

Audio-Visual Variational Fusion for Multi-Person Tracking with Robots

Xavier Alameda-Pineda

Soraya Arias

Yutong Ban

Guillaume Delorme

Laurent Girin

Radu Horaud

Xiaofei Li

Bastien Mourgue

Guillaume Sarrazin

ABSTRACT

Robust multi-person tracking with robots opens the door to analysing engagement and social signals in real-world environments. Multi-person scenarios are characterised by (i) a time-varying number of people, (ii) intermittent auditory (e.g. speech turns) and visual cues (e.g. person appearing/disappearing) and (iii) impact of the robot actions in perception. The various sensors (cameras and microphones) available for perception, provide a rich flow of information of intermittent and complementary nature. How to jointly exploit these cues to tackle the multi-person tracking problem with an autonomous system has been an intense research line of the Perception Team in the past few years. In this demo we want to present our, now mature, achievements in the field, and demonstrate two robotic systems able to track multiple persons using auditory and visual cues, when they are available. We will bring the two robots and the necessary computing resources with us, as well as the required presentation materials to discuss the models, methods and tools supporting this technology with the attendants.

CCS CONCEPTS

• **Mathematics of computing** → **Variational methods**; • **Information systems** → **Multimedia and multimodal retrieval**; • **Computing methodologies** → **Tracking**; *Scene understanding*; *Vision for robotics*.

ACM Reference Format:

Xavier Alameda-Pineda, Soraya Arias, Yutong Ban, Guillaume Delorme, Laurent Girin, Radu Horaud, Xiaofei Li, Bastien Mourgue, and Guillaume Sarrazin. 2019. Audio-Visual Variational Fusion for Multi-Person Tracking with Robots. In *Proceedings of the 27th ACM Int'l Conf. on Multimedia (MM'19)*, Oct. 21–25, 2019, Nice, France. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3343031.3350590>

1 INTRODUCTION

The automatic understanding of populated spaces in terms of individual as well as collective behavior is a topic of great interest for the multimedia community for several reasons. First, multimedia systems performing surveillance or live recommendation need to

Perception Team, Inria firstname.lastname@inria.fr, Authors in alphabetical order.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

MM '19, October 21–25, 2019, Nice, France

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-6889-6/19/10.

<https://doi.org/10.1145/3343031.3350590>

firstly extract the user or users of interest, before addressing any further processing. Similarly, they need to understand the behavioral cues common to the collectivity and characteristic of each individual. Second, the analysis of social and emotional signals in real-world scenarios, requires the identification of the relevant person(s) beforehand. Third, for many real-world applications the response-time is crucial, as well as the flexibility to accommodate other inputs and produce other outputs. Therefore, the design of the system architecture must be done with care and requires proper modularisation. We chose to address all these three issues at once by proposing a software architecture for audio-visual multi-person tracking and demonstrate it in two robotic platforms.

In the following, we provide a brief description of the tracking formulation, of the software architecture and finally of the demo that we would like to bring to ACM Multimedia 2019.

2 AUDIO-VISUAL TRACKING SYSTEM

Our formulation lies within the tracking-by-detection framework, meaning that, at every time frame t we assume the existence of K_t auditory and M_t visual detections (that will be presented later). The set of auditory and visual detections (observations) is formally denoted as $\mathbf{o}_t = (o_{t,1}, \dots, o_{t,K_t}, \dots, o_{t,K_t+M_t})$. These observations are supposed to be generated by N_t persons with (unobserved) position $\mathbf{s}_t = (s_{t,1}, \dots, s_{t,N_t})$. Additional latent assignment variables serve to indicate which person generated each observation: $z_{t,k} = n$ denotes that observation k was generated by person n at time t , $z_{t,k} = 0$ denotes clutter and $\mathbf{z}_t = (z_{t,1}, \dots, z_{t,K_t+M_t})$. Within a probabilistic framework, the tracking problem writes:

$$\hat{\mathbf{s}}_t = \arg \max_{\mathbf{s}_t} \sum_{\mathbf{z}_t} p(\mathbf{s}_t, \mathbf{z}_t | \mathbf{o}_{1:t}), \quad (1)$$

where $1:t$ indicates from 1 to t .

We now need to define the observation and dynamic models:

$$p(o_{t,k} | z_{t,k}, \mathbf{s}_t) \quad \text{and} \quad p(\mathbf{s}_t | \mathbf{s}_{t-1}). \quad (2)$$

Depending on the applications this probability distributions could be e.g. Gaussian [5], von Mises [3] or von Mises-Fisher [11]. However, in any of these cases, the combinatorial problem associated to the marginalisation of (1), leads to a mixture model with a number of components that grows exponentially with time [4]. Therefore, the exact probability distribution is not computationally tractable after a few visual frames: an approximation is required.

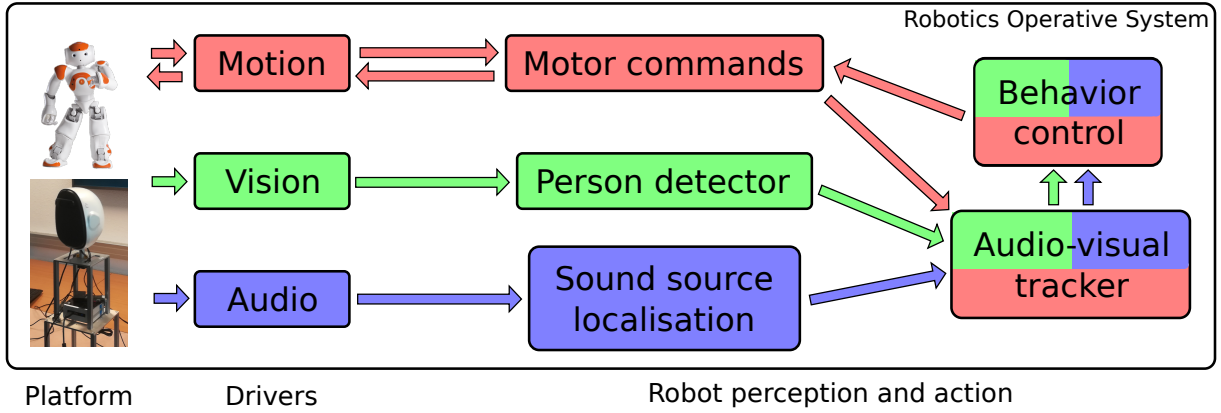


Figure 1: Software architecture of the proposed AV tracking demonstrator.

2.1 The variational approximation

Our privileged choice is a variational approximation [10], in which the a posteriori distribution is assumed to be separable in \mathbf{s}_t and \mathbf{z}_t :

$$p(\mathbf{s}_t, \mathbf{z}_t | \mathbf{o}_{1:t}) \approx q_s(\mathbf{s}_t) q_z(\mathbf{z}_t). \quad (3)$$

There general principles of the variational Bayes theory state that the optimal value (in the sense of the KL divergence) for q_s is given by the following expression:

$$q_s(\mathbf{s}_t) \propto \exp \left(\int_{\mathbf{z}_t} q_z(\mathbf{z}_t) \log p(\mathbf{s}_t, \mathbf{z}_t | \mathbf{o}_{1:t}) d\mathbf{z}_t \right), \quad (4)$$

and the reciprocal holds for q_z .

2.2 Time-varying number of persons

As briefly discussed above, the number of persons N_t varies over time. Therefore, we require a track birth and death process. We tackle this issue through a hypothesis testing problem. At every frame t , we check all the auditory and visual observations considered to be clutter ($z = 0$) in the previous t_B frames. We test two hypothesis: the null hypothesis is that these observations are unstructured, and the alternate hypothesis is that part of the clutter observations are consistent enough to be generated by a person not yet tracked. If that is the case, a new person is created and N_t is increased by one. A track with no assigned detections during t_D frames is discarded (see [4] for more details).

2.3 Auditory and visual observations

Depending on the application the auditory and visual observations can be different, actually there might be several of them. For instance face [8] or body [6] detections for video, and directions of arrival [3] or direct-path identification [9]. Independently of the features used, one must always find a correspondence between the auditory and visual feature spaces. Indeed, since the different cues will naturally live in different mathematical spaces, geometrical [9] or learned [7] mappings must be exploited to be able to fuse the auditory and visual information.

2.4 Self-motion and servoing

When exploiting this technology for tracking with autonomous systems, one must be aware that the actions taken by the system may have an impact in the environment and in the way it is perceived. In our case, when the robot turns the head, or more generally moves, the positions of the persons in the image plane or in the auditory feature space may vary. These variations can often be learned or modeled [2]. This relation can be used for servoing, meaning that we can implement robot behavior rules able to *e.g.* center the person of interest in the image, scan the scene for new potential interactees, move to better hear what the person is saying.

2.5 Software architecture

An overview of the software architecture used in the Demo is shown in Figure 1. We use the Robotics Operative System [1], and develop two kind of nodes: the driver nodes, which are platform-dependent, and the robot perception and action nodes, which are platform-independent. The audio-visual perception and tracking modules constitute the main contribution of the demo, while the behavior control and motor commands consist of basic features.

3 DEMO DESCRIPTION

During this demo, we will demonstrate the audio-visual multi-person tracking capabilities of our system in two different robotic platforms, a commercial robot and a prototype robotic head (shown in the figure). We will also bring with us the necessary support material to explain and describe the various methods and technologies used in our tracking system. Together with the robots, we will bring the necessary computing resources (computer with a GPU) necessary to make the tracking run in real-time while keeping high tracking robustness and quality.

Up to our knowledge, ours is the first multi-person tracking system able to exploit intermittent auditory and visual cues, based on a sound mathematical formulation that has been translated into an on-line algorithm working on robotic platforms. With this demonstrator we would like (i) to show the scientific and technological challenges associated to the problem, (ii) to describe the proposed solution able to tackle all these issues and (iii) to present the software architecture running in two different robotic platforms.

REFERENCES

- [1] Robotics operative system. <https://www.ros.org/>.
- [2] Yutong Ban, Xavier Alameda-Pineda, Fabien Badeig, Sileye Ba, and Radu Horaud. Tracking a varying number of people with a visually-controlled robotic head. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017.
- [3] Yutong Ban, Xavier Alameda-Pineda, Christine Evers, and Radu Horaud. Tracking multiple audio sources with the von mises distribution and variational em. *IEEE Signal Processing Letters*, 26(6):798–802, 2019.
- [4] Yutong Ban, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud. Variational bayesian inference for audio-visual tracking of multiple speakers. *arXiv preprint arXiv:1809.10961*, 2018.
- [5] Yutong Ban, Laurent Girin, Xavier Alameda-Pineda, and Radu Horaud. Exploiting the complementarity of audio and visual data in multi-speaker tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 446–454, 2017.
- [6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [7] Antoine Deleforge, Radu Horaud, Yoav Y Schechner, and Laurent Girin. Co-localization of audio sources in images using binaural features and locally-linear regression. *IEEE Transactions on Audio, Speech, and Language Processing*, 23(4):718–731, 2015.
- [8] Huaizu Jiang and Erik Learned-Miller. Face detection with the faster R-CNN. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition*, 2017.
- [9] Xiaofei Li, Yutong Ban, Laurent Girin, Xavier Alameda-Pineda, and Radu Horaud. Online localization and tracking of multiple moving speakers in reverberant environment. *IEEE Journal of Selected Topics in Signal Processing*, 2019.
- [10] V. Smidl and A. Quinn. *Variational Bayes Method in Signal Processing, The. Signals and Communication Technology*. Springer, 2006.
- [11] Johannes Traa and Paris Smaragdis. Multiple speaker tracking with the factorial von mises-fisher filter. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, 2014.