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

Florian Jacob and Reinhold Haeb-Umbach

Department of Communications Engineering - University of Paderborn

26.09.2014

Table of contents

Introduction

Application areas of audiovisual sensor networks

- Joint speaker localization and tracking
- Advanced teleconferencing systems

Problem statement

- Applications require joint coordinate system
- Existing algorithms: Separate calibration of acoustic or visual sensor networks
- Lack of joint calibration algorithms

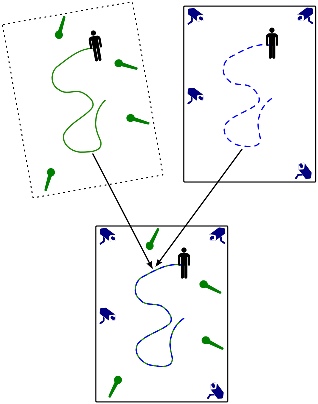
Task

1. Estimate acoustic and visual sensor locations in a joint coordinate frame
2. Use self-calibrating sensor network for audiovisual tracking

Introduction

Our approach

1. Calibration of the acoustic sensor network
2. Use each sensor network to estimate the trajectory of a moving speaker separately
3. Estimate a **coordinate mapping** between both trajectories to obtain joint coordinate frame
4. Perform a cross modality localization and tracking



Coordinate mappingCoordinate mapping in complex space

Conventional approach

- Singular Value Decomposition based [Cha95]
- XY coordinates

Mapping between coordinate frames

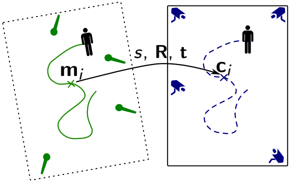
- Acoustic location estimates $\mathbf{m}_i \Rightarrow u_i$
- Visual location estimates $\mathbf{c}_i \Rightarrow v_i$

$$\mathbf{c}_i = s\mathbf{R}\mathbf{m}_i + \mathbf{t} \Rightarrow v_i = \alpha u_i + \beta$$

$\Rightarrow$ Estimate Rigid Body Transformation (RBT):

- Translation $\mathbf{t} \Rightarrow \beta$
- Rotation $\mathbf{R}$
- Scale s

$\left.\begin{matrix}\text{Rotation } \mathbf{R} \\ \text{Scale } s\end{matrix}\right\} \Rightarrow \alpha$



Least squares objective function

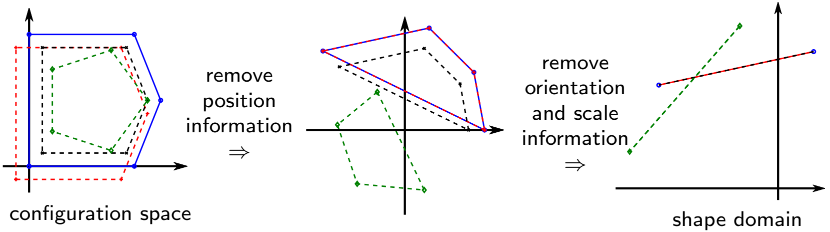
$$\langle \alpha^*, \beta^* \rangle = \underset{\alpha,\beta}{\operatorname{argmin}} \left(\alpha\mathbf{u} + \beta\mathbf{1} - \mathbf{v}\right)^{\mathrm{H}} \left(\alpha\mathbf{u} + \beta\mathbf{1} - \mathbf{v}\right)$$

Introduction to shape domain

Shape domain

- Describe objects irrespective to
  - Location
  - Orientation
  - Scale

} ⇒ Coordinate transformation



e.g. Kendall coordinates [DM98]

1. Remove translation by multiplication with special orthogonal matrix (e.g. Helmert-matrix)
2. Projection to new base vectors

RBT parameter estimation

Use transformation into shape domain to estimate RBT parameter

- Discrete Fourier Transformation (DFT) matrix provides same orthogonal properties as Helmert-matrix
- Transformation into shape decouples joint optimization problem

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

$$\langle \alpha^*, \beta^* \rangle = \operatorname*{argmin}_{\alpha,\beta} \left(\alpha \mathbf{x} + \beta \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \mathbf{y} \right)^{\mathrm{H}} \left(\alpha \mathbf{x} + \beta \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \mathbf{y} \right)$$

into two separate

$$\alpha^* = \frac{\mathbf{x}_{2:N}^{\mathrm{H}} \mathbf{y}_{2:N}}{\left(\mathbf{x}_{2:N}^{\mathrm{H}} \mathbf{x}_{2:N} \right)} \qquad \beta^* = y_1 - \alpha^* x_1$$

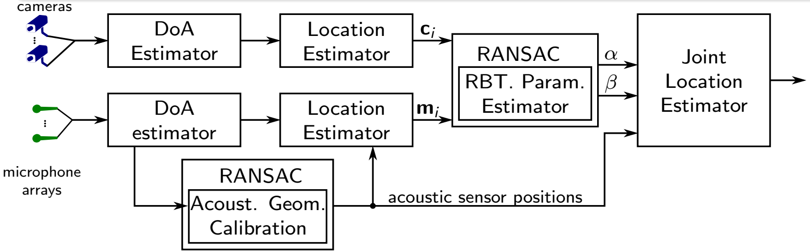
- Computationally efficient realization by FFT
- Shape domain realization ≈ 2.5 times faster than SVD

F: Fourier-transformation matrix

Self-calibration framework

Overview

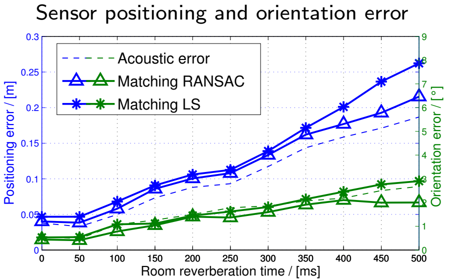
- Acoustic direction of arrival estimation:
  - Filter-and-sum Beamformer, continuously adapt to dominant source
- Visual Direction of arrival estimation:
  - Histogram of orientated gradients of head and shoulder + support vector machine
- Location estimator:
  - Simple intersection based approach
- Random sample consensus algorithm (RANSAC) [FB81]
  - Outlier rejection scheme
- Acoustic sensor positions determined by self-calibration algorithm [JSHU12]



Simulation results

Simulated environment

- Scenario:
  - 4 microphone arrays (2-elements)
  - 4 virtual cameras
  - 2 sensors setups
  - 5 trajectories for each setup
- Data generation:
  - Acoustic signal: Image-method
  - Visual DoA: HMM based error model trained on AV16.3 [LOGP05]



Speaker localization error

	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
	combined	0.08	0.15	0.22	0.26	0.30	0.33
	combined, visual	0.07	0.13	0.16	0.17	0.18	0.18

Summary

Summary

- Estimation of joint coordinate frame:
  - Acoustic self-calibration + separate speaker tracking
- Estimation of mapping parameters:
  - Conventional approach: SVD
  - Proposed: shape domain approach
- Calibration framework:
  - RANSAC significantly decreases calibration error
  - Calibration error: < 0.25 m even at $T_{60} = 500$ ms
  - Joint tracking error with self calibrating sensor network: < 0.18 m

References I

CHALLIS, J. H.:

A procedure for determining rigid body transformation parameters.

In: *Journal of Biomechanics* (1995)

DRYDEN, I. L. ; MARDIA, K. V.:

Statistical shape analysis.

1998 (Wiley series in probability and statistics). –

ISBN 0471958166

FISCHLER, M. ; BOLLES, R.:

Random Sample Consensus: a paradigm for model fitting with applications to image analysis and automated cartography.

In: *Communications of the ACM* (1981)

JACOB, F. ; SCHMALENSTROEER, J. ; HAEB-UMBACH, R.:

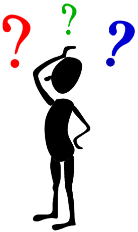
Microphone Array Position Self-Calibration from Reverberant Speech Input.

In: *Int. Workshop on Acoustic Signal Enhancement*, 2012

LATHOUD, G. ; ODOBEZ, J.-M. ; GATICA-PEREZ, D.:

AV16.3: an Audio-Visual Corpus for Speaker Localization and Tracking.

In: *Workshop for Machine Learning for Multimodal Interaction*, 2005



Thank you for your attention!

Questions ?

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

Florian Jacob

University of Paderborn
Department of Communications Engineering

jacob@nt.uni-paderborn.de
nt.uni-paderborn.de