Signal Processing 129 (2016) 137–149

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Robust indoor speaker recognition in a network of audio and video sensors[☆]

Eleonora D'Arca[a,*], Neil M. Robertson[a], James R. Hopgood[b]

[a] Visionlab, ISSS, Heriot Watt University, Edinburgh EH14 4AS, UK
[b] University of Edinburgh, Edinburgh EH9 3JG, UK

ARTICLE INFO

Article history:
Received 4 October 2015
Received in revised form
26 April 2016
Accepted 28 April 2016
Available online 4 June 2016

Keywords:
Surveillance
Speaker diarisation
Security biometric
Audio–video speaker tracking
Multimodal fusion

ABSTRACT

Situational awareness is achieved naturally by the human senses of sight and hearing in combination. Automatic scene understanding aims at replicating this human ability using microphones and cameras in cooperation. In this paper, audio and video signals are fused and integrated at different levels of semantic abstractions. We detect and track a speaker who is relatively unconstrained, i.e., free to move indoors within an area larger than the comparable reported work, which is usually limited to round table meetings. The system is relatively simple: consisting of just 4 microphone pairs and a single camera. Results show that the overall multimodal tracker is more reliable than single modality systems, tolerating large occlusions and cross-talk. System evaluation is performed on both single and multi-modality tracking. The performance improvement given by the audio–video integration and fusion is quantified in terms of tracking precision and accuracy as well as speaker diarisation error rate and precision–recall (recognition). Improvements vs. the closest works are evaluated: 56% sound source localisation computational cost over an audio only system, 8% speaker diarisation error rate over an audio only speaker recognition unit and 36% on the precision–recall metric over an audio–video dominant speaker recognition method.

1. Introduction

The establishment of the digital era has created applications which combine audio and video to automate human activity analysis and understanding. We highlight the main areas of interest. First, for surveillance applications, i.e., detecting a person's biometric features to ensure that there are no intruders in a restricted area [1]. Second, understanding people social behaviour and interaction to determine their "role" and their intentions [2]. Third, detecting a possible threat in a public place [3] and, consequently, beam-forming and segmenting a dialogue [4]. Typical surveillance scenarios are characterised by the use of many wide area, distributed sensors covering unconstrained scenarios. Scene monitoring is often required to be real-time, thus computationally inexpensive algorithms are fundamental to the development of an effective system, but this is not always evident in the literature at present, and this is the challenge we address in this work.

1.1. Related work

The first step to full audio–video (AV) human activity analysis and understanding systems is detecting and tracking speakers through significant occlusions. State-of-the-art sound source localisation algorithms [5,6] are still computationally expensive, hence they are not suitable for "real-time" (or frame-rate) applications. Solving large video occlusions is still an inherently challenging research problem: many existing papers solve the problem by using advanced multi-camera 3-dimensional (3D) systems [7] which are prone to error when the camera views do not overlap. They are computationally expensive, requiring GPU/FPGA implementations to function at frame-rate when parallelisation is possible. Complementary use of audio and video is able to compensate for noisy, missing and erroneous data, reducing the number of sensors and the computational resources required at the expense of minimal effort in integrating or fusing signals [8–11,2,12–15,3,16–22].

Audio and video fusion can be achieved in several ways chiefly using variations of sampling techniques [8,14,15,19,21]. Existing AV person tracking system architectures work well only in highly sanitised, i.e., constrained and predictable scenarios: principally meeting rooms and diarisation [13,16,18,20] in which the person motion is either stationary, e.g., when people are talking seated

[☆] This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/J015180/1 and the MOD University Defence Research Collaboration in Signal Processing.

[*] Corresponding author.
E-mail addresses: eleonoradarca@hotmail.it (E. D'Arca), N.Robertson@qub.ac.uk (N.M. Robertson), James.Hopgood@ed.ac.uk (J.R. Hopgood).

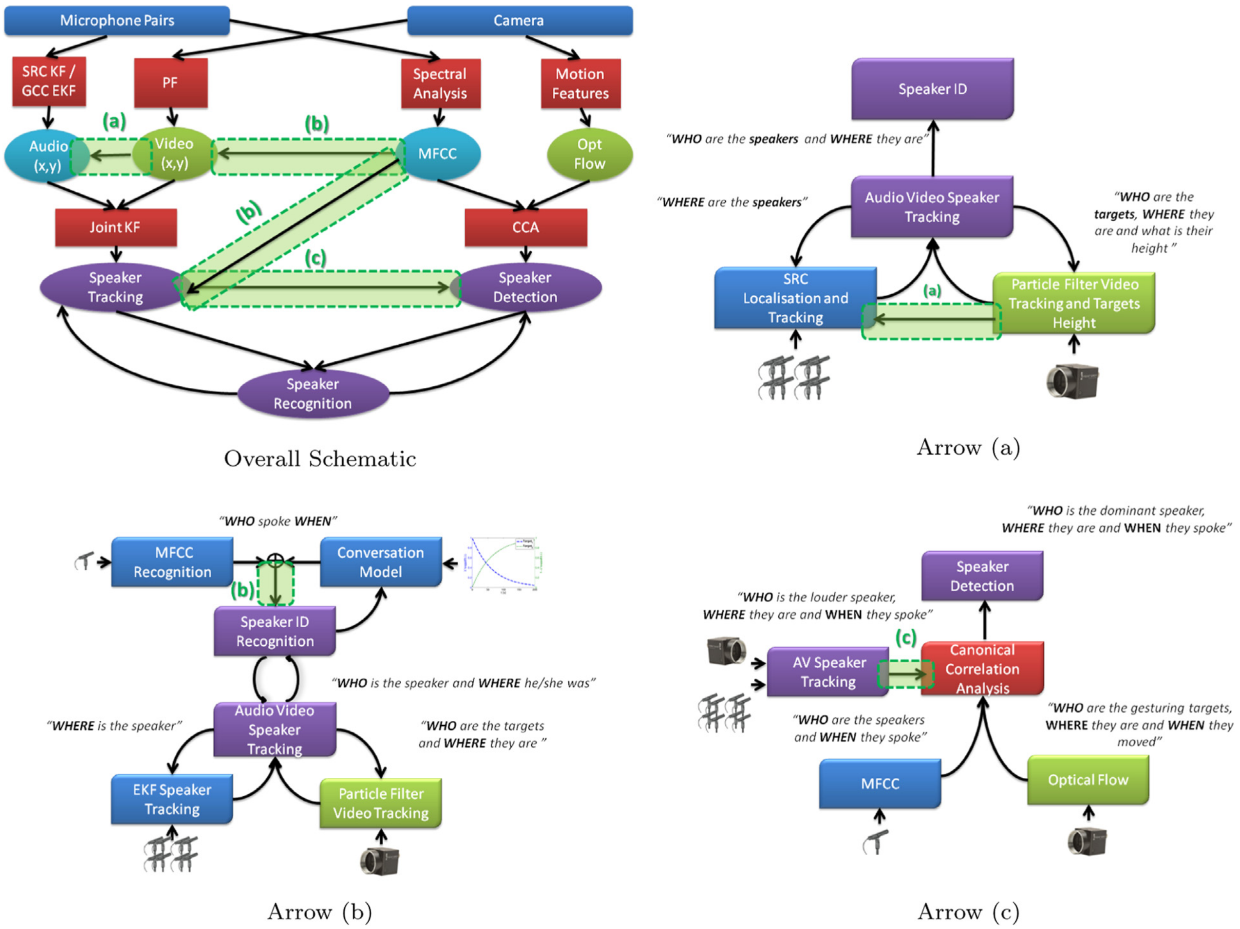
http://dx.doi.org/10.1016/j.sigpro.2016.04.014
0165-1684/© 2016 The Authors. Published by Elsevier B.V. All rights reserved.

around a fixed table. Existing systems use large sensor networks in which microphones and cameras are often very close or even *attached* to people [13–16]. A hierarchical system is more likely to achieve robust situational awareness. These are more robust, as accurate and require lower algorithmic and hardware complexity [2,16]. The weakness is that such systems often treat the two signals as if they were derived from truly independent processes: assuming one source of noise can affect only one kind of signal. None of the previous work explores whether an underlying relation between audio and video exists and seeks to exploit it fully.

AV event or anomaly detection literature is generally based on inferring AV signal correlations to recognise whether a relevant event has happened in some scenario of interest [3,11,12,17]. The same correlation approach may be used to pick out the dominant speaker from a group of speaking people, without audio beamforming, filtering, blind source separation and data association [23,24]. The definition of dominant speaker is clearly useful: a high degree of gesticulation and speaking activity are the fundamental cues to define dominance [25–28]. In fact, gesturing is 80–90% of the time associated to speaking activity [29]. Focussing on gesticulation detection is particularly suitable for low resolution video, where fine lip motion detection is not applicable and where close microphones may not be available.

To aid the reader, a schematic is shown in Fig. 1 and links to the different sections of the papers are explicitly made in the caption. Section 2 presents the integration of audio and video data at the signal level and their fusion at decision level for speaker detection and tracking (see [30]). A speaker voice recognition unit is implemented to make the multimodal tracking robust to occlusions (see also [31]). In Section 3, the experiments and the results related to the first part of the system are described. Here, the benefits of fusing multimodal data are highlighted remarking that standalone trackers have worse performances than the AV solution. Then, a possible solution to the problem of tracking the current speaker identity through occlusions by recognising speakers voices is demonstrated. Section 4 presents how to visualise in large indoor surveillance-like scenarios the dominant speaker identity when multiple people speak contemporaneously without resorting to sophisticated algorithms (see also [32]). Finally, in Section 5 the conclusions of this research study are highlighted and future avenues of research enumerated.

The exact contributions of this work relative to the published literature are: (a) definition of a new, high accuracy, fast audio source localisation algorithm augmented by video (stochastic region contraction with height estimation (SRC-HE)) which outperforms the baseline method stochastic region contraction (SRC) of Do et al. [6]; (b) extension of AV techniques for speaker tracking and event detection where people dynamically move and interact which outperforms the baseline method of Izadinia et al. [17]; (c) exploitation of a small sensor network, deploying only a single


Microphone Pairs
Camera
SRC KF / GCC EKF
PF
Spectral Analysis
Motion Features
(a)
(b)
Audio (x,y)
Video (x,y)
MFCC
Opt Flow
(b)
Joint KF
CCA
(c)
Speaker Tracking
Speaker Detection
Speaker Recognition
Speaker ID
"WHO are the speakers and WHERE they are"
"WHERE are the speakers"
Audio Video Speaker Tracking
"WHO are the targets, WHERE they are and what is their height"
SRC Localisation and Tracking
(a)
Particle Filter Video Tracking and Targets Height
"WHO spoke WHEN"
MFCC Recognition
Conversation Model
(b)
Speaker ID Recognition
"WHO is the speaker and WHERE he/she was"
"WHERE is the speaker"
Audio Video Speaker Tracking
"WHO are the targets and WHERE they are"
EKF Speaker Tracking
Particle Filter Video Tracking
"WHO is the dominant speaker, WHERE they are and WHEN they spoke"
Speaker Detection
"WHO is the louder speaker, WHERE they are and WHEN they spoke"
(c)
AV Speaker Tracking
Canonical Correlation Analysis
"WHO are the speakers and WHEN they spoke"
"WHO are the gesturing targets, WHERE they are and WHEN they moved"
MFCC
Optical Flow

Overall Schematic

Arrow (a)

Arrow (b)

Arrow (c)

Fig. 1. A detailed schematic diagram of the overall system presented in this paper. The schematic in (a) shows how the audio and video features cooperate at different levels of semantic abstraction. Block cooperations are represented by highlighted arrows which coincide with the novelties of this work. In (b) an audio localisation algorithm is cued by video data which becomes faster and not less accurate (see Section 2.1). In (c) it is shown how Mel-frequency cepstral coefficients (MFCC) voice signature recognition helps video ID tracking to be consistent through occlusions and ID swaps (see Section 2.7). In (d) the system describes how the correlation between optical flow associated with gesturing and sound signature of the scene helps the speaker ID recognition through speech interferences (see Section 4.3). Fundamentally, this system represents the combination of the detections of three "weak" classifiers into one robust process.