Coordinate Mapping Between an Acoustic and Visual Sensor Network in the Shape Domain for a Joint Self-Calibrating Speaker Tracking

Florian Jacob[1]*, Reinhold Haeb-Umbach*

Department of Communications Engineering, University of Paderborn, 33098 Paderborn, Germany
Email: {jacob, haeb}@nt.uni-paderborn.de
Web: nt.uni-paderborn.de

Abstract

Several self-localization algorithms have been proposed, that determine the positions of either acoustic or visual sensors autonomously. Usually these positions are given in a modality specific coordinate system, with an unknown rotation, translation and scale between the different systems. For a joint audiovisual tracking, where the different modalities support each other, the two modalities need to be mapped into a common coordinate system. In this paper we propose to estimate this mapping based on audiovisual correlates, i.e., a speaker that can be localized by both, a microphone and a camera network, separately. The voice is tracked by a microphone network, which had to be calibrated by a self-localization algorithm at first, and the head is tracked by a calibrated camera network. Unlike existing Singular Value Decomposition based approaches to estimate the coordinate system mapping, we propose to perform an estimation in the shape domain, which turns out to be computationally more efficient. The estimation process is embedded into a Random Sample Consensus (RANASC) framework to obtain a noise robust mapping. Simulations of the self-localization of an acoustic sensor network and a following coordinate mapping for a joint speaker localization showed a significant improvement of the localization performance, since the modalities were able to support each other.

1 Introduction

Audiovisual sensor networks have found widespread use in many applications such as teleconferencing systems, smart rooms or surveillance and monitoring systems [1]. Distributed cameras and microphones are used to localize and track a source of interest, and to capture and enhance audio and video signals. In order to steer these sensors to their targets, many algorithms require knowledge about the sensor positions. While the sensor positions can be determined manually, there exist several self-localization techniques, which, determine the position of the sensors automatically. Such a calibration is typically realized by localizing and tracking an object and then determining the position of the sensors such that the location estimates are most plausible.

Visual calibration approaches rely on a known and easily recognizable object that is imaged by all cameras. Features extracted from this object are then used to estimate the camera pose, while Kalman filters are able to handle temporary occlusion of the calibration object [2]. Automatic calibration algorithms like [3] often use the Scale-Invariant Feature transform algorithm to obtain the required features from an arbitrary scene.

On the acoustic side, time of flight (ToF) based algorithms, which require a tight clock synchronization between transmitter and receiver, in combination with special calibration hardware [4] achieve high positioning accuracies. Time difference of arrival (TDoA) based algorithms use signals with appropriate correlation properties to reduce the synchronization requirements and to accomplish precise results [5].

Self-localization algorithms, that determine the position of either acoustic or visual sensors, are usually unable to return absolute position coordinates in a common world coordinate system. However a common coordinate system is mandatory for a cross-modality localization and tracking of events. In [6] the authors estimate joint coordinate system of a stereo camera and an acoustic echo localization system by matching the trajectories of both modalities using a particle filter.

In this work, the visual sensor network is assumed to be the reference coordinate system. The speech signal of a moving speaker is used as input for an acoustic self-localization algorithm to estimate the positions of the acoustic sensor network. Applying audiovisual correlates, with this we denote events, that can be localized by both sensor networks separately, the mapping from the acoustic to the visual sensor network can be revealed. The event locations in modality specific coordinate systems are described by a set of points. Mapping a set of points from one coordinate frame to another is known as Rigid Body Transformation (RBT). In contrast to the widespread approach from [7] to compute the RBT parameters (scale, rotation and translation) via a Singular Value Decomposition (SVD) we suggest a different and computationally more efficient way. The RBT parameters are computed in the shape domain [8], which is given by the discrete Fourier transform (DFT) for the two-dimensional mapping problem considered here.

In the next section, we review the coordinate mapping problem and its solution by SVD, while Sec. 3 outlines the DFT-based approach. Sec. 4 presents the overall calibration framework to estimate the acoustic sensor positions in its modality specific coordinate system, followed by mapping it to the visual sensor network. Simulation are carried out in Sec. 5, while Sec. 6 concludes this paper.

2 Formulation of the Coordinate Mapping Problem

Our goal is the estimation of a common coordinate system for an audiovisual sensor network in order to perform a joint localization and tracking, where the modalities support each other. After each modality has been calibrated by its domain specific algorithm, we have to estimate the RBT parameters to map the acoustic sensor positions into the visual coordinate system. To this end, a moving speaker is tracked by the microphone and camera network separately. Let the i-th speaker position estimate of the acoustic sensor network be denoted by $\mathbf{m}_i$ and the corresponding estimate

[1]This work has been supported by Deutsche Forschungsgemeinschaft (DFG) under contract no. Ha3455/7-2.

from the visual network by $\mathbf{c}_i$, where each estimate is represented in its modality-specific coordinate system.

To map one coordinate system to the other we have to estimate the rotation matrix $\mathbf{R}$, the translation vector $\mathbf{t}$ and the scale factor s between the two. These parameters are used to transform every point $\mathbf{m}_i$ measured in the acoustic coordinate system to a point $\mathbf{c}_i$ measured in the visual coordinate system:

$$\mathbf{c}_i = s\mathbf{R}\mathbf{m}_i + \mathbf{t}; \quad i = 0, \ldots, N-1, \tag{1}$$

where N is the total number of observations.

According to the comparison of different techniques to estimate the RBT parameters ($\mathbf{R}$, $\mathbf{t}$ and s), SVD based approaches stand out due to their stability [9].

The RBT parameters can be found by minimizing the following least squares objective function:

$$\langle \mathbf{R}^*, \mathbf{t}^*, s^* \rangle = \underset{\mathbf{R},\mathbf{t},s}{\operatorname{argmin}} \frac{1}{N} \sum_{i=0}^{N-1} \|s\mathbf{R}\mathbf{m}_i + \mathbf{t} - \mathbf{c}_i\|^2. \tag{2}$$

Following [7] the parameters that minimize Eq. (2) can be revealed by a SVD applied to the cross-dispersion matrix

$$\mathbf{D} = \frac{1}{N} \sum_{i=0}^{N-1} (\mathbf{c}_i - \bar{\mathbf{c}})(\mathbf{m}_i - \bar{\mathbf{m}})^{\mathrm{T}}, \tag{3}$$

where $\bar{\mathbf{m}}$ and $\bar{\mathbf{c}}$ denote the average of all $\mathbf{m}_i$ and $\mathbf{c}_i$ respectively. The SVD decomposes the cross-dispersion matrix $\mathbf{D}$ into three matrices: $\mathbf{D} = \mathbf{U}\mathbf{W}\mathbf{V}^{\mathrm{T}}$.

Now the RBT parameters can then be recovered as:

$$\mathbf{R}^* = \mathbf{U}\mathbf{V}^{\mathrm{T}}, \;\; \mathbf{t}^* = \bar{\mathbf{c}} - s\mathbf{R}^*\bar{\mathbf{m}}, \;\; s^* = \sigma_m^{-2} \operatorname{trace}\left(\mathbf{R}^{\mathrm{T}}\mathbf{D}\right), \tag{4}$$

where σ_m^2 is the trace of sample covariance matrix of all points $\mathbf{m}_i, i = 0, \ldots, N-1$.

3 Estimation in the Shape Domain

Statistical shape analysis is used in many areas of research, mainly in geodesy and biology, to measure the similarity of objects, irrespective of their orientation, translation and scale. Each object is described by a set of Cartesian coordinates called landmarks. To obtain a representation of these landmarks irrespective of orientation, translation and scale these landmarks are transformed into the shape domain.

A particular instance of shape analysis is the distortion-free transformation from one coordinate system into another, that is described by Eq. (1), and which is known as Helmert-Transformation. In 3-dimensional space this transformation is specified by 7 parameters which are computed by pairs of landmarks, that are given in both coordinate systems. In this paper we concentrate on a coordinate mapping in the plane, i.e. in two dimensions. For this case a particularly simple way of computing the coordinate mapping will be presented, which is based on the discrete Fourier transform.

To this end, we introduce a notation to describe the acoustic position estimates as complex numbers $u_i = m_{i,1} + \mathrm{j}m_{i,2}$, and the visual estimates as $v_i = c_{i,1} + \mathrm{j}c_{i,2}$, respectively. Thus the coordinate mapping is expressed as:

$$v_i = \alpha u_i + \beta \tag{5}$$

where α corresponds to rotation and scale ($s\mathbf{R}$) and β to the translation $\mathbf{t}$. Stacking all observations into vectors $\mathbf{u} = [u_0, \ldots, u_{N-1}]$ and $\mathbf{v} = [v_0, \ldots, v_{N-1}]$, and applying the previous notation to the objective function of Eq. (2) leads to:

$$\langle \alpha^*, \beta^* \rangle = \underset{\alpha,\beta}{\operatorname{argmin}} (\alpha\mathbf{u} + \beta\mathbf{1} - \mathbf{v})^{\mathrm{H}} (\alpha\mathbf{u} + \beta\mathbf{1} - \mathbf{v}), \tag{6}$$

where $\mathbf{1}$ denotes an N-element vector of ones and $(\cdot)^{\mathrm{H}}$ the complex conjugate transpose of a matrix or vector.

The discrete Fourier transform (DFT), which can be computed in a very efficient way using the FFT algorithm, provides an attractive algorithm to obtain a shape domain representation of our landmarks. Let $\mathbf{x}$ be the DFT of $\mathbf{u}$ and $\mathbf{y}$ the DFT of $\mathbf{v}$ respectively, then the optimization problem of Eq. (6) can be reformulated using the orthogonal properties of the DFT as follows:

$$\langle \alpha^*, \beta^* \rangle = \underset{\alpha,\beta}{\operatorname{argmin}} (\alpha\mathbf{x} + \beta\mathbf{e} - \mathbf{y})^{\mathrm{H}} (\alpha\mathbf{x} + \beta\mathbf{e} - \mathbf{y}), \tag{7}$$

where $\mathbf{e} = [1, 0, \ldots, 0]^{\mathrm{T}}$.

Due to the orthogonal properties of the DFT the 2-dimensional optimization problem is decoupled into two separate optimizations:

$$\alpha^* = \underset{\alpha}{\operatorname{argmin}} (\alpha\mathbf{x}_{2:N} - \mathbf{y}_{2:N})^{\mathrm{H}} (\alpha\mathbf{x}_{2:N} - \mathbf{y}_{2:N}) \text{ and} \tag{8}$$

$$\beta^* = \underset{\beta}{\operatorname{argmin}} (\alpha^* x_1 + \beta - y_1)^{\mathrm{H}} (\alpha^* x_1 + \beta - y_1), \tag{9}$$

where the first bin of the DFTs is denoted by $(\cdot)_1$ and all other bins by $(\cdot)_{2:N}$. Since Eq. (8) is a general least squares problem the solution is found to be

$$\alpha^* = \mathbf{x}_{2:N}^{\mathrm{H}} \mathbf{y}_{2:N} / \left(\mathbf{x}_{2:N}^{\mathrm{H}} \mathbf{x}_{2:N}\right), \tag{10}$$

whereas Eq. (9) is just a linear equation, which is minimized by

$$\beta^* = y_1 - \alpha^* x_1. \tag{11}$$

The resulting algorithm consists of three steps. First compute the DFTs of the complex observation vectors, then evaluate Eq. (10) and finally determine the translation using Eq. (11). If only the rotation and translation are to be estimated the factor α^* obtained for Eq. (10) has to be normalized to unit length ($\bar{\alpha}$) to preserve the original scale, before it is plugged into Eq. (11). The RBT parameters in the Cartesian domain can be recovered from the shape domain result as follows:

$$s = |\alpha|, \; \mathbf{R} = \begin{bmatrix} \Re\{\frac{\bar{\alpha}}{s}\} & -\Im\{\frac{\bar{\alpha}}{s}\} \\ \Im\{\frac{\bar{\alpha}}{s}\} & \Re\{\frac{\bar{\alpha}}{s}\} \end{bmatrix} \text{ and } \mathbf{t} = \frac{1}{N} \begin{bmatrix} \Re\{\beta\} \\ \Im\{\beta\} \end{bmatrix}, \tag{12}$$

where $\Re$ and $\Im$ denote real part and imaginary part, respectively. The RBT parameter estimation in the shape domain has been derived from the same objective function as the SVD based approach. Thus, both algorithms deliver the same results, up to the numerical precision boundary.

4 System Overview

Before the RBT parameters can be estimated based on common event localizations in both modalities, one has to acquire the location data. Here we assume the visual sensor

network has been calibrated in advance, such that the camera positions are perfectly known. The visual system therefore provides the reference coordinate system for the coordinate mapping. Each node of the camera network is connected to a Histogram of Oriented Gradient (HOG) detector, which detects the head and shoulder of a person [10] to get visual DoA estimates. The impinging angles are used in an intersection based approach from [11] to obtain the speaker location.

In case of the acoustic sensor network, each node consists of two microphones, thus, DoA estimates are available. The DoA estimation for the acoustic modality is realized by cross correlating the filter coefficients of a filter-and-sum beamformer (FSB) [12]. The beamformer operates with a sampling rate of 16 kHz at a block length of 128 samples, resulting in a new DoA estimate every 8 ms. To obtain the event locations for the acoustic part with the same intersection based technique, as in the visual part, the locations of the acoustic sensor nodes need to be known. To obtain them, we apply our self-localization algorithm for acoustic sensor networks, which we have originally presented in [13]. In this self-localization algorithm the DoA estimates are used to formulate geometric relations between the sensor and event locations, from which a large nonlinear system of equations is obtained. The solution of this system delivers the most plausible sensor localizations. To obtain good results even in case of imperfect DoA estimates due to room reverberation, the algorithm is embedded into a RANSAC framework for outlier rejection, which has led to a significant error reduction.

Since the algorithm uses solely DoA information only relative sensor positions can be revealed, with an arbitrary scale factor and an unknown rotation and translation to the visual sensor network, which is assumed to be the reference. As it is shown in [14], TDoA measurements can be employed to estimate this scale factor, thus, there remains only the estimation of a rotation and translation to map the acoustic sensor locations to the visual coordinate system. To estimate these RBT parameters the proposed coordinate mapping algorithm in the shape domain can be applied.

It is very critical to estimate precise RBT parameters to be able to benefit from the two modalities in a joint localization and tracking. In the noise free case with perfect event localizations the RBT parameters can be easily recovered, but measurement errors have an impact on the performance. Since the RANSAC framework boosts the performance of the acoustic self-localization process, we also use a RANSAC framework for the RBT parameter estimation. At first, an initial set of position estimates is selected at random to perform the RBT parameter estimation. Two measurements would be sufficient, but our experiments revealed that the best performance is achieved with approximately 10 measurements, since a sufficient variability of the input samples is required for precise results. These parameters are used to determine a subset of measurements, that fit to the model, up to a threshold. Here we used the Euclidean distance between the measurements of the two modalities after transformation to the same coordinate system. If enough measurements are compatible with the model parameters, we incorporate all compatible observations into the RBT parameter estimation, otherwise we start again. This procedure is repeated until either a predefined amount of iterations is reached or enough observations fit to the RBT parameters.

Based on the automatically found locations of the acoustic sensors and the estimated RBT parameters we are able to map the acoustic trajectory into the visual reference coordinate system and then perform a joint localization and tracking. The same intersection based algorithm as before is used to combine both modalities and their corresponding DoA estimates. The overall block diagram of this system is shown in Fig. 1.

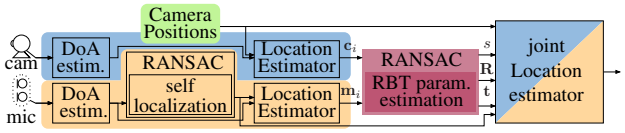
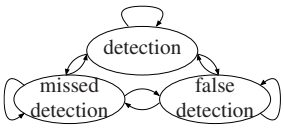


Figure 1: Block diagram of the self-calibrating joint location system.

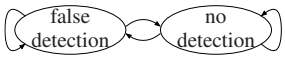
5 Experimental Results

We simulated 5 random speaker trajectories consisting of ≈150 positions, with an approximate duration of 20 min each for 2 different sensor configurations located in a room of 5 m×6 m. 4 virtual cameras and 4 two-element microphone arrays are located in the vicinity of the walls, oriented towards the center of the room. The room impulse response for reverberation times from 0 ms up to 500 ms is simulated using the image method [15]. Acoustic DoA estimates, which reside in the plane, are then obtained from the FSB coefficients, which continuously adapt to the moving sound source.

For the visual part, DoA estimates are simulated as follows. We use Hidden-Markov-Models (HMM) to describe the errors in the DoA estimation. A limited field of view for a camera is taken into account by dropping all angles outside a window of ±30° relative to the mean camera orientation. If the person is located in the visible region of the camera the HMM of Fig. 2a is used, whereas in case of a speaker position outside the visible region the HMM of Fig. 2b is used.



(a) Person in visible region.



(b) No person in visible region.

Figure 2: Error model for visual person detection.

The transition probabilities of these models and the distribution of the error in the DoA estimation process had been estimated on sequences (*seq01-1p-0000* and *seq15-1p-0100*) of the AV16.3 corpus [16]. These sequences provided a total video duration of 12.5 min. We performed a visual DoA estimation using a HOG detection [10] and derived the transition probabilities and DoA error distribution by matching the HMM model to the detection results.

Our first goal is the estimation of the positions of the acoustic sensors in a joint coordinate system with the visual sensors, where the visual sensors are assumed to form the reference coordinate system. The relative positions of the acoustic sensor network, which had been determined by a self-calibration algorithm, need to be mapped onto the visual coordinate system. The error introduced by this acoustic self-calibration procedure is shown in Fig. 3 by a dashed line. Using these noisy position estimates for the acoustic sensor network, each modality localizes the mov-

ing speaker. Then the location estimates of both modalities are used to estimate the RBT parameters. The results in Fig. 3 show that a RBT parameter estimation embedded into a RANASC framework (RBT RANSAC) can outperform a conventional least squares RBT parameter estimation (RBT LS) incorporating all available measurements. Comparing the computational complexity of the SVD and FFT based least squares implementation, the RBT parameter estimation using the FFT is twice as fast as the SVD.

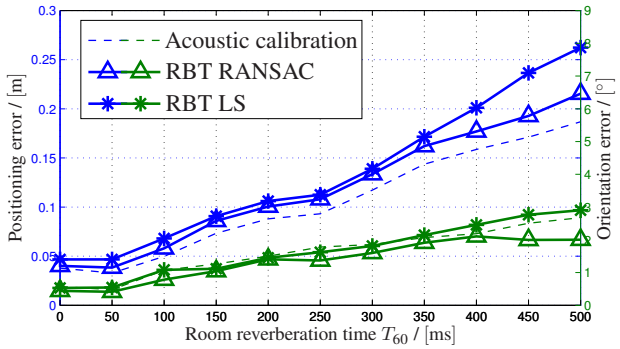


Figure 3: Comparison of positioning (blue) and orientation (green) error for the calibration of the acoustic sensor network (dashed) and the mapped sensor locations using a least squares approach or RANSAC.

Secondly the self-calibrated sensor network is used to localize a moving speaker. The localization performance is shown in Tab. 1. The first two rows show the localization error if either the acoustic or the visual sensor network is used to localize the speaker. Since the speaker is continuously speaking nearly 100 % coverage is achieved for the acoustic part. But in the visual part there is only approx. 57 % coverage, due to event locations outside the field of view or sequences where the speaker is not recognized. If the information of both modalities are combined (third row), whenever at least one modality delivers estimates, again almost 100 % coverage is achieved. In this case the localization accuracy is significantly increased compared to an acoustic only localization. The average localization precision is slightly worse than the average precision of the video localization, this is unsurprising, since the error is dominated by sections where only audio information are available. If only those parts of the trajectory, where both modalities are available at the same time, are considered the joint localization outperforms both acoustic and video only localization.

	T_{60} / [ms]	0	100	200	300	400	500
Error / [m]	audio	0.09	0.18	0.32	0.43	0.55	0.66
	video	0.20	0.20	0.20	0.20	0.20	0.20
	audio $\vee$ video	0.08	0.15	0.22	0.26	0.30	0.33
	audio $\wedge$ video	0.07	0.13	0.16	0.17	0.18	0.18

Table 1: Average localization error in [m] for acoustic only localization (audio), visual only localization (video), combined localization using either acoustic, visual or both modalities, depending on availability (audio $\vee$ video), and joint audiovisual localization, measured on those parts of the trajectory where both estimates are available (audio $\wedge$ video).

6 Conclusions

We have derived a general framework to combine the calibration results of two sensor systems, which are originally given in separate coordinate systems, into a common coordinate system. The positions of the acoustic sensors had been determined by a self-calibration algorithm. We started with an existing approach to estimate the RBT parameters and derived a new strategy to estimate the parameters in the shape domain. Our simulations showed that the RBT parameters can be estimated from events that had been localized by different modalities separately. If this process is used to calibrate an acoustic sensor network to a visual sensor network the localization performance is boosted.

References

[1] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, "A joint particle filter for audio-visual speaker tracking," in *Int. Conf. on Multimodal Interfaces*, 2005.

[2] X. Chen, J. Davis, and P. Slusallek, "Wide area camera calibration using virtual calibration objects," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2000.

[3] M. Brückner and J. Denzler, "Active Self-calibration of Multi-camera Systems," in *DAGM Symp. on Pattern Recognition*, 2010.

[4] M. Crocco, A. Del Bue, M. Bustreo, and V. Murino, "A Closed Form Solution to the Microphone Position Self-Calibration Problem," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2012.

[5] S. Valente, M. Tagliasacchi, F. Antonacci, P. Bestagini, A. Sarti, and S. Tubaro, "Geometric calibration of distributed microphone arrays from acoustic source correspondences," in *IEEE Int. Workshop on Multimedia Signal Processing*, 2010.

[6] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP J. Appl. Signal Process.*, 2002.

[7] J. H. Challis, "A procedure for determining rigid body transformation parameters," *J. of Biomechanics*, 1995.

[8] I. Dryden and K. Mardia, *Statistical shape analysis*. Wiley series in probability and statistics, 1998.

[9] D. W. Eggert, A. Lorusso, and R. B. Fisher, "Estimating 3-D rigid body transformations: a comparison of four major algorithms," *Machine Vision and Applications*, 1997.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2005.

[11] M. Brandstein, J. Adcock, and H. Silverman, "A closed-form method for finding source locations from microphone-array time-decay estimates," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 1995.

[12] E. Warsitz and R. Haeb-Umbach, "Acoustic filter-and-sum beamforming by adaptive principal component analysis," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2005.

[13] F. Jacob, J. Schmalenstroeer, and R. Haeb-Umbach, "Microphone Array Position Self-Calibration from Reverberant Speech Input," in *Int. Workshop on Acoustic Signal Enhancement*, 2012.

[14] J. Schmalenstroeer, F. Jacob, R. Haeb-Umbach, M. Hennecke, and G. Fink, "Unsupervised Geometry Calibration of Acoustic Sensor Networks Using Source Correspondences," in *Interspeech 2011*, 2011.

[15] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustical Society of America*, 1979.

[16] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: an Audio-Visual Corpus for Speaker Localization and Tracking," in *Workshop for Machine Learning for Multimodal Interaction*, 2005.