13, 14]. The goal of SfM is to accurately estimate the 3D scene structure and the camera parameters (motion) using as input multiple 2D views (images). After obtaining an initial estimation of the 3D scene points and the camera motion through feature point matching, SfM uses BA [12], which minimizes the error between the feature points and 3D points re-projected in each view.

While SfM relies on multiple correspondences of feature points extracted from each view, in our multi-modal scenario we obtain one multi-modal correspondence for each target position $x_{n,t}$ by grouping the corresponding annotations from the respective frame of each sensor s. We treat each 3D point in a trajectory as an independent point and denote it with $x_j = x_{n,t}$. We thus rewrite $\mathcal{X}$ as $\mathcal{X} = \{x_j : j = 1, \ldots, J\}$, where $J = \sum_{n=1}^{N}(t_{T_n} - t_{n,0} + 1)$. In the rest of the paper, depending on the context, we will refer to $\mathcal{X}$ as set of trajectories or set of (independent) 3D points.

Let $\mathcal{C} = \{(..., b_{j,s}, \nu_{j,s}, ...) : j = 1, ..., J; s = 1, ..., M+V\}$ be the set of all the correspondences. Given a sufficient number of frames with a correspondence, we use 3D reconstruction to jointly optimize the calibration parameters, Θ, and trajectories in 3D, $\mathcal{X}$, by minimizing the re-projection error on all the streams simultaneously, such that

$$(\mathcal{X}^*, \Theta^*) = \underset{(\mathcal{X},\Theta)}{\arg\min} \sum_{s=1}^{M+V} \sum_{j=1}^{J} \nu_{j,s} \| b_{j,s} - f(x_j, \theta_s) \|_2^2. \qquad (5)$$

This objective function leads to a non-linear optimization problem that is usually solved with the Levenberg-Marquardt algorithm [15, 16], also known as BA [12]. The large number of parameters involved ($3 \times J$ speaker parameters, $9 \times V$ camera calibration parameters, and $6 \times M$ audio calibration parameters) makes the optimization computationally expensive [17]. However, as subgroups of parameters are uncorrelated, the Jacobian is sparse thus increasing the efficiency of the solution [17]. We exploit this sparseness, as done in Sparse Bundle Adjustment (SBA) [17, 18], to support the parametrization of the reconstruction problem using arbitrary sensing models[4] $f(\cdot)$. We model a planar circular microphone array with radius R and known height from the ground H, and derive the corresponding Jacobian.

If c is the speed of sound, the expected TDoA for a general microphone pair at sampling frequency F_s is

$$b_j = f(x_j, \theta) = \frac{F_s}{c}(\|x_j - \mu_1\| - \|x_j - \mu_2\|), \qquad (6)$$

with Jacobian

$$\frac{F_s}{c}\left[\frac{x-\mu_1}{\|x-\mu_1\|} - \frac{x-\mu_2}{\|x-\mu_2\|}, \frac{x-\mu_2}{\|x-\mu_2\|}\frac{\partial \mu_2}{\partial \theta} - \frac{x-\mu_1}{\|x-\mu_1\|}\frac{\partial \mu_1}{\partial \theta}\right].$$

For the circular microphone array, the position of each microphone is parametrized with respect to the centre of the array, $\mu_{m,q} = h(\xi, m, q)$ where $\xi = (p_x, p_y, H, \psi, 0, 0) \in \mathbb{R}^6$ is the 6D pose (position and orientation) of the array, with only $(p_x, p_y, \psi) \in R^3$ as free parameters to optimize[5] in Eq. 5. For each pair of opposite microphones, m, the TDoAs, $b_{j,m}$ is defined with positions

$$\begin{aligned}\mu_{m,q} &= [p_x, p_y, H]^T + \\ &R[\cos(\psi + \frac{\pi}{4}(m+q-1)), \sin(\psi + \frac{\pi}{4}(m+q-1)), 0]^T, \qquad (7)\end{aligned}$$

having Jacobian

$$\left[\mathbb{1}_{2\times 3}, R[-\sin(\psi + \frac{\pi}{4}(m+q-1)), \cos(\psi + \frac{\pi}{4}(m+q-1)), 0]^T\right],$$

where $\mathbb{1}_{2\times 3}$ is the 2×3 identity matrix.

This model for the microphone array allows us to project 3D trajectories in the TDoA domain (Eq. 5) and enables MM-BA to accurately calibrate the array coherently with the rest of the system as well as with the reconstructed trajectories.

3.2. Initialization

We initialize the 3D trajectories, $\mathcal{X}^0$, the calibration parameters, Θ^0, and the correspondences, $\mathcal{C}$, as follows (see Fig. 2).

[3] When we cannot validate the measurement due to silence, interfering sources, or non-stationarity of the speaker, then $\nu_{n,t,m} = 0$.

[4] For the camera model and its Jacobian, we refer the reader to [18].

[5] The audio calibration parameters are reduced to 3 under this model.

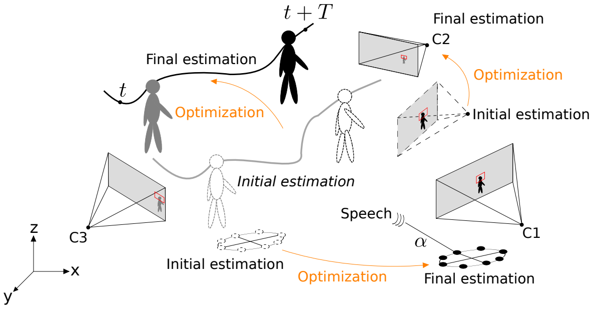


Fig. 2. A speaker observed by multiple cameras (C1, C2, C3) and by a circular microphone array. Our multi-modal approach optimizes an initial sensors calibration (position and orientation) and the 3D trajectory while keeping fixed correspondences of multi-view annotations (red boxes) and audio observations (angle of arrival, α).

After estimating the initial camera calibration parameters using Zhang's method [19], we localize the mouths (faces) in multiple views manually or automatically [20]. Given the calibration and the visual correspondences $\mathcal{C}^v \subset \mathcal{C}$, we obtain x_j by back-projecting each (visible) annotation $b_{j,k}$ of the multi-view correspondence to 3D rays and computing the spatial least-squares intersection of the rays using singular value decomposition [13]. The triangulation of all the correspondences $\mathcal{C}^v$ results in an initial guess of $\mathcal{X}^0$. Next, we randomly initialize on a 2D plane limited by the size of the environment (i.e. the room where the sensors operate) the pose of the microphone array, ξ, and obtain Θ^0, the initial guess for the parameters of all the sensors.

A manual annotation of the TDoA measurements is not possible and therefore we estimate $\mathcal{B}^a_m$, for each microphone pair, m, as the peak of the Generalized Cross Correlation with Phase Transform (GCC-PHAT) [21]. To estimate speech activity, $\nu_{j,m}$, for each target location, x_j, we threshold the GCC-PHAT values associated to the estimated $b_{j,m}$.

We initialize the set of multi-modal correspondences, $\mathcal{C}$, by associating $\mathcal{B}^v$, the manual visual annotations, across cameras and with the estimated audio annotations, $\mathcal{B}^a$.

3.3. Cross-modal correspondence selection

While the set of correspondences $\mathcal{C}$ is fixed during the SBA optimization, after convergence a *selection mechanism* updates $\mathcal{V}^a = \{\nu_{j,m}\} \subset \mathcal{B}^a$ to infer which TDoA measurements $b_{j,m}$ are valid observations originating from a target. We re-iterate SBA and the cross-modal selection mechanism until the audio flag no longer changes, i.e. $\nu^{(i+1)}_{j,m} = \nu^{(i)}_{j,m}$, where i is the iteration index[6].

To consider only valid measurements in the SBA step, we update $\mathcal{V}^a_m$ at iteration i by applying a threshold η_i to the residual between the re-projection of the target positions $x^{(i)}_j$ in the TDoA domain and the estimated TDoAs, $b_{j,m}$ as

$$\nu^{(i+1)}_{j,m} = \begin{cases} 1 & \text{if } \|b_{j,m} - f(x^{(i)}_j, \theta^{(i)}_m)\|_2 < \eta_i \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

By updating η_i according to a pre-defined annealing schedule, we filter out measurements that do not comply with the acoustic sensing model $f(\cdot)$ and the current 3D estimates $\mathcal{X}^{(i)}$, and also recover TDoAs that were previously excluded from the optimization.

Algorithm 1: MM-BA with correspondence selection

```
Input :  $\mathcal{B}^v, \mathcal{B}^a, \eta$
Initialize:  $\Theta^0, \mathcal{X}^0, \mathcal{V}^{a,0} \subset \mathcal{B}^a, \quad \mathcal{C} = (\mathcal{B}^a, \mathcal{B}^v), \quad i \leftarrow 0$
repeat
    // optimize 3D trajectories and sensor parameters
    $(\mathcal{X}^{(i+1)}, \Theta^{(i+1)}) \leftarrow \text{SBA}(\mathcal{X}^{(i)}, \Theta^{(i)}, \mathcal{C})$
    // update correspondences with Eq. 8
    $\nu^{a,(i+1)}_{j,m} \leftarrow (\|b^a_{j,m} - f(x^{(i+1)}_j, \theta^{a,(i+1)}_m)\|_2 < \eta_i)$
    $i+1 \leftarrow i$
until $\mathcal{V}^{a,(i+1)} = \mathcal{V}^{a,(i)}$
Output: $(\mathcal{X}, \Theta, \mathcal{V}^a)$
```

Algorithm 1 summarizes the proposed MM-BA with cross-modal correspondence selection.

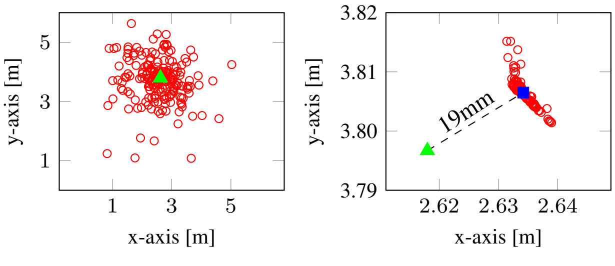
4. RESULTS

We demonstrate MM-BA on CAV3D, a new dataset for 3D speaker tracking collected with a sensing platform that consists of a monocular camera co-located with a circular microphone array [20]. The sensing platform is positioned on a table in a 4.77 x 5.95 x 4.5 m room with reverberation time of approximately 0.7s. The microphone array is composed of 8 high-quality omnidirectional microphones (Shure MX391/O), connected to a pre-amplifier (Focusrite OctoPre LE). We then obtain an analog to digital conversion at 96 kHz 24 bits with a RME ADI-8 DS board. A Marlin F-080C color camera with progressive scan SONY CCD ICX-204AL image device and vari-focal lense recorded the video at 15 frames per second. In addition to the sensing platform, we used other four CCD F-080C cameras installed at the top corners of the room. The field of view of the cameras is about 90°. All five cameras are hardware-synchronized using external trigger shutter. We manually synchronize the audio with the visual streams.

The dataset includes 20 sequences whose duration varies from 15 to 80s and are organized in three sessions: CAV3D-SOT: nine sequences with a single speaker; CAV3D-SOT2: six sequences with a single active speaker but another person is in the scene as interferer; CAV3D-MOT: five sequences with up to three targets speaking simultaneously. In each session, speakers undergo occlusions and abrupt direction and pose changes thus creating non-frontal views, and may also walk outside the field of view of the camera. Furthermore, an air conditioner produces noise that compounds with human-made noise when a speaker arranges objects, claps or stomps.

We use MM-BA to annotate CAV3D-SOT and CAV3D-SOT2 using all the streams from the microphone array and the cameras, whereas we annotate CAV3D-MOT using the visual modality only as it contains overlapping speech signals for long periods. For the visual streams, we manually annotate on the image plane the position of the mouth of each speaker for each frame from the five cameras. For the annotator to easily update the position of the mouth with a mouse click, we display sequentially in a graphical user interface frames from the same camera with a superimposed 50 x 50 zoomed-in view of the candidate region, centred at the position annotated in the previous frame. For the audio streams, we use the four microphone pairs with the longest distance from each other from the 8-element circular array of radius $R = 10$ cm to obtain the TDoAs with the GCC-PHAT. For calibrating the cameras, we use scene

[6] Note that iterations of the cross-modal correspondence selection differ from the internal iterations of SBA.



(a) (b)

iteration	$\varepsilon_{avg}(\xi)$	$\varepsilon_{var}(\xi)$
i=0	-	$(5.09, 6.61, 8.99) \times 10^{-1}$
i=1	$(3.92, 1.99, 1.97) \times 10^{-2}$	$(1.63, 2.33, 0.48) \times 10^{-3}$
i=2	$(2.05, 1.25, 0.95) \times 10^{-2}$	$(0.38, 1.41, 0.20) \times 10^{-4}$
conv	$(1.65, 1.00, 0.75) \times 10^{-2}$	$(2.29, 1.06, 1.25) \times 10^{-5}$

(c)

Fig. 3. Convergence of MM-BA with 100 random initializations of the position of the microphone array (MA). (a) Sampled initial positions (initialization); (b) final positions at convergence (zoom); (c) mean and variance of the pose of the MA for the first iterations and at convergence (*conv*). KEY – ▲ reference; ○ estimated; ■ mean position of the MA.

markers with their measured 3D positions to estimate intrinsic and extrinsic parameters using the Zhang's method [19]. While MM-BA does not depend on any time index and any data stream, the annotated correspondences, $\mathcal{C}$, provide the temporal and target indexes consistency between domain-dependent annotations, $\mathcal{B}$, along with the independence of the target locations, x_j. This allows us to perform the optimization of all sequences jointly to obtain all the 3D trajectories in one run.

To analyse the robustness of MM-BA to random initializations of the microphone array, we sample 100 initial poses centered around the reference position $\hat{\xi} = (\hat{p}_x, \hat{p}_y, H, \hat{\psi}, 0, 0)$ (see Fig. 3(a)),

$$\xi^0 = [\hat{p}_x + \rho\cos(\phi), \hat{p}_y + \rho\sin(\phi), H, \hat{\psi}, 0, 0], \tag{9}$$

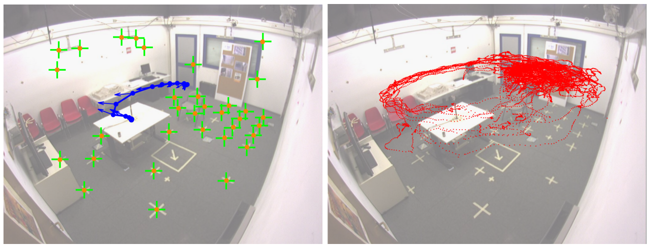
where $H = 0.72$ m, $\hat{\psi}$ and ϕ are sampled from a uniform distribution in $[0, 2\pi]$ and ρ is sampled from a gamma distribution with shape parameter 2.0 and scale parameter 0.5. The reference microphone array position, $\hat{\xi}$, is manually measured with respect to the walls with an expected uncertainty range within 2-3 cm.

We use the same settings of SBA [18] and update the threshold η_i for the speech presence flags, $\mathcal{V}^a$, with the following schedule: 5.0, 2.0, 1.5, 1.2, 1.0,. . . ,1.0. MM-BA requires on average 5-7 outer loop iterations and 120 inner (SBA) iterations to converge. As performance measures we compute the re-projection error $\varepsilon(r)$ (with r being the argument of the minimization in Eq. 5) and the microphone array pose error $\varepsilon(\xi)$ in [m,m,rad] averaged across all the initializations.

Table 1 shows the re-projection error of the estimated trajectories across all initializations. Note that the final estimated trajectories accurately converge to the same positions, as the variance of the 3D positions is only 10^{-7} over iterations (and therefore we do not report it in the table). Moreover, the re-projection error decreases with the iterations and MM-BA achieves at convergence an average pose error, $\varepsilon(\xi)$, of about 19mm around the measured reference, i.e. accurately within the uncertainty range of $\hat{\xi}$ (see Fig. 3(b)).

Table 1. Effect of the optimization process with 100 random initializations of the pose of the microphone array. KEY – $\varepsilon(r)$: re-projection error of all 3D locations; acc($\mathcal{V}^a$): speech presence accuracy (in %); *conv.*: convergence.

iteration	$\varepsilon(r)$	acc($\mathcal{V}^a$)
i=0	17.75	-
i=1	$6.28 \pm 5.21 \times 10^{-2}$	73.55%
i=2	$4.14 \pm 2.61 \times 10^{-4}$	74.01%
conv.	$3.57 \pm 5.33 \times 10^{-6}$	75.16%



(a) (b)

Fig. 4. Annotation of all CAV3D sequences for one random initialization of the microphone array pose. (a) Trajectory of the array during the optimization converging on the table, and scene markers. The arrows represent the orientation of the array. (b) Projected 3D trajectories after optimization. KEY – + scene markers; ● scene markers re-projections; ● microphone array poses; ● projected 3D target locations.

We also measure the accuracy of our cross-modal correspondence selection for the speech presence estimation, acc($\mathcal{V}^a$), using as reference the speech/non-speech segments annotation provided with CAV3D. We note in Table 1 that the accuracy improves as the optimization progresses, despite a more stringent threshold on the TDoA re-projection error is used to select the valid samples. Note that some TDoAs may be unreliable also in presence of speech, due to overlapped coherent noise sources. Therefore the algorithm may appropriately discard some of these frames.

Finally, Fig. 4 shows a view from the CAV3D dataset with the result from one experiment of Table 1. We overlay the trajectory of the array traced during the optimization, the re-projections of scene markers, and the estimated trajectories at convergence.

5. CONCLUSION

We presented an optimization method for the joint estimation of 3D target trajectories and the calibration parameters of multi-modal sensors. The core of the method is a multi-modal extension of Bundle Adjustment, complemented by a cross-modal correspondence validation step. We applied the method to accurately annotate a novel audio-visual dataset for speaker tracking in 3D, CAV3D, even in the presence of only rough planar initializations of the pose of the microphone array.

CAV3D is available with 2D mouth locations manually annotated on the image plane, calibration and synchronization of all sensors, annotations of 3D trajectories and of speech-activity segments. The dataset also provides the audio-visual recordings and the manual annotations of the 2D mouth locations on the frames of four cameras placed on the top-corners of the room.

6. REFERENCES

[1] J. Carletta, “Announcing the AMI meeting corpus,” *The ELRA Newsletter*, vol. 11, no. 1, pp. 3–5, Jan.–Mar. 2006.

[2] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. Chu, A. Tyagi, J. Casas, J. Turmo, L. Cristoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet, “The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms,” *J. Lang. Res. Eval.*, vol. 41, no. 3, pp. 389–407, Dec. 2007.

[3] G. Lathoud, J. Odobez, and D. Gatica-Perez, “AV16.3: An audio-visual corpus for speaker localization and tracking,” in *Int. Workshop Mach. Learn. Multimodal Interaction*, Martigny, Switzerland, 22–23 June 2004.

[4] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and S. Sebe, “SALSA: A novel dataset for multimodal group behavior analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1707–1720, Aug. 2016.

[5] X. Alameda-Pineda, J. Sanchez-Riera, J. Wienke, V. Franc, J. Cech, K. Kulkarni, A. Deleforge, and R. Horaud, “RAVEL: An annotated corpus for training robots with audiovisual abilities,” *J. Multimodal User Interfaces*, vol. 7, no. 1-2, pp. 79–91, Mar. 2013.

[6] E. Arnaud, H. Christensen, Y. Lu, J. Barker, V. Khalidov, M. Hansard, B. Holveck, H. Mathieu, R. Narasimha, E. Taillant, F. Forbes, and R. Horaud, “The CAVA corpus: synchronised stereoscopic and binaural datasets with head movements,” in *Proc. Int. Conf. Multimodal Interfaces*, Chania, Crete, Greece, 20–22 Oct. 2008.

[7] R. Sanchez-Matilla and A. Cavallaro, “Confidence intervals for tracking performance scores,” in *Proc. IEEE Conf. Image Process.*, Athens, Greece, 7–10 Oct. 2018.

[8] I. D. Gebru, S. Ba, G. Evangelidis, and R. Horaud, “Tracking the active speaker based on a joint audio-visual observation model,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Santiago, Chile, 11–18 Dec. 2015.

[9] A. Deleforge, R. Horaud, Y. Y. Schechner, and L. Girin, “Co-localization of audio sources in images using binaural features and locally-linear regression,” *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 4, pp. 718–731, Apr. 2015.

[10] I. D. Gebru, S. Ba, X. Li, and R. Horaud, “Audio-visual speaker diarization based on spatiotemporal Bayesian fusion,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1086–1099, May 2017.

[11] “Surveillance performance evaluation initiative (SPEVI): Audiovisual people dataset,” `http://www.eecs.qmul.ac.uk/~andrea/spevi.html`, 2007, [Online].

[12] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, “Bundle adjustment – a modern synthesis,” in *Vision Algorithms: Theory and Practice*, vol. 1883 of *Lecture Notes in Computer Science*, pp. 298–372. Springer-Verlag Berlin Heidelberg, 2000.

[13] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, New York, NY, USA, second edition, 2003.

[14] J. L. Schönberger and J. Frahm, “Structure-from-motion revisited,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 27–30 June 2016.

[15] D. W. Marquardt, “An algorithm for least-squares estimation of nonlinear parameters,” *J. Appl. Math.*, vol. 11, no. 2, pp. 431–441, 1963.

[16] K. Levenberg, “A method for the solution of certain non-linear problems in least squares,” *Quart. Appl. Math.*, vol. 2, no. 2, pp. 164–168, 1944.

[17] M. I. A. Lourakis and A. A. Argyros, “The design and implementation of a generic sparse bundle adjustment software package based on the levenberg-marquardt algorithm,” Tech. Rep., FORTH-ICS, 2004.

[18] M. I. A. Lourakis and A. A. Argyros, “SBA: A software package for generic sparse bundle adjustment,” *ACM Trans. Math. Software*, vol. 36, no. 1, pp. 29–58, Mar. 2009.

[19] Z. Zhang, “A flexible new technique for camera calibration,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.

[20] X. Qian, A. Xompero, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, “3D mouth tracking from a compact microphone array co-located with a camera,” in *Proc. IEEE Int. Conf. Audio, Speech and Signal Process.*, Calgary, Canada, 15–20 Apr. 2018.

[21] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.