Multimodal Tracking for Smart Videoconferencing and Video Surveillance

Dmitry N. Zotkin, Vikas C. Raykar, Ramani Duraiswami, and Larry S. Davis
Perceptual Interfaces and Reality Lab, Institute for Advanced Computer Studies (UMIACS)
University of Maryland, College Park, MD 20742 USA
{dz,vikas,ramani,lsd}@umiacs.umd.edu

Abstract

Many applications require the ability to track the 3-D motion of the subjects. We build a particle filter based framework for multimodal tracking using multiple cameras and multiple microphone arrays. In order to calibrate the resulting system, we propose a method to determine the locations of all microphones using at least five loudspeakers and under assumption that for each loudspeaker there exists a microphone very close to it. We derive the maximum likelihood (ML) estimator, which reduces to the solution of the non-linear least squares problem. We verify the correctness and robustness of the multimodal tracker and of the self-calibration algorithm both with Monte-Carlo simulations and on real data from three experimental setups.

1. Introduction

Many applications require an ability to localize and track a person in an environment. Integration of information obtained from sensors of multiple modalities can lead to improvement of the tracking accuracy and to a practical design for the development of a surveillance, augmented reality, or smart videoconferencing system. In particular, in this work we focus on integration of audio and video measurements for joint audio-visual tracking.

To allow for self-calibration of the tracking system, we propose a method to automatically determine the 3-D positions of multiple microphones by measuring time of flight (TOF) from a few loudspeakers to all microphones. The method does not require knowledge of loudspeaker positions; the only assumption we make is that each loudspeaker has a microphone attached to it. We obtain implicit expression for loudspeaker and microphone positions using ML estimator [1] and derive its mean and covariance.

We then propose a multimodal information fusion algorithm for audio and video measurements obtained from multiple microphone arrays and calibrated cameras using sequential Monte-Carlo methods (also known as particle filters [2]). The proposed tracker is able to seamlessly handle temporary absence of some measurements and to recover dynamically changing self-configuration of the tracking system. We describe a particular setup of the audio-visual tracking system and show simulated and experimental tracking and occlusion handling results.

2. Overview of the work

2.1. Autocalibration of multi-microphone setup

Given a set of M microphones and S loudspeakers in unknown locations, our goal is to estimate their 3-D coordinates. We assume that each of S speakers has a microphone attached to it so the corresponding TOF is very small. A speaker with an attached microphone is a *speaker-microphone pair*; each of the remaining $M-S$ microphones is a *single microphone*.

Let $\mathbf{m}_i$ and $\mathbf{s}_j$ be the coordinates of the i^{th} microphone and j^{th} loudspeaker respectively. We excite each of the S speakers one at a time and measure the TOF at each of the M microphones using PHAT-weighted generalized cross correlation. The TOF_{ij} for the i^{th} microphone and the j^{th} speaker is defined as the time taken for the acoustic signal to travel from the j^{th} speaker to the i^{th} microphone. Let $TOF_{ij}^{estimated}$ and TOF_{ij}^{actual} be the estimated and the actual TOF respectively for the i^{th} microphone and j^{th} speaker. The actual TOF is written as

$$TOF_{ij}^{actual} = \frac{\parallel \mathbf{m}_i - \mathbf{s}_j \parallel}{c}, \tag{1}$$

where $\|\|$ is the Euclidean norm and c is the sound speed.

Let Θ be a vector representing all the unknown non-random parameters to be estimated (microphone and loudspeaker coordinates). Assuming that each TOF is independently corrupted by zero-mean additive white Gaussian noise of variance σ_{ij}^2, we show that the ML estimator reduces to a non-linear least squares formulation:

$$\hat{\Theta}_{ML} = \arg_{\Theta} \min \sum_{i=1}^{M} \sum_{j=1}^{S} \frac{(TOF_{ij}^{estimated} - TOF_{ij}^{actual})^2}{\sigma_{ij}^2}. \tag{2}$$

As the solution depends only on pairwise TOFs, it can be translated and rotated arbitrarily. To eliminate ambiguity, we fix four arbitrary nodes to provide origin, positive X-axis, positive Y-axis, and positive-Z half-space, respectively. Also, we show that it is necessary to have at least five speaker-microphone pairs to supply sufficient information to determine all unknowns.

As the ML estimate is implicitly defined as a minimum of the non-linear function, the minimization has to be performed using numerical optimization methods (*e.g.* Levenberg-Marquardt), requiring a good initial guess for convergence. Using measured TOFs, we obtain an approximate closed-form solution for the speaker-microphone pair locations using multidimensional scaling [3] and then approximately localize the remaining single microphones. We then slightly perturb the obtained coordinates and use them as an initial guess for the Levenberg-Marquardt method to determine final locations of each speaker and each microphone. Using Taylor series expansion and implicit function theorem [4], we also derive expressions for the estimator mean and covariance and plot uncertainty ellipses for estimated coordinates (not given here for brevity). The plots show that the estimator uncertainty is minimized when the loudspeakers are positioned as far away from each other as possible and in such a way that all microphones lie in the convex hull formed by the loudspeakers.

2.2. Multimodal tracking algorithm

The particle filter tracker, also known as a CONDENSATION tracker, was first introduced in the computer vision area by Isard and Blake [5]. The *state vector* X_s describes the state of the tracked object. The *measurement vector* X_m consists of the measurement values obtained from the sensors, which are related to and carry some information about the state of the object X_s. X_m is related to X_s via *measurement equation*. The algorithm maintains a set of points in X_s space along with the weight for each point; these points approximate the PDF on X_s space. The algorithm also defines rules for updating the PDF as new X_m becomes available. During the update, the measurement equation defining how X_m depends on X_s is used. The measurement equation can usually be expressed in the simple form (it is just the projection equation for video measurement or the definition of TDOA for audio measurement) and does not need to be inverted (complicated and error-prone procedure).

We use object position and speed as the state vector and image coordinates and pairwise time differences of arrival (TDOAs) between the microphones as the measurements. To handle unavailability of some data (*e.g.* camera occlusion or silence), we marginalize the PDF update equation to operate on available values only. We are also able to handle the case of the sensor moving and/or rotating itself by including its parameters (*e.g.* rotation angle and rotational velocity) in the state vector.

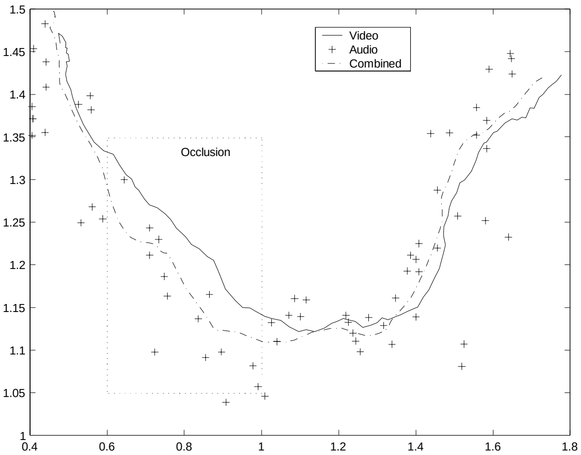

Figure 1. Multimodal speaker tracking with occlusions.

3. Results

We test our algorithms on synthetic data and on three experimental setups (a 32-microphone array for HRTF measurements; a 7-microphone 2-camera array for bat tracking in a quiet room; and a 14-microphone 2-camera array for videoconferencing in an office). We show that the self-calibration algorithm is able to derive the microphone positions with minimal error compared to ground-truth and that the audio-visual tracker is able to meaningfully combine multimodal data to enhance accuracy of the tracker, to support tracking during temporary absence of some measurements, and to recover sensor motion *together* with the object tracking even in the case when only one sensor (out of three) is stationary. Figure 1 shows a track of the speaker position in an office environment. The accuracy of the tracker is somewhat decreased during video occlusion but still stays acceptable for surveillance/videoconferencing purposes.

References

[1] A. J. Weiss and B. Friedlander (1989). “Array shape calibration using sources in unknown locations – a maximum-likelihood approach”, IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-37, no. 12, pp. 1958-1966.

[2] A. Doucet, N. de Freitas, and N. Gordons (eds.) (2001). “Sequential Monte-Carlo Methods in Practice”, Springer, New York, NY.

[3] M. Steyvers (2002). “Multidimensional Scaling”, Nature Publishing Group, London, UK.

[4] A. K. Roy Chowdhury and R. Chellappa (2003). “Stochastic approximation and rate distortion analysis for robust structure and motion estimation”, Intl. J. Computer Vision, vol. 55, no. 1, pp. 27-53.

[5] M. Isard and A. Blake (1996). “CONDENSATION conditional density propagation for visual tracking”, Intl. J. Computer Vision, vol. 29, no. 1, pp. 5-28.