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

1  Introduction & motivation

2  Coordinate mapping

3  Calibration framework

4  Simulation results

5  Conclusion
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
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 mapping in complex space

Conventional approach

- Singular Value Decomposition based [Cha95]
- XY coordinates

Mapping between coordinate frames

- Acoustic location estimates $m_i \Rightarrow u_i$
- Visual location estimates $c_i \Rightarrow v_i$
  
  $c_i = sRm_i + t \Rightarrow v_i = \alpha u_i + \beta$

  $\Rightarrow$ Estimate Rigid Body Transformation (RBT):
  
  - Translation $t \Rightarrow \beta$
  - Rotation $R \Rightarrow \alpha$
  - Scale $s$

Least squares objective function

$$\langle \alpha^*, \beta^* \rangle = \arg\min_{\alpha, \beta} (\alpha u + \beta 1 - v)^H (\alpha u + \beta 1 - v)$$
Shape domain

- Describe objects irrespective to
  - Location
  - Orientation
  - Scale

\[ \Rightarrow \text{Coordinate transformation} \]

configuration space

- Remove position information

- Remove orientation and scale information

shape domain

e.g. Kendall coordinates [DM98]

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

\[
\langle \alpha^*, \beta^* \rangle = \arg \min_{\alpha, \beta} (\alpha u + \beta \mathbf{1} - v)^{\mathsf{H}} F^{\mathsf{H}} F (\alpha u + \beta \mathbf{1} - v)
\]

\[
\langle \alpha^*, \beta^* \rangle = \arg \min_{\alpha, \beta} \left( \alpha x + \beta \begin{bmatrix} 1 \\ \vdots \\ 0 \end{bmatrix} - y \right)^{\mathsf{H}} \left( \alpha x + \beta \begin{bmatrix} 1 \\ \vdots \\ 0 \end{bmatrix} - y \right)
\] into two separate

\[
\alpha^* = \frac{x_{2:N}^{\mathsf{H}} y_{2:N}}{(x_{2:N}^{\mathsf{H}} x_{2:N})} \quad \beta^* = y_1 - \alpha^* x_1
\]

- Computationally efficient realization by FFT
- Shape domain realization \(\approx 2.5\) 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]

Sensor positioning and orientation error

Speaker localization error

<table>
<thead>
<tr>
<th>$T_{60}$ [ms]</th>
<th>0</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Error [m]</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>audio</td>
<td>0.09</td>
<td>0.18</td>
<td>0.32</td>
<td>0.43</td>
<td>0.55</td>
<td>0.66</td>
</tr>
<tr>
<td>video</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>combined</td>
<td>0.08</td>
<td>0.15</td>
<td>0.22</td>
<td>0.26</td>
<td>0.30</td>
<td>0.33</td>
</tr>
<tr>
<td>combined, visual</td>
<td>0.07</td>
<td>0.13</td>
<td>0.16</td>
<td>0.17</td>
<td>0.18</td>
<td>0.18</td>
</tr>
</tbody>
</table>
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: $< 0.25 \text{ m even at } T_{60} = 500 \text{ ms}$
  ▶ Joint tracking error with self-calibrating sensor network: $< 0.18 \text{ m}$
Challis, J. H.:  
A procedure for determining rigid body transformation parameters.  

Dryden, I. L.; Mardia, K. V.:  
*Statistical shape analysis.*  

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.:  
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