Image and Vision Computing 27 (2009) 1775–1787

Contents lists available at ScienceDirect

Image and Vision Computing

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

Automatic nonverbal analysis of social interaction in small groups: A review

Daniel Gatica-Perez

Idiap Research Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), Martigny, Switzerland

ARTICLE INFO

Article history:
Received 17 May 2008
Received in revised form 9 December 2008
Accepted 16 January 2009

Keywords:
Social interaction analysis
Small group conversations
Nonverbal behavior

ABSTRACT

An increasing awareness of the scientific and technological value of the automatic understanding of face-to-face social interaction has motivated in the past few years a surge of interest in the devising of computational techniques for conversational analysis. As an alternative to existing linguistic approaches for the automatic analysis of conversations, a relatively recent domain is using findings in social cognition, social psychology, and communication that have established the key role that nonverbal communication plays in the formation, maintenance, and evolution of a number of fundamental social constructs, which emerge from face-to-face interactions in time scales that range from short glimpses all the way to long-term encounters. Small group conversations are a specific case on which much of this work has been conducted. This paper reviews the existing literature on automatic analysis of small group conversations using nonverbal communication, and aims at bridging the current fragmentation of the work in this domain, currently split among half a dozen technical communities. The review is organized around the main themes studied in the literature and discusses, in a comparative fashion, about 100 works addressing problems related to the computational modeling of interaction management, internal states, personality traits, and social relationships in small group conversations, along with pointers to the relevant literature in social science. Some of the many open challenges and opportunities in this domain are also discussed.

1. Introduction

The automatic analysis of face-to-face conversational interaction from sensor data is a domain spanning research in audio, speech, and language processing, visual processing, multimodal processing, human–computer interaction, and ubiquitous computing. Face-to-face conversations represent a fundamental case of social interaction as they are ubiquitous and constitute by far – despite the increased use of computed-mediated communication tools – the most natural, enjoyable, and effective way to fulfill our social needs. More specifically, the computational analysis of group conversations has an enormous value on their own for several social sciences [8,92], and could open doors to a number of relevant applications that support interaction and communication, including self-assessment, training and educational tools, and systems to support group collaboration [37,101,53,103], through the automatic sensing, analysis, and interpretation of social behavior.

As documented by a significant amount of work in social psychology and cognition [8,92], groups both in professional and social settings proceed through diverse communication phases in the course of a conversation sharing information, engaging in discussions, making decisions, or dominating outcomes. Group conversations involve multiple participants effectively constrained by each other through complex conscious and unconscious social rules, and in the workplace they range from casual peer chatting to regular group discussions, formal meetings, and presentations; many other forms exist in the personal sphere.

While spoken language constitutes a very strong communication channel in group conversations [118], it is known that a wealth of information is conveyed nonverbally in parallel to the spoken words [80,89,93]. Nonverbal signals include features that are perceived aurally – through tone of voice and prosody – and visually – through body gestures and posture, eye gaze, and facial expressions [80,89]. Substantial work on social cognition regarding the mechanisms of nonverbal communication has suggested that, although some social cues are intentional (i.e., responding to specific motivations or goals), many others are the result of automatic processes [59]. Furthermore, it is known that people are also able to interpret social cues rapidly, correctly, and often automatically, accessing in this way information related to "the internal states, social identities, and relationships of those who make up our social world" [28] (p. 309), three social categories often used in social psychology and cognition. Experimental evidence shows that many of our social constructs and actions are in good part determined by the display and interpretation of nonverbal cues, in some cases without relying in speech understanding [59].

E-mail address: gatica@idiap.ch

0262-8856/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2009.01.004

This paper represents an attempt to draw a map of the existing work in the domain of automatic analysis of group interaction from nonverbal communicative cues, focusing on the small group setting. The main goal of the paper is to respond to the current fragmentation of this domain by gathering and briefly discussing works which, given its multi-faceted nature, have appeared in the literature spread over several communities, including speech and language processing, computer vision, multimodal processing, machine learning, human–computer interaction, and ubiquitous computing. As discussed in this review, initial progress has been made towards the detection, discovery, and recognition of patterns of multi-party *interaction management*, including turn-taking [30,91,27] and addressing [74]; group members' *internal states*, including interest and attraction [131,52,102]; individuals' *personality traits* including dominance and extroversion [11,111,106]; and *social relationships* in small groups including roles [133,128].

This review paper is focused on the discussion of computational models for the nonverbal analysis of physically collocated small groups (between three and six people). The definition of this concrete scope has several implications on the material chosen for discussion:

- *Focus on small groups*. It is well known that the size of a group has a definite influence in its dynamics, and that small groups tend to be more dynamic than large ones [49]. The small group case has produced an increasing body of work in this decade. With a few exceptions – which have been chosen as they have a clear relation to the small group case – the paper does not discuss cases of research in nonverbal modeling of dyadic conversations (e.g. [103]) or of large groups (e.g. [29]), which deserve a separate treatment.
- *Focus on nonverbal behavior*. The paper mainly discusses works that have targeted the modeling of nonverbal information (rather than speech and language) as their main goal. In a few cases, however, it will touch upon research that has relied on transcribed speech whenever this information was jointly used with nonverbal behavior.
- *Focus on computational models*. Rather than summarizing the well-established field of nonverbal communication, for which excellent textbooks have existed for years – as one notable example, the first edition of the popular book by Knapp, and later coauthored by Hall, dates from the early seventies [80] – the review aims at introducing, in a comparative fashion, a number of computational modeling works regarded as representative either by the addressed research problem or by the proposed solution, while providing up-to-date pointers to the literature (ca. Jan. 2009) to a non-expert reader. Whenever possible, pointers to social psychology and cognition literature are provided, which can be seen both as a motivating factor for some of the research described here and as a source of knowledge to support the design of computational models.
- *Focus on social constructs, not on cues*. This review focuses on the review of computational models for social constructs that can be identified with nonverbal behavior, rather than on the specific perceptual processing methods that can be used to extract such cues from audio and video, and which has spanned a considerable amount of research in audio processing (paralinguistics) and computer vision (face, gaze, body, and gesture analysis) over many years. The reader can refer to [129] for a recent attempt to recount a few of the existing cue extraction methods.
- *Focus on face-to-face conversations*. The paper reviews work on physically collocated groups that exclusively involve people, and therefore does not include as part of the discussion (with limited exceptions) the significant amount of work conducted in the Computer-Supported Collaborative Work (CSCW), Embodied Conversational Agents (ECA), and social robotics communities, which have also addressed group interaction from related but different perspectives and emphases.

The definition of the scope of the review according to the above criteria resulted in the body of technical work summarized in Fig. 1 (close to 100 papers published in journals, magazines, conferences, workshops, and other sources). Fig. 1(a) shows the distribution of this set of publications over time. The earliest references in this review date from 2001, a jump in the number of publications can be appreciated at 2003, and from then on a constant flow of new work has appeared in the literature. The work reviewed in 2009 is incomplete due to the date on which this paper was submitted for printing. Fig. 1(b) shows the distribution of publications per research field. It can be observed that roughly 36% of the reviewed papers have appeared in audio, speech, and language venues (labeled ASL in Fig. 1 and including TASLP, ICASSP, ICSLP, and LREC, among others), and 39% have appeared in multimedia and multimodal processing venues (labeled MM and including TMM, ACM

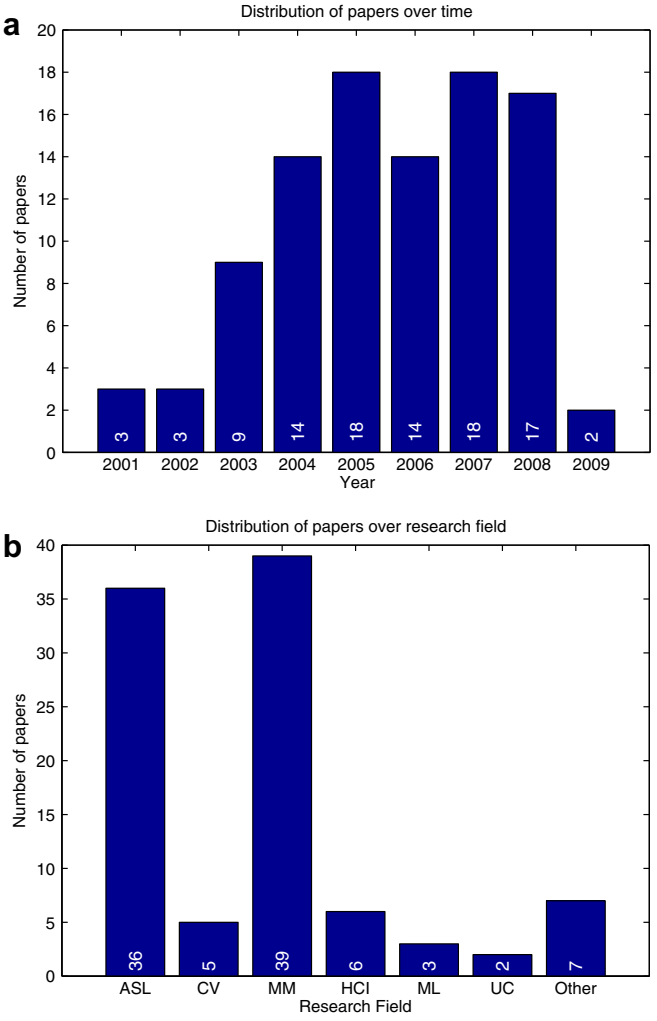

Fig. 1. Statistics of the 98 technical references on group interaction modeling reviewed in this paper. All papers were located in mainstream publication sources. The exact number for all bars is shown inside them. (a) Yearly number of publications. (b) Number of publications per research field (in journals, conferences, and workshops): audio, speech, and language (ASL); computer vision (CV); multimodal and multimedia processing (MM), human–computer interaction (HCI); machine learning and pattern recognition (ML); ubiquitous computing (UC); other (includes theses, technical reports, general computing magazines, and books).