Image and Vision Computing xxx (2009) xxx–xxx

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
Available online xxxx

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

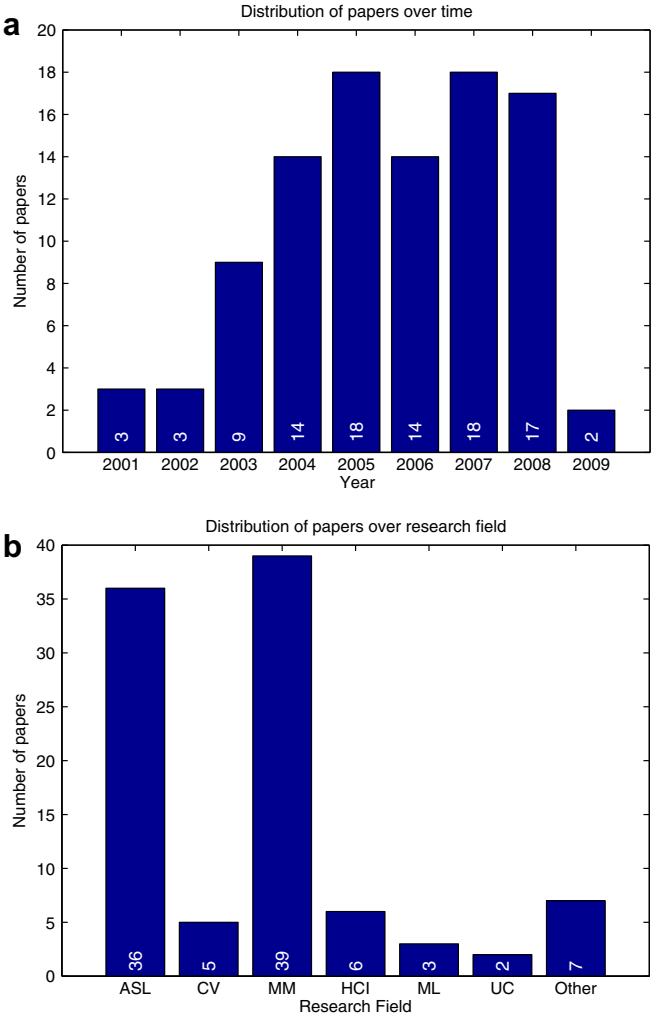

Fig. 1. Statistics of the 98 technical references on group interaction modeling reviewed in this paper. All papers were located in mainstream publication sources. The exact number for all bars is shown inside them. (a) Yearly number of publications. (b) Number of publications per research field (in journals, 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).

Please cite this article in press as: D. Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: A review, Image Vis. Comput. (2009), doi:10.1016/j.imavis.2009.01.004

MM, ICMI, MLMI, and ICME). Nevertheless, a sizable amount of work has been published in other domains in computing. Given the number of existing works and the perspective of their continuing increase in the future, a review of this subject seems timely. Table 1 provides a summary of the content of this review for rapid access to references of specific interest. As mentioned earlier, four categories of social constructs (interaction management, internal states, personality traits, and relationships) are discussed. Within each of them, specific cases for which there is work on computational models are discussed. The list of references for each case are listed chronologically.

A preliminary version of this paper, which to the author's knowledge constitutes the first survey of this domain, appeared in a conference version format in 2006 [54].

The remainder of the paper is organized as follows. Section 2 discusses work on computational models for interaction management. Section 3 reviews work on automatic modeling of internal states. Section 4 discusses the existing work towards the automatic modeling of personality traits. Section 5 discusses work on computational modeling of social relations. Section 6 summarizes the existing research infrastructure resources in this domain. Finally, Section 7 provides a final discussion and some concluding remarks about the challenges and opportunities that lie ahead in this domain.

2. Modeling interaction management

The analysis of the mechanisms to manage conversational interaction, which give rise to specific dynamics, is a fundamental area in social psychology and nonverbal communication [56,31]. Conversational patterns exist at multiple time scales, ranging from addressing (i.e., who speaks to whom), to a large variety of turn-taking patterns of longer temporal support including floor control mechanisms, discussion types, etc. In this section, we discuss two aspects that have been studied in the literature of nonverbal computational modeling, namely addressing and turn-taking behavior.

2.1. Addressing

In a conversation, an addressee is the person at whom the speech is directed [31]. In social psychology, it is known that the addressing phenomenon occurs through different communication channels, including speech, gaze, and gesture, e.g. listeners manifest attention by orienting their gaze to speakers, who in turn use gaze to indicate whom they address, and to ensure visual attention from addressees to hold the floor [56]. It is also known that participants in group conversations, interacting and exchanging roles as speakers, addressees, and side participants, contribute to the emergence of conversational events that characterize the flow of a meeting (for instance monologues, group discussions, or side conversations).

A good part of the body of work on automatic analysis of head pose as a surrogate for gaze and of visual focus of attention (VFOA) in group conversations [121,120,4,5] could be applied towards the automatic identification of addressees in multi-party cases. In brief, the goals of the existing works in addressing are to identify what participant(s) in a conversation the current speaker is talking to, and to explore the connections between addressee modeling and other conversational activities, like the ones described in Section 2.2. Katzenmeier et al. [76] presented a study about the identification addressees between two people and a simulated robot, with the further goal of discriminating human–human interaction from human-robot interaction. Three cases of perceptual integration were studied: audio-only using speech-derived features, visual-only based on head pose estimated from video, and audio-visual combining both types of cues. A Bayesian classification technique was used, in which neural networks were used to learn the head pose and audio representations. In this three-participant scenario, it was found that head pose is indeed a strong cue for addressee identification, and that the best performance was obtained with the multimodal approach, despite a relatively low performance obtained with the audio modality.

To the author's knowledge, the most comprehensive study on addressee estimation in small group conversations is the one by Jovanovic et al. [72–75]. In [72], the authors presented an initial scheme of verbal, nonverbal, and contextual features for addressee identification, but no experiments were conducted to validate the proposed scheme. In subsequent work, Jovanovic et al. [73] annotated with respect to addressee behavior a five-hour data set from the Augmented Multiparty Interaction (AMI) meeting corpus, a publicly available collection of four-person meetings with multi-microphone audio, multi-camera video, slides, whiteboard and handwritten notes [22] (see more about data corpora in Section 6). The data used in [73] was annotated with respect to discrete VFOA for each participant, addressee information, and dialogue acts (DA), that is speech utterances labeled as backchannels, floor grabbers, questions, and statements. The annotation of addressees used dialog acts as basic units, assigning one of four possible tags to each DA, to indicate whether the speaker addresses a single person, a subgroup, the whole audience, or if the addressee is unknown. The detailed discussion about the reliability of the manual annotation process in [73] indicates that the annotation ranges in quality from acceptable to good for those DAs whose boundaries are agreed upon by manual annotators; that the reliability is higher on those meeting segments where the speaker addresses a single person; and that annotators had problems distinguishing between subgroup and group addressing. All these findings were useful to assess the type of performance that could be expected with automatic processing, a task that was addressed in [75] using Bayesian Networks and a combination of automatic and manual cues, and that highlighted the difficulty of the task. The best reported performance was around 75% classification accuracy. A recent description of advances by the same group can be found in [2].

Takemae et al. also studied the addressing problem in small groups [124], using a small data set of three- to five-person meetings, manually annotated gaze and close-talk microphones for each group member. This work also analyzed separately the single-person and multi-person addressee cases, and reported 89% accuracy in classifying these two addressing classes, using speech utterances as data units, and basic features derived from each speaker's gaze patterns. In work closely related to the addressing task, Otsuka

Table 1
A summary of the content of this review. Four main categories of social constructs are discussed. Within each of them, specific attributes are studied. The list of references on computational models discussed for each case are listed chronologically.

Social category	Construct	Technical references
Interaction management	Addressing	[121,120,4,76,124,72,73,96] [5,74,98,75,99,61,2,6,100]
	Turn-taking	[90,35,134,135,10,91,3,26] [136,15,19,137,110,27,36,20]
Internal states	Interest	[131,132,77,44,62,52,102,86] [87,122,34,103]
	Other states	[78,125,109,60,126,82,83,104]
Personality traits	Dominance	[11,37,111,81,112,97,123,113] [63,13,64–66,69,7,79,70]
	Extroversion	[88,106]
	Locus-of-control	[88,106]
Relationships	Roles	[10,133,105,88,38,128,68,48] [50,108]

et al. [96] proposed a Dynamic Bayesian Network (DBN) approach which jointly infers the gaze pattern for multiple participants and the conversational gaze regime responsible for specific speaking activity and gaze patterns (e.g. all participants converging onto one person, or two people looking at each other). Gaze was approximated by head pose. The work relied on clean head pose and speaking activity extracted from magnetic head trackers attached to each participant and manual speaking-turn segmentations, respectively. The same model was later used with head pose estimated automatically from video [98]. In more recent work [99], Otsuka et al. extended their DBN model in an attempt to respond to the questions of who responds to whom, when, and how in a joint manner. The employed cues were head pose extracted from magnetic sensors, automatic speaker segmentations derived from lapel microphone signals, and an elementary form of binary head gestures automatically recognized. Finally, Otsuka et al. developed an automatic system that integrates head pose tracking and speaker diarization that re-uses several of the ideas from previous works for real-time usage [100].

In other related research, an interesting interplay between speaking activity and visual attention (two of the nonverbal components of the group addressing phenomenon) was recently explored by Ba and Odobez with a different purpose [6]. In this work, the authors proposed a DBN model for the estimation of the joint focus of attention of group participants by using people's speaking activity as a contextual cue, defining a prior distribution on the potential VFOA of each participant. This observation resulted in improved VFOA recognition from head orientation automatically estimated from a single camera on a subset of the AMI corpus.

Another conversational construct clearly related to addressing and attention is listening. Listening is in principle a multimodal phenomenon, and some recent works have started to investigate computational models that could be useful both to improve visual attention estimation and addressing, but also for synthesis purposes in embodied agents. Heylen et al. [61] presented an overview of initial work towards building a Sensitive Artificial Listener. This includes the manual annotation of basic nonverbal behaviors displayed by listeners in multi-party conversations (including gaze patterns, head movements, and facial expressions extracted from the AMI corpus), and their use as part of a manual or semi-automatic model to animate embodied agents that can have different personalities (e.g. optimistic and active vs. negative and passive). Needless to say, this line of work faces the difficulty of automatically extracting (possibly) subtle patterns related to eye gaze and facial expressions, which are open problems in computer vision.

2.2. Turn-taking behavior

A group in a conversation can be seen as proceeding through diverse communication phases. Assuming that the presence of a group in a multi-sensor room is either known or inferred [15], a sequential model for a group conversation could assume a discrete set of group conversational activities and view a conversation as a sequence of such activities. In a formal meeting scenario, where people discuss around a table and use artifacts including a whiteboard and a projector screen, McCowan et al. [90,91] first investigated this approach and targeted the joint segmentation and recognition of meetings into a number of location-sensitive turn-taking patterns, including monologues, discussions, and presentations. The approach relied on supervised learning techniques, namely Hidden Markov Models (HMMs) [107], and used a number of simple audio and visual nonverbal features automatically derived from multiple cameras and microphones. The audio cues were derived from lapel microphones (pitch, energy, speaking rate) and from microphone arrays (speech activity estimated by the steered power response phase transform (SRP-PHAT)). The visual cues were extracted from each participant's skin-color blobs motion and location, as a very rough proxy for head and body motion and pose. Using the MultiModal Meeting Manager (M4) corpus, the problem was studied as a multistream system, where data streams can correspond either to the features extracted from each person or to each perceptual modality (audio or visual). A number of HMM variations were tested, including multistream HMMs [43], coupled HMMs [14], and asynchronous HMMs [12]. The results, measured in terms of Action Recognition Rate, were encouraging and showed the benefits of audio-visual fusion. The approach, however, has two limitations. First, HMMs can be challenged by a large number of parameters, and by the risk of overfitting when learned from limited data [95]. This situation might occur in the case of turn-taking recognition where, in the simplest fusion scenario, large vectors of audio-visual features from each participant are concatenated to define the observation space. Second, the framework does not explicitly model other patterns at different semantic levels, despite the known fact that models in social psychology describe group meetings as comprising both individual and group-level activities [92].

Zhang et al. [134,137] addressed the above limitations with a two-layer HMM framework [95]. In the first layer, individual activities performed by each person, such as *writing* and *speaking*, are recognized from raw audio-visual observations (i.e., estimating the posterior probabilities of the individual activities). Then, the second layer recognizes the group turn-taking patterns using as input the results of the low-layer recognizers for all meeting participants and a set of additional features, directly extracted from the raw streams and not associated to any person. Compared with single-layer HMMs, layered HMMs have several advantages including the use of much smaller observation spaces; the fact that the low-layer HMMs can in practice be better estimated as much more data (arising from multiple people) is available; a reduction in sensitivity for group activity recognition as the observations for the high-layer are posterior-based features; and the possibility of exploring different HMM options for each layer. The experiments in [134,137] led to three findings. First, the two-layer HMM approach outperformed the single-layer one. Second, the use of audio-visual features outperformed the use of single modalities for both single-layer and two-layer HMMs, supporting the hypothesis that the target interaction patterns are inherently multimodal. Third, the best low-layer model was the asynchronous HMM (a model that explicitly accounts for variations of alignment between two data streams), which suggested that, while some asynchrony might be present, it was reasonably captured by the model. A best performance of 85% action recognition rate was obtained on the M4 corpus.

Other works have also targeted the same task and data with hierarchical representations. Dielmann and Renals [35,36] proposed two approaches using audio-only cues and multilevel Dynamic Bayesian Networks (DBNs). The first DBN decomposed the interaction patterns as sequences of sub-activities with no explicit meaning. The second DBN processed features of different nature independently, and integrated them at a higher level. In this work, the sub-activities have no obvious interpretation, and their number is a model parameter learned during training or set by hand, which makes the structure of the models more difficult to interpret. An initial comparison of various recognition models on the same task, including the layered HMM, the multilevel DBN, and other approaches, was presented by Al-Hames et al. [3]. More recently, an approach by Reiter et al. [110] exploited the flexibility of the layered architecture to go beyond the use of HMMs, more specifically using a discriminative approach (neural networks) in one of the layers, showing minor performance improvement.

Please cite this article in press as: D. Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: A review, Image Vis. Comput. (2009), doi:10.1016/j.imavis.2009.01.004

Two variations of the same problem were also explored by Zhang et al. [135,136]. These approaches take a practical perspective: the manual labeling of group turn-taking patterns for training purposes is both difficult and expensive. The use of unsupervised or partially supervised approaches could thus be attractive alternatives. The approach in [135] proposed a two-layer framework where the low-layer is identical to the one presented in [134], and the high-layer is a fully unsupervised HMM that discovers (rather than recognizing) group activity patterns. The method in [136] uses model adaptation techniques, where instead of directly training one model for each group pattern (as done in [134]), a general model is first estimated using unlabeled data, and then adapted to each individual pattern using its own labeled data using Bayesian adaptation. Both methods define a tradeoff between performance and availability of labeled data. The investigation of models that rely on unsupervised or lightly supervised approaches remains as a research area of practical relevance, given the increasing availability of unlabeled data and the annotation costs required by supervised methods.

In work with essentially similar objectives to [90,91], Banerjee and Rudnicky [10] proposed a simple method to recognize three classes of meeting activities, including discussion, presentation, and briefing, in groups meetings recorded with close-talk microphones. In the method, three inexpensive features are computed over a running time window for each participant, using speaker segmentations as input. The features include the number of speaker changes, the number of speakers, the number of overlapping turns, and the average length of the overlaps within the running window. All features were extracted from manual speaker segmentations. A C4.5 decision tree was used to classify time windows at one-second time steps. Experiments on 45 min of the data corpus reported in [9] resulted in a best classification accuracy of 51%, which suggested that relatively simple cues and standard supervised learning approaches could perform better than random, but certainly far from very high performance.

Campbell and collaborators [19,20], building on his long-term work on prosodic analysis [18] have also explored the problem of nonverbal determination of participation status of individuals in group conversations. Concretely, the work by Campbell and Douxchamps [20] extracted nonverbal visual cues derived from standard face detection and motion analysis techniques, using a single video signal captured with a parabolic camera on a table. Furthermore, sound cues were derived from a single microphone and included speech utterances which, depending on their duration, were labeled as being either opinion or backchannel utterances. The authors first performed an analysis of overlapping speech and back-channeling for three different group conversation types (formal meeting, relaxed conversation, party situation) and found clear differences in the amounts of overlapping. They also reported preliminary experiments on audio-visual cue fusion for recognition of conversational activity labels.

Finally, there has been some work on modeling of floor control and turn-taking transitions in group conversations. A speaker in charge of ensuring the progress of a conversation is said to have control of the floor. As a conversation progresses, the active speaker holds the floor, while other speakers participate either cooperating or competing to share the floor. Floor control is a phenomenon studied in psychology and linguistics for many years [42], and has been observed that multimodal verbal and nonverbal cues (including gaze exchanges between the floor holder and the interlocutors, and discourse markers) are related to floor control changes. Regarding computational models, Chen et al. [26] described initial efforts to combine gaze, gesture, and speech for floor control modeling, using the Video Analysis and Content Extraction (VACE) meeting data collected with multiple cameras, microphones, and magnetic sensors. This study was later extended in [27]. The work includes the definition of a scheme for floor control annotation, and the use of a small labeled corpus to identify multimodal cues correlated with floor changes. The analysis of the corpus suggests that discourse markers occur frequently at the beginning of a floor; that mutual gaze between the current floor holder and the next one offer occurs during floor transitions; and that gestures related to floor capturing also occur. However, no attempt to perform automatic processing was reported.

3. Modeling internal states

The number of a person's internal states that can be revealed by conscious or unconscious nonverbal behavior is large [80,28]. In small group interaction, a list by no means exhaustive could include nervousness, anxiety, embarrassment, frustration, anger, joy, attraction, confusion, consistency, boredom, and interest. While the work on computational modeling of internal states in HCI is large and has explored single and multimodal approaches [33], there has been relatively little work in the specific context of small face-to-face group interaction. In this review, we focus on one construct that is particularly relevant in multi-party conversations, namely interest, and also touch upon other cases currently being studied.

3.1. Interest

The term "interest" is used here to designate people's internal states related to the degree of engagement that individuals display, consciously or not, during their interaction. Such displayed engagement can be the result of many factors, ranging from interest in a conversation, attraction to the interlocutor(s), and social rapport. Social displays of interest through nonverbal cues have been widely studied in social psychology and include mimicry [24,25] (a complex phenomenon displayed through sound cues but also via body postures and mannerisms, and facial expressions), elevated displays of voice and kinesic activity, and higher conversational dynamics.

In the specific context of group interaction, the degree of interest that the members of a group collectively display during their interaction is an important state to extract from formal meetings and other conversational settings. Segments of conversations where participants are highly engaged (e.g. in a discussion) are likely to be of interest to other observers too. Interest level can therefore define a form of relevance around which conversations could potentially be summarized or indexed. The computational modeling of interest has started to be explored in group conversational settings from a number of different perspectives [131,132,77,44,52,102].

Most existing work has focused on the analysis of the relation between interest and the speech modality, using both verbal and nonverbal cues. Wrede and Shriberg [131,132] introduced the notion of hot-spots in group meetings, defining it in terms of participants highly involved in a discussion, and relating it to the concept of activation in emotion modeling, i.e., "the strength of a person's disposition to take action" [33]. Using data from the International Computer Science Institute (ICSI) meeting corpus [67] containing four- to eight-person conversations, close-talk microphones, and speech utterances as the basic units, the authors first developed an annotation scheme that included a category of non-involvement and three categories of involvement (amused, disagreeing, other), and then found that human annotators could reliably perceive involvement at the utterance level, i.e., whether a speaker sounded "especially interested, surprised or enthusiastic about what is being said, or he or she could express strong disagreement, amusement, or stress" [132]. They also found that a very small proportion of utterances (about 2%) corresponded to involved utterances

Please cite this article in press as: D. Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: A review, Image Vis. Comput. (2009), doi:10.1016/j.imavis.2009.01.004

[132], and that certain nonverbal prosodic features, including energy and pitch, were correlated with involved utterances. Wrede and Shriberg extended their work to study the relation between hot-spots and dialog acts, using both contextual features (e.g. speaker identity and meeting type) and lexical features (e.g. utterance length and perplexity) [132].

In a related line of work, Kennedy and Ellis [77] defined emphasis for speech utterances in meetings, acknowledging that this concept and emotional involvement might be acoustically and perceptually similar. A very simple approach was used to measure emphasis, using normalized pitch as the only cue, and a performance of 92% accuracy was reported for utterances that were consistently annotated by people as having high-emphasis. Other works in the speech processing community can be also be related to the detection of certain types of high-interest segments. As one example, Hillard et al. [62] proposed a method to recognize a specific kind of interaction in meetings (agreement vs. disagreement) that is likely related to high interest. The work used both word-based features (e.g. total number of words, and the number of "positive" and "negative" keywords), as well as prosodic cues in a machine learning approach that made use of unlabeled data.

A number of works in the wearable computing community have also dealt with the estimation of interest and related quantities, notably the work by Pentland and collaborators [44,102,86,87,122,103]. Much of this work has been conducted for dyadic cases, but some group conversational cases have also been studied. The two-person cases have included prediction of interest-level in 3-min conversations between same-gender people about random topics [102]; prediction of attraction (that is, romantic or friendly interest) between different-gender strangers in 5-min speed-dating encounters [86,87,122]; and prediction of the outcome of dyadic salary package negotiations using the first five minutes of the conversation [34], a situation clearly related to constructs of rapport and interest. The multi-party scenarios have included small group conversations [44] and brief conversational exchanges among large groups of attendees of a conference [102]. A very interesting feature of this line of work is that it has often been grounded in social situations with concrete behavioral outcomes – people declaring common attraction in a speed dating situation, or people exchanging business cards at a conference as a sign of mutual interest – which substantially reduces the need for third-party annotation of interest. Regarding computational models, the estimation of interest has ranged from introducing it manually as in the work by Eagle and Pentland [44] to computing it automatically from a number of acoustic nonverbal cues (so-called activity, engagement, stress, and mirroring signals), and in some cases complemented by body motion estimated from accelerometers [86,87,122]. These cues, in different combinations, and used as input to standard classifiers like linear classifiers or Support Vector Machines (SVM) have resulted in promising performance (74% accuracy for the random-topic conversations, 70–80% accuracy for the speed date cases, and 80–85% accuracy for the conference attendee case).

To the author's knowledge, very few works have studied the use of audio-visual cues for interest estimation in group conversations. In [52], Gatica-Perez et al. presented an investigation of the performance of audio-visual cues on discriminating high vs. neutral group interest-level segments, i.e., on estimating single labels for meeting segments, much like hot-spots, using a supervised learning approach that simultaneously produces a temporal segmentation of the meeting and the binary classification of the segments into high or neutral interest-level classes. Experiments were conducted on the four-person conversations from the M4 corpus. Two classic HMM recognition strategies were investigated: early integration, where all cues are synchronized and concatenated to form the observation vectors; and multistream HMMs used for audio-visual fusion. The investigated cues were the same as the ones used in [91]. Various combinations of voice, kinesic, and multimodal cues and HMM models were investigated. The results were promising, and while the audio modality turned out to be dominant, audio-visual fusion could improve performance. Furthermore, the investigation of visual cues that are better correlated with displays of interest remained as an open issue.

3.2. Other internal states

There has been work towards finding other internal states in group conversations. Using the AMI corpus, Reidsma et al. [109,60] initially investigated procedures to manually annotate emotional states in meetings. The authors found that the problem is very complex, and the use of existing emotion annotation schemes that have worked in other scenarios (FeelTrace) did not provide good results in the meeting case. As a more viable alternative, the authors developed a procedure based on marking distinctive changes in the "state of mind" of a person being observed, and then annotating the state and its intensity and quality. The list of state labels, which are likely to occur in group conversations, included "neutral", "curious", "amused", "distracted", "frustrated" and "confused", among others, and often do not correspond to prototypical emotion categories but rather to cognitive and social states. In experiments with subjects, the authors found the quality of the resulting annotation (measured by inter-annotator agreement techniques) to be comparable to that of other reported studies on emotion annotation, and concluded that further work was necessary to improve the annotation scheme.

One particular phenomenon that has studied in group conversations is laughter, which is a behavior that might correspond to several internal states. This case is interesting because in principle it occurs frequently enough in group interaction, and might be amenable for reliable annotation for training and evaluation purposes. Existing approaches investigating detection and classification of laughter segments include [78,125,126,82,83,104]. Most work has studied the problem using only audio cues, and often using the ICSI meeting corpus. With the availability of audio-visual corpora of meetings, the study of multimodal approaches for laugther detection starts to be feasible. An example of initial work in this direction is the work by Petridis and Pantic [104], based on a small data set from the AMI corpus and on joint modeling of acoustic and facial features.

4. Modeling personality

While nonverbal behavior and personality have important connections [28], there is also strong evidence that the problem of linking both consistently is a challenging problem, due to a number of complex factors [55]. The work on computational models of personality traits in small groups has begun to explore a few fundamental cases. In this section, the existing work in three dimensions of personality traits, namely dominance, extroversion, and locus-of-control, is discussed.

4.1. Dominance

Dominance is a key concept in social interaction and has been well studied in social psychology [46,16], as one component of the so-called vertical dimension of social relationships [57]. Dominance is often seen in two ways, both "as a personality characteristic" (a trait) and "to indicate a person's hierarchical position within a group" (a state) [116] (p. 421). Although dominance and related terms like power have multiple definitions and are often used as equivalent, a distinguishing approach taken in [40] defines power as "the capacity to produce intended effects, and in partic-

Please cite this article in press as: D. Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: A review, Image Vis. Comput. (2009), doi:10.1016/j.imavis.2009.01.004

ular, the ability to influence the behavior of another person" (p. 208), and dominance as a set of "expressive, relationally based communicative acts by which power is exerted and influence achieved", "one behavioral manifestation of the relational construct of power", and "necessarily manifest" (pp. 208–209). For the development of computational approaches, two key findings from social psychology are the use of specific nonverbal cues to often display dominance in conversations, and the ability to correctly interpret such cues by interaction partners and external observers. The fact that people can often correctly perceive dominance is fundamental towards generating reliable human annotations and developing computational models.

Nonverbal displays of dominance involve sound and motion [40]. The first type includes amount of speaking time [116], speech energy, pitch, rate [127], vocal control [40], and interruptions [85]. Among these, speaking time has shown to be a particularly robust cue to perceive dominance, as dominant people tend to talk more [116]. Kinesic cues include body movement, posture, and elevation, gestures, facial expressions, and eye gaze [40]. Dominant people are often more active, and gestures associated with speech are correlated with dominance [41,16]. It has also been found that in dyadic conversations high-status persons receive more visual attention than low-status people [45], and that people who rarely look at others are perceived as weak [32]. Furthermore, there is evidence of dominance-related patterns of joint visual attention and speaking activity in dyadic exchanges via the visual dominance ratio [47,39], in which high-status people often display a higher looking-while-speaking to looking-while-listening ratio (the proportion between the time they gaze at the other while talking and the time they gaze at the other while listening) [47].

Although studies on nonverbal display and interpretation of dominance have existed for decades, the problem of automatically estimating dominance-related measures has begun relatively recently. All the works discussed below studied small groups recorded with multiple cameras and microphones, and with the exception of [11], they all analyzed four-people conversations.

Basu et al. [11] described an approach to estimate the most influential participant in a debate. The influence model (IM), a DBN which models the members of a group as a set of Markov chains influencing the state transitions of one another, was applied to automatically determine the degree of influence a person has on the others on a pairwise basis. Cues related to speaking activity (manually labeled speaker turns and automatically extracted speaker energy and voicing information) and visual activity (region-based motion derived from skin-color blobs) were used. Although the IM (and other related models, like the one proposed by Choudhury and Basu [30]) is a tractable alternative to model group interactions, it only models pairwise interactions between individual players, not explicitly modeling the group as a whole.

Rienks and Heylen [111] proposed a supervised learning approach based on SVMs, addressing a three-class classification task in which meeting participants were labeled as having high, normal, or low dominance. A number of manually produced audio-only cues, both nonverbal (speaking time, number of speaker turns, number of successful floor grabbing attempts) and verbal (number of spoken words) were used. A best performance of 75% classification accuracy was reported on a data set containing meetings from the M4 and AMI corpora. Later, Rienks et al. [112] compared the approaches from [111] and a variation of the influence model for the same three-class dominance-level task, on a data set from the AMI corpus larger than the one used in [111] but with similar audio features. The SVM approach showed to outperformed the influence model, reporting a best performance of 70% classification accuracy. An analysis of participant influence using the same data as in [112] and *verbal* information, including manually annotations of dialog acts and argumentation categories, was conducted in [113]. The authors reported that using argumentation did not succeed at predicting influence better than a naive assumption that assigned the most common class to all test data points.

Following the ideas of [11], Otsuka et al. proposed in [97] the use of the output of the DBN model discussed in [96] and Section 2.2 to estimate pairwise speaker influence. A number of influence-related simple features (called incoming/outgoing influence, interactivity score, and centralization) is estimated from the conversational regimes and gaze patterns inferred by the DBN, and use to characterize each person. This work is one of the first ones to use the fact that gaze and speaking activity are both related to dominance perception, as suggested in social psychology [45,39]. On the other hand, while the features proposed in [97] are conceptually appealing, this work presented neither an objective performance evaluation nor a comparison to previous methods.

Although not necessarily addressing automatic dominance estimation, some recent approaches [37,81,123,13,7,79] have been developed to actively influence the dynamics of a conversation by providing on-line feedback to the group members, through the estimation of nonverbal cues known to be correlated to dominant behavior. Dominant people might overcontrol the floor and negatively affect a conversation where the ideas of others might be important but overlooked, for instance in a brainstorming meeting or as part of a learning process. It has been documented in social psychology that people who hold the floor too much are perceived as overcontrolling [21].

DiMicco et al. [37] proposed an approach that estimates the speaking time of each participant on-line from headset microphones and visualizes this information publicly, finding that such type of feedback tends to promote a more even participation of the group members. Sturm et al. [123] built on a previous work by the same authors [81] to develop a system that automatically estimates speaking time from headset microphones and focus of attention from headbands with reflective pieces tracked by infrared cameras. Their system builds on the assumption that speaking time and visual attention time are strong cues to decide who controls the conversation, and so visualizes these cumulative cues on the conversation table in real-time. Rather than automatically predicting the dominant people, the system aims at regulating the flow of the conversation in a way that facilitates individual participation. Bachour et al. [7] explored a similar idea, employing the meeting table as both a sensor and a display, estimating each person's speaking activity with a microphone array implemented on the table, and displaying the proportion of speaking time via LEDs placed under the table. Kim et al. [79] opted for a portable solution for both sensing and displaying of group interaction, in which participants wear a badge that extracts speaking time, prosody, and motion, and displays measures of the interaction on a cell phone. Overall, while the initial evaluation of these prototypes has been promising in terms of their individual and social acceptability and their ability to change behavior in the intended ways (e.g. improving participation from all the group members), the usability of such tools to consistently improve group communication in a variety of social situations is still an open question.

Recent work by the author's research group [63–66,68–70] has aimed at studying dominance in small groups in a systematic way, examining the effects of specific nonverbal cues, dominance estimation models, and sensor settings, as well as the variability of the human perception of dominance, all with a common data set and using fully automatic nonverbal cues. Hung et al. [63] addressed the task of estimating the most-dominant person in a group using automatically extracted voice cues (speaking time, energy) from headset microphones, and kinesic cues (coarse visual activity measures) computed from compressed-domain video recorded by close-up view cameras. A simple yet effective approach assumed that higher activity corresponded to higher dominance.

Please cite this article in press as: D. Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: A review, Image Vis. Comput. (2009), doi:10.1016/j.imavis.2009.01.004

Speaking time proved to be the strongest single cue, providing a classification accuracy of 85% over 5 h of the AMI corpus divided in 5-min meeting segments. A more thorough analysis was conducted by Jayagopi et al. [70]. The study included an improved set of nonverbal activity cues, an additional SVM-based approach, and two classification tasks (most-dominant person and least-dominant person) divided into two conditions, each of which evaluated data with a different degree of variability with respect to human perception of dominance. The results suggested that, while audio is the most informative modality, visual activity also carried some discriminative power (e.g. best performance of 79% classification accuracy for the most-dominant task), and also that nonverbal cue fusion in the supervised setting was beneficial in some cases (e.g. best performance of 91% accuracy for the most-dominant task). Furthermore, more challenging data in terms of higher variability of dominance judgment by people did translate into a consistent decrease of performance for the automatic methods. Using the same data, Hung et al. [66] investigated the automation of the visual dominance ratio studied in social psychology [47], extending it to the multi-party case, revisiting the "looking-while-speaking" definition to include all people whom a person looks at when she/he talks, and the "looking-while-listening" case to include all cases when a person does not talk and looks at any speaker. Using visual attention automatically estimated from monocular video [6], and speaker turns derived from close-talk microphones, the results for estimating the most dominant person showed that the visual dominance ratio outperformed both its individual components and the total amount of received attention, but also that despite this good performance, certain audio-only cues were still the most discriminant ones. In a different research line, Jayagopi et al. [69] applied the same methodology to the task of classifying the dominant clique (i.e, the subgroup of people who are most dominant) in a conversation, achieving similar performance levels (best obtained accuracy of 90%), and observing similar trends regarding the discrimination of single and fused cues. Finally, one key issue for moving towards real-life situations is robustness with respect to the sensor setting. The works discussed in this section rely on good-quality audio signals extracted from multiple close-talk microphones, one per speaker. A more desirable setting would involve non-obtrusive, possibly single, distant microphones. The case of a single microphone would involve the need for speaker diarization. This is the problem of segmenting an audio track from a single audio channel, producing a set of speech segments and cluster labels for each of the segments in an unsupervised way (i.e., without information about the number or identity of the speakers) [1]. Hung et al. [64] studied the problem of estimating the most-dominant person for the single distant microphone case, using a fast speaker diarization algorithm and speaking time as only cue. Not surprisingly, the results showed a decrease of performance in the estimation of the most-dominant person compared to the close-talk microphone signals, given the challenge of accurately segmenting speaker turns from a single audio source. An important problem when using a single audio channel and diarization is the lack of direct ways of associating people identities with the speaker clusters produced by diarization. One way of addressing this problem, in the context of multi-sensor spaces, is by exploiting the correlation between visual activity and speaking activity. This has been studied by Hung et al. in [65], using some of the nonverbal cues developed in [70] and combining it with the framework developed in [64].

4.2. Extroversion and locus-of-control

Recent work by Pianesi et al. [106] has addressed the recognition of two different types of personality traits in small groups, using the Mission Survival 2 Corpus (MSC-2), a corpus of 13 meetings (about 7 h of data) recorded with cameras and microphones in which four-person groups discussed and agreed on an itemized strategy for survival in a disaster scenario [88]. The first studied trait is extroversion, one of the components of the Big Five model in social psychology [71], which posits that a person's personality is constituted by five general traits, namely extroversion vs. introversion, neuroticism vs. emotional stability, openness vs. un-openness, agreeableness vs. disagreeableness, and conscientiousness vs. un-conscientiousness. The second personality trait is locus-of-control [114], which quantifies if an individual's behavior is assumed to be dependent on his/her own actions (internal orientation) or on external factors outside the person's control (external orientation). For each of the above two traits, Pianesi et al. defined a three-class classification task, using people's self-reports based on standard assessment procedures as ground truth, to classify a person's trait as being "high", "middle", or "low". The authors used a relatively large number of acoustic features following the work by Pentland [103] and Stoltzman [122], and a small number of visual cues related to fidgeting, and investigated two questions: whether social context (i.e., classifying the observations of a person using also the observations of the other group members) was useful; and whether a feature selection method applied on a person's cues was beneficial for classification. Using one-minute time slices as units of processing, the authors found that the first hypothesis seemed to hold, while the second one did not. For both extroversion and locus-of-control, the best achieved performance for classification accuracy was 94%, which is clearly promising.

5. Modeling social relations

5.1. Roles

Much research has been conducted in social psychology and sociology on the subject of roles in small groups for the past 70 years [58,115]. As proposed by Hare, a primary definition of a role "is that is associated with a position in a group (or status) with rights and duties to one or more group members [...] for formal group roles that members perform consciously" [58] (p. 434). At the same time, informal roles can emerge and be played during group interaction. Many role categorization systems have been proposed in the literature, each from a different perspective. For instance, following [58], Hare distinguishes between functional roles based on forms of differentiation that exist or emerge among group members, and that include functions like control of others, access to resources, status, and identification with the group; sociometric roles, where people can occupy central, friendly, or isolated roles based on the analysis of the group's network; emotional roles, which involve roles that are not consciously acted or realized by a group, and can include prototypical roles such as hero, clown, or scapegoat; and dramaturgical roles, which involve traditional roles played in social drama, including protagonists, antagonists, or audience members. In addition to the variety of role categorization systems, it is also known from social science that people's roles in group interaction structure nonverbal behavior in important ways [84,57]. As an example, it is known that people that play high-status roles in groups are often more vocally and kinesically expressive than their counterparts, and that often receive more visual attention [57].

Given the multiplicity in perspectives (and associated definitions) of roles in small group interaction, the existing work on computational models for role recognition has essentially examined disparate cases [10,133,38,128,68,48,50,108]. The rest of this section examines each of these approaches.

As part of the work by Banerjee et al. [10] discussed in Section 2.2, the authors also used a decision tree to recognize a number of intuitively useful roles, including discussion participator, presenter, information provider, and information consumer. The

Please cite this article in press as: D. Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: A review, Image Vis. Comput. (2009), doi:10.1016/j.imavis.2009.01.004

method used six features computed over a sliding time window for each participant, four of which are the same ones used for meeting phase recognition in Section 2.2, complemented by the amount of speaking time and the amount of overlapped speaking time with other participants within the running window. All features were manually extracted. Experiments on the same data set produced a best classification accuracy of 53%.

More recently, the work by Zancanaro, Pianesi, et al. [133,105,38] have investigated the recognition of functional roles in small groups. Their work first involved the definition of a coding scheme, described in detail in [105], which included both task-based roles (i.e., roles related to the coordination and implementation of the tasks the group is undertaking and to the skills of each individual towards doing so) and socioemotional-based roles (i.e., roles related to the maintenance and regulation of the relationships between the group members). The first type of roles includes five categories (orienteer, giver, seeker, recorder, and follower); the second one also spans five classes (gatekeeper, protagonist, supporter, attacker, and neutral). As data, the authors used the Mission Survival Corpus (MSC-1), an audio-visual corpus of 11 meetings (approx. 4 h of data) that is actually the precursor of the one presented in [88], and was recorded with the same scenario [105]. Regarding nonverbal cues, semi-automatic audio cues (including speaker segmentation from close-talk microphones and the number of simultaneous speakers within the analyzed window), and automatic visual cues (two visual fidgeting measures, one for the body and one for the hand, extracted from motion features computed on skin regions) were extracted and used jointly. An SVM was used as a classifier for both types of roles, using short temporal sliding windows with observations for each participant as data samples. The best reported performance was 65% classification accuracy and 0.52 *F*-score for the task-based roles, and 70% classification accuracy and 0.55 *F*-score for the social-based roles. No analysis of the discrimination power of each perceptual modality was reported. A second approach, reported in [105], used the features for all participants to classify the role of an individual, which resulted in a significant improvement of performance. Dong et al. [38] extended this work by addressing the same two tasks using an influence model. Using a subset of eight meetings of the MSC-1 corpus, the authors found that this generative approach did not outperform an SVM when using features corresponding to a single individual and thus neither outperformed the best results obtained in [105] with multi-person features. However, Dong et al. suggested that the influence model is potentially less prone to overfitting, and also potentially convenient to handle case of recognizing roles in meetings with varying numbers of participants.

Vinciarelli studied the problem of role recognition in audio recordings of professional radio news shows [128]. Six roles corresponded to the different sections that people are responsible for or part of in a news show, including primary and secondary anchorman, guest, interviewee, headline person, and weather person. Unlike a regular meeting scenario, in this type of data not every person is conversing with each other (they might not even be present at the same time). Rather, the conversations are often dyadic and the sections of the show follow a regular structure, which facilitates the role recognition role problem compared to meetings like the ones discussed earlier in this section. The method uses an approach that extracts features based on basic concepts of social network analysis and on the duration of each of the role segments. The reported performance was 85% frame-based classification accuracy on 96 bulletins (with 12-min average duration). Experiments with a variation of the approach and another source of radio shows (talk-shows) was presented by Favre et al. with similar performance [48].

Jayagopi et al. [68] recently addressed a role-related problem, namely the estimation of role-based status in small groups, contrasting it to the related (but not equivalent) problem of estimating dominant people. While dominance has been defined in Section 4.1, status can be defined as "an ascribed or achieved quality implying respect or privilege, [but] does not necessarily include the ability to control others or their resources" [57] (p. 898). In the workplace, status often corresponds to a person's position in a group or an organization's hierarchy, and it is often defined by a role (e.g. a project manager). Dominance and status are related constructs: dominant-personality people often occupy high-status positions in an organization; conversely, high-status people are often allowed to behave dominantly with their subordinates. However, these two concepts do not always coincide, and can even contradict. Using the same meeting data as in previous work (5 h of AMI data divided into 5-min time slices), Jayagopi et al. presented a study on prediction of role-based status (the project manager of the team) and dominance using a rich number of automatic nonverbal cues that characterize speaking activity, visual activity, and visual attention. The work showed that although dominance and role-based high status might be related in terms of the associated nonverbal behavior, they are better explained by different cues; and that the best single nonverbal cues can correctly predict the person with highest dominance or role-based status with 70% segment-based classification accuracy.

Favre et al. [48] also attempted the audio-only recognition of the project manager and the other three pre-defined roles in a larger portion of the AMI corpus (138 meetings, 45 h). Using the data from full meetings, as opposed to thin slices as in the previous paragraph, the approach extracts features of each person's occurrence on a set of temporal windows, as well as the proportion of speaking time, and uses a simple Bayesian classifier. Unlike [68], evaluation was not done at the meeting or meeting segment level, but at the frame level, like in [133,128]. For the four-role task, a best performance of 44% classification accuracy was reported, but interestingly, the project manager class is recognized with 79% frame-based accuracy. Garg et al. [50] combined this approach with one that uses verbal information (words derived from manual or automatic speech transcripts). The results using automatic features showed a significant improvement over the use of nonverbal information only, with frame-based classification accuracy of 68% for the four roles, and of 84% for the project manager-only. This work shows that for this case, the fusion of verbal and nonverbal information was beneficial.

Finally, while much of the existing research in role modeling in small groups has been conducted in social situations where the fundamental goal is teamwork, very recent work has started to examine cases where competition, rather than cooperation or coordination, is the main goal. Raducanu et al. [108] proposed to investigate the case of role analysis in competitive meetings coming from a popular US reality TV show, where participants aim at getting a real job in a firm. In each episode, after participating in a business-related assigned task among two opposing teams, one participant is fired based on his/her performance in a group meeting led by a strong-minded boss. Raducanu et al. investigated simple approaches based on manually extracted cues related to high social status (speaking time and turns, interruptions, and centrality), and reported performance for the estimation of both the meeting chairman and the fired person of 85% and 92%, respectively, using 90 min of meeting data corresponding to a full season of the TV show.

6. Research infrastructure resources

The research discussed in this paper has been conducted using a number of collections, each of which varies with respect to the sensor setup, the type of recorded group conversations, the collection structure, and the type of existing annotations. Existing corpora

Please cite this article in press as: D. Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: A review, Image Vis. Comput. (2009), doi:10.1016/j.imavis.2009.01.004

designed and collected with the explicit goal of studying group interaction include, in rough chronological order:

- The meeting corpus from CMU Interactive Systems Lab (ISL) [130,17].
- The meeting corpus from the International Computer Science Institute (ICSI) [94,67,117].
- The meeting corpus from the Multimodal Meeting Manager (M4) European Project [91].
- The meeting corpus from the US National Institute of Standards and Technology (NIST) [119,51].
- The meeting corpus from the Augmented-Multi-Party Interaction (AMI) European Project [22,23].
- The Video Analysis and Content Extraction (VACE) meeting corpus from Virginia Tech [26].
- The meeting corpus from NTT's Communication Science Labs [96].
- The meeting corpus from the Advanced Telecommunications Lab (ATR) [19].
- The Mission Survivor Corpora from the Foundation Bruno Kessler (FBK-irst) [105,88].
- A suite of data sets recorded by MIT's Human Dynamics Lab [79,103].

A summary of information about these corpora appears in Table 2. It is important to notice that these collections have, in general, a disparate set of available annotations (some of which have been enriched over time), as is their degree of availability to the research community. For purposes of completeness, annotations related to both nonverbal and verbal information are included. Furthermore, the data sets vary widely in their degree of ecological validity, ranging from scripted conversations (e.g. the M4 corpus), to task-based conversations with pre-assigned roles (e.g. the AMI corpus), to task-based conversations with no roles (e.g. the MSC corpora), to real-life interactions that would have happened irrespectively of the recording process (e.g. in the ICSI and NIST corpora).

In addition, other data sets involving professional media (either TV or radio shows) have also been used for the research described here [128,48,108]. While in principle professional media provides certain characteristics that might facilitate analysis (good resolution and quality of data, and controlled settings), at the same time this source of data might impose a number of challenges (including edited multi-camera video that does not allow to obtain continuous observations for all the group members, single sound sources, music and other non-speech sources). Furthermore, most professional material often involves copyright issues and this limits their public distribution for research purposes.

7. Final discussion and conclusions

This paper has presented a review of the current facets of research on automatic nonverbal analysis of social interaction in small group conversations from sensor data. The number of research problems presented here is by no means exhaustive. The domain is challenging and still disorganized. In the author's opinion, some factors that contribute to this high entropy are the following:

- There is still not a clearly identified (and integrated) community within computer science on this subject. As the review has shown, work in this domain has been appearing in half a dozen communities in computing for the past years, although, as the statistics in Section 1 show, a significant portion of the literature appears in audio and speech and in multimodal and multimedia processing publications. Overall, researchers working in this domain have often come from different technical backgrounds, sometimes speak different technical jargons, and have different expectations on what should (or could) be investigated. Given the inertia of some of the traditional academic fields that have spanned this research, a degree of fragmentation will likely continue to exist for several years to come. This is not necessarily a negative factor, as the knowledge generated in the different communities has resulted in a richness of algorithms and tools that have been partly responsible for some of the most promising work so far.
- The list of relevant nonverbal cues, group interaction patterns, and social scenarios is large, as are our perception and participation in the social world [80]. The list of potential research problems is therefore large, and certainly larger than the size of the community working on them today. There is no current consensus on the core problems that ought to be investigated, although it seems clear that attempting to map every single problem related to nonverbal communication in small groups using computational approaches is not possible (or even desirable).
- Despite the large progress in social psychology and cognition, no single theory can answer the questions of what specific nonverbal cues and what concrete integration mechanisms are used to make sense of each social situation. Furthermore, such theories might not exist at all. As pointed out by Hall et al., "specific nonverbal behaviors often cannot be mapped onto specific meanings with any certainty" [57] (p. 898), and a social construct can be expressed or perceived differently depending on the specific conversational situation.
- Partly due to the previous point, the work in this field has, overall, a strong trial-and error, empirical flavor.

Table 2
Data sets explicitly recorded for small-group interaction research. The data modalities include audio (A), video (V), slides (S), handwritten notes (H), whiteboard notes (W), and motion (M). The annotations include speaker segmentation (SS), speech transcripts (ST), speaking-turn types (TT), dialog acts (DA), addressing (AD), visual focus (VF), body motion (BD), interest (IN), dominance (DO), roles (RO), personality (PE), and performance (PF). Only manual or semi-automatic annotations have been included. NA indicates that a piece of information was not located in the literature.

Corpus	Group size	No. meetings/Duration	Modalities	Annotations
ISL [17]	6.4 (avg.)	104/103 h	A,V	SS,ST
ICSI [67]	6 (avg.)	75/72 h	A	SS,ST,DA,IN
M4 [91]	4	60/5 h	A,V	SS,ST,TT,IN
NIST [51]	5.4 (avg.)	19/15 h	A,V	SS,ST
AMI [22]	4	167/100 h	A,V,S,H,W	SS,ST,DA,RO
AMI-12 [73]	4	12/5 h 30 min	A,V,S,H,W	AMI + AD,VF,DO
AMI-40 [112]	4	40/20 h	A	SS,ST,DO,RO
NTT [96]	4	4/22 min	A,V,M	SS
VACE [26]	5	NA	A,V,M	SS,ST,VF,DO
ATR [19]	4–9	10/10 h	A,V	SS,TT,BM,RO
MSC-1 [105]	4	11/3 h 45 min	A,V	SS,RO
MSC-2 [88]	4	13/6 h 47 min	A,V	SS,RO,PE
MIT [79]	4	36/10 h 48 mim	A,M	DO,PE,PF

Please cite this article in press as: D. Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: A review, Image Vis. Comput. (2009), doi:10.1016/j.imavis.2009.01.004

- Regarding perceptual processing, much of the existing work is still oriented to single modalities and/or often relies on (at least partially) manual cues. The degree of robustness for the extraction of social cues vary widely with respect to the data modality and the sensor setting. Audio cues seem to have an advantage in this direction. Furthermore, fully automatic multimodal approaches are still not the most common trend in the literature, and the true value of multimodal integration in this domain is still an open issue, even though there is support in the social psychology literature regarding the connections between multimodality and nonverbal communication.
- Research resources, including data, annotations, research tasks, and performance evaluation protocols are by no means mature. As of today, group conversational data is still, in general, not easy to record, maintain, distribute, and process. Annotation in many cases is expensive. Data might not be available in large quantities to make strong conclusions in terms of statistical significance.
- An issue that stands out regarding research resources is privacy. Viewed both as a human right and as a technical problem, privacy is a fundamental topic that needs to be addressed seriously in order to make steps towards systematic advances in research. At the moment, some of the research presented here is not reproducible or portable across data sets and conditions, due to the limitations in distributing data for public use when real-life data is used. On the other hand, data sets collected using role-playing techniques, scenarios, etc., might be easier to distribute publicly but might also have more limitations in terms of ecological validity.

Finally, reviewing a subject like the one discussed here constitutes a challenging enterprise, and so this article obviously has limitations. The author acknowledges that, inevitably, relevant works might have escaped his attention, and also that new papers will have appeared by the time this review is published. Nevertheless, the author hopes that the research discussed here will provide a concise entry point to newcomers to this domain, and will encourage others to continue making progress along several of the many research threads that are available today.

Acknowledgements

The author thanks the support of the Swiss National Center of Competence in Research on Interactive Multimodal Information Management (IM2), the EC project Augmented Multi-Party Interaction with Distant Access (AMIDA), and the US research program on Video Analysis and Content Extraction (VACE). He also thanks Jean-Marc Odobez, Hayley Hung, and Sileye Ba (Idiap) for discussions about several of the topics presented here, and Dinesh Jayagopi (Idiap) for his comments and technical help with the manuscript.

References

[1] J. Ajmera, C. Wooters, A robust speaker clustering algorithm, in: Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), St. Thomas, Dec. 2003.
[2] H.J.A. op den Akker, M. Theune, How do I address you? Modelling addressing behavior based on an analysis of a multi-modal corpus of conversational discourse, in: Proc. of the AISB Symposium on Multimodal Output Generation (MOG 2008), Aberdeen, Apr. 2008.
[3] M. Al-Hames, A. Dielmann, D. Gatica-Perez, S. Reiter, S. Renals, G. Rigoll, D. Zhang, Multimodal integration for meeting group action segmentation and recognition, in: Proc. of the Workshop on Machine Learning for Multimodal Interaction (MLMI), Edinburgh, May 2005.
[4] S.O. Ba, J.-M. Odobez, A probabilistic framework for joint head tracking and pose estimation, in: Proc. of the Int. Conf. on Pattern Recognition (ICPR), Cambridge, Aug. 2004.
[5] S.O. Ba, J.-M. Odobez, A study on visual focus of attention modeling using head pose, in: Proc. of the Workshop on Machine Learning for Multimodal Interaction (MLMI), Washington, DC, May 2006.
[6] S.O. Ba, J.M. Odobez, Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues, in: Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, Mar. 2008.
[7] K. Bachour, F. Kaplan, P. Dillenbourg, An interactive table for regulating face-to-face collaborative learning, in: Proc. of the European Conf. on Technology-Enhanced Learning (ECTEL), Maastricht, Sep. 2008.
[8] R.F. Bales, Interaction Process Analysis: A Method for the Study of Small Groups, Addison-Wesley, Reading, MA, 1951.
[9] S. Banerjee, J. Cohen, T. Quisel, A. Chan, Y. Patodia, Z. Al-Bawab, R. Zhang, P. Rybsi, M. Veloso, A. Black, R. Stern, R. Rosenfeld, A. Rudnicky, Creating multi-modal, user-centric records of meetings with the Carnegie Mellon meeting recording architecture, in: Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Meeting Recognition Workshop, Montreal, Mar. 2004.
[10] S. Banerjee, A. Rudnicky, Using simple speech-based features to detect the state of a meeting and the roles of the meeting participants, in: Proc. of the Int. Conf. on Spoken Language Processing (ICSLP), Jeju Island, Oct. 2004.
[11] S. Basu, T. Choudhury, B. Clarkson, A. Pentland, Towards measuring human interactions in conversational settings, in: Proc. of the IEEE CVPR Int. Workshop on Cues in Communication (CVPR-CUES), Kauai, Dec. 2001.
[12] S. Bengio, An asynchronous Hidden Markov Model for audio-visual speech recognition, in: Proc. of the Conf. on Advances in Neural Information Processing Systems, NIPS 15, Vancouver, Dec. 2002.
[13] T. Bergstrom, K. Karahalios, Conversation clock: visualizing audio patterns in co-located groups, in: Proc. of the Hawaii Int. Conf. on Systems Science (HICSS), Hawaii, 2007.
[14] M. Brand, N. Oliver, A. Pentland, Coupled hidden Markov models for complex action recognition, in: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, Jun. 1997.
[15] O. Brdiczka, J. Maisonnasse, P. Reigner, Automatic detection of interaction groups, in: Proc. of the Int. Conf. on Multimodal Interfaces (ICMI), Trento, Oct. 2005.
[16] J.K. Burgoon, N.E. Dunbar, Nonverbal expressions of dominance and power in human relationships, in: V. Manusov, M. Patterson (Eds.), The Sage Handbook of Nonverbal Communication, Sage, Beverley Hills, CA, 2006.
[17] S. Burger, V. MacLaren, H. Yu, The ISL meeting corpus: the impact of meeting type on speech style, in: Proc. of the ICSLP, Denver, Sep. 2002.
[18] N. Campbell, On the use of nonverbal speech sounds in human communication, in: Proc. of the COST 2102 Workshop on Verbal and Nonverbal Communication Behaviours, Vietri sul Mare, Mar. 2007.
[19] N. Campbell, T. Sadanobu, M. Imura, N. Iwahashi, S. Noriko, D. Douxchamps, A multimedia database of meeting and informal interactions for tracking participant involvement and discourse flow, in: Proc. of the Language Resources and Evaluation Conference (LREC), Genoa, May 2006.
[20] N. Campbell, D. Douxchamps, Processing image and audio information for recognizing discourse participation status through features of face and voice, in: Proc. of the INTERSPEECH, 2007.
[21] J.N. Cappella, Controlling the floor in conversation, in: A.W. Siegman, S. Feldstein (Eds.), Multichannel Integrations of Nonverbal Behavior, Erlbaum, Hillsdale, NJ, 1985.
[22] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec,V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, P. Wellner, The AMI meeting corpus: a pre-announcement, in: Proc. of the Workshop on Machine Learning for Multimodal Interaction (MLMI), Edinburgh, Jul. 2005.
[23] J. Carletta, Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus, Language Resources and Evaluation Journal 41 (2) (2007) 181–190.
[24] T.L. Chartrand, J.A. Bargh, The chameleon effect: the perception-behavior link and social interaction, Journal of Personality and Social Psychology 76 (6) (1999) 893–910.
[25] T.L. Chartrand, W. Maddux, J. Lakin, Beyond the perception-behavior link: the ubiquitous utility and motivational moderators of nonconscious mimicry, in: R. Hassin, J. Uleman, J.A. Bargh (Eds.), The New Unconscious, Oxford University Press, Berlin, 2005.
[26] L. Chen, T.R. Rose, F. Parrill, X. Han, J. Tu, Z. Huang, M. Harper, F. Quek, D. McNeill, R. Tuttle, T. Huang, VACE multimodal meeting corpus, in: Proc. of the Workshop on Machine Learning for Multimodal Interaction (MLMI), Edinburgh, Jul. 2005.
[27] L. Chen, M. Harper, A. Franklin, T.R. Rose, I. Kimbara, Z. Huang, F. Quek, A multimodal analysis of floor control in meetings, in: Proc. of the Workshop on Machine Learning for Multimodal Interaction (MLMI), Washington, DC, May 2006.
[28] Y.S. Choi, H.M. Gray, N. Ambady, The glimpsed world: unintended communication and unintended perception, in: R.H. Hassin, J.S. Uleman, J.A. Bargh (Eds.), The New Unconscious, Oxford University Press, Oxford, 2005.
[29] T. Choudhury, A. Pentland, Modeling face-to-face communication using the sociometer, in: Proc. of the Int. Conf. on Ubiquitous Computing, Seattle, Oct. 2003.
[30] T. Choudhury, S. Basu, Modeling conversational dynamics as a mixed memory Markov process, in: Proc. of the NIPS, Dec. 2004.
[31] H.H. Clark, T.B. Carlson, Hearers and speech acts, Language 58 (2) (1982) 332–373.
[32] M. Cook, J.M.C. Smith, The role of gaze in impression formation, British Journal of Social and Clinical Psychology (1975).

Please cite this article in press as: D. Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: A review, Image Vis. Comput. (2009), doi:10.1016/j.imavis.2009.01.004

[33] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J. Taylor, Emotion recognition in human–computer interaction, IEEE Signal Processing Magazine (2001).
[34] J.R. Curhan, A. Pentland, Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes, Journal of Applied Psychology 92 (3) (2007) 802–811.
[35] A. Dielmann, S. Renals, Dynamic Bayesian networks for meeting structuring, in: Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Montreal, May 2004.
[36] A. Dielmann, S. Renals, Dynamic Bayesian networks for meeting structuring, IEEE Transactions in Multimedia 9 (1) (2007) 25–36.
[37] J.M. DiMicco, A. Pandolfo, W. Bender influencing group participation with a shared display, in: Proc. of the ACM Conf. on Computer Supported Cooperative Work (CSCW), Chicago, Nov. 2004.
[38] W. Dong, B. Lepri, A. Capelletti, A. Pentland, F. Pianesi, M. Zancanaro, Using the influence model to recognize functional roles in meetings, in: Proc. of the Int. Conf. on Multimodal Interfaces (ICMI), Nagoya, Nov. 2007.
[39] J.F. Dovidio, S.L. Ellyson, Decoding visual dominance: atributions of power based on relative percentages of looking while speaking and looking while listening, Social Psychology Quarterly 45 (2) (1982) 106–113.
[40] N.E. Dunbar, J.K. Burgoon, Perceptions of power and interactional dominance in interpersonal relationships, Journal of Social and Personal Relationships 22 (2) (2005) 207–233.
[41] N.E. Dunbar, J.K. Burgoon, Measuring nonverbal dominance, in: V. Manusov (Ed.), The Sourcebook of Nonverbal Measures: Going Beyond Words, Erlbaum, Hillsdale, NJ, 2005.
[42] S. Duncan, Some signals and rules for taking speaker turns in conversations, Journal of Personality and Social Psychology 23 (2) (1972) 283–292.
[43] S. Dupont, J. Luettin, Audio-visual speech modeling for continuous speech recognition, IEEE Transactions on Multimedia 2 (3) (2000) 141–151.
[44] N. Eagle, A. Pentland, Social network computing, in: Proc. of the Int. Conf. on Ubiquituous Computing (UBICOMP), Seattle, Oct. 2003.
[45] J.S. Efran, Looking for approval: effects of visual behavior of approbation from persons differing in importance, Journal of Personality and Social Psychology 10 (1) (1968) 21–25.
[46] S.L. Ellyson, J.F. Dovidio (Eds.), Power Dominance and Nonverbal Behavior, Springer, Berlin, 1985.
[47] R.V. Exline, S.L. Ellyson, B. Long, Visual behavior as an aspect of power role relationships, in: Advances in the Study of Communication and Affect, Plenum Press, New York, 1975.
[48] S. Favre, H. Salamin, J. Dines, A. Vinciarelli. Role recognition in multiparty recordings using social affiliation networks and discrete distributions, in: Proc. of the Int. Conf. on Multimodal Interfaces (ICMI), Chania, Oct. 2008.
[49] N. Fay, S. Garod, J. Carletta, Group discussion as interactive dialogue or serial monologue: the influence of group size, Psychological Science 11 (6) (2000) 487–492.
[50] N. Garg, S. Favre, H. Salamin, D. Hakkani-Tr, A. Vinciarelli, Role recognition for meeting participants: an approach based on lexical information and social network analysis, in: Proc. of the ACM Int. Conf. on Multimedia, Vancouver, Oct. 2008.
[51] J. Garofolo, M. Michel, C. Laprun, V. Stanford, E. Tabassi, The NIST meeting room pilot corpus, in: Proc. of the Language Resources and Evaluation Conference (LREC), Lisbon, May 2004.
[52] D. Gatica-Perez, I. McCowan, D. Zhang, S. Bengio, Detecting group interest-level in meetings, in: Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Philadelphia, Mar. 2005.
[53] D. Gatica-Perez, D. Zhang, S. Bengio, Extracting information from multimedia meeting collections, in: Proc. of the ACM Int. Conf. on Multimedia, Workshop on Multimedia Information Retrieval (ACM MM MIR), Singapore, Nov. 2005.
[54] D. Gatica-Perez, Analyzing human interaction in conversations: a review, in: Proc. of the IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI), Heidelberg, Sep. 2006.
[55] R. Gifford, Personality and nonverbal behavior: a complex conundrum, in: V. Manusov, M.L. Patterson (Eds.), The SAGE Handbook of Nonverbal Communication, Sage, Beverly Hills, CA, 2006.
[56] C. Goodwin, Conversational Organization: Interaction Between Speakers and Hearers, vol. 11, Academic Press, New York, NY, 1981.
[57] J.A. Hall, E.J. Coats, L.S. LeBeau, Nonverbal behavior and the vertical dimension of social relations: a meta-analysis, Psychological Bulletin 131 (6) (2005) 898–924.
[58] A.P. Hare, Types of roles in small groups: a bit of history and a current perspective, Small Group Research 25 (Aug.) (1994) 433–448.
[59] R.H. Hassin, J.S. Uleman, J.A. Bargh (Eds.), The New Unconscious, Oxford University Press, Oxford, 2005.
[60] D. Heylen, D. Reidsma, R. Ordelman, Annotating state of mind in meeting data, in: Proc. of the LREC Workshop on Corpora for Research on Emotion and Affect, Genoa, May 2006.
[61] D. Heylen, A. Nijholt, M. Poel, Generating nonverbal signals for a sensitive artificial listener, in: Proc. of the COST 2102 Workshop on Verbal and Nonverbal Communication Behaviours, Vietri sul Mare, Mar. 2007.
[62] D. Hillard, M. Ostendorf, E. Shriberg, Detection of agreement vs. disagreement in meetings: training with unlabeled data, in: Proc. of the HLT-NAACL Conference, Edmonton, May 2003.
[63] H. Hung, D. Jayagopi, C. Yeo, G. Friedland, S. Ba, J.-M. Odobez, K. Ramchandran, N. Mirghafori, D. Gatica-Perez, Using audio and video features to classify the most dominant person in a group meeting, in: Proc. of the ACM Int. Conf. on Multimedia (ACM MM), Augsburg, Sep. 2007.
[64] H. Hung, Y. Huang, G. Friedland, D. Gatica-Perez, Estimating the dominant person in multi-party conversations using speaker diarization strategies, in: Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, Mar. 2008.
[65] H. Hung, Y. Huang, C. Yeo, D. Gatica-Perez. Associating audio-visual activity cues in a dominance estimation framework, in: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB), Anchorage, Jun. 2008.
[66] H. Hung, D. Jayagopi, S. Ba, J.-M. Odobez, D. Gatica-Perez, Investigating automatic dominance estimation in groups from visual attention and speaking activity, in: Proc. of the Int. Conf. on Multimodal Interfaces (ICMI), Chania, Oct. 2008.
[67] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters, The ICSI meeting corpus, in: Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Hong-Kong, Apr. 2003.
[68] D. Jayagopi, S. Ba, J.-M. Odobez, D. Gatica-Perez Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues, in: Proc. of the Int. Conf. on Multimodal Interfaces (ICMI), Chania, Oct. 2008.
[69] D. Jayagopi, H. Hung, C. Yeo, D. Gatica-Perez, Predicting the dominant clique in group conversations with nonverbal cues, in: Proc. of the ACM Int. Conf. on Multimedia, Vancouver, Oct. 2008.
[70] D. Jayagopi, H. Hung, C. Yeo, D. Gatica-Perez Modeling dominance in group conversations using nonverbal activity cues, IEEE Trans. on Pattern Analysis and Machine Intelligence, Special Issue on Multimodal Processing for Speech-based Interactions, vol. 17, No. 3, Mar. 2009.
[71] O.P. John, S. Srivastava, The big five trait taxonomy: history measurement and theoretical perspectives, in: L.A. Pervian, O.P. John (Eds.), Handbook of Personality Theory and Research, Guilford Press, New York, 1999.
[72] N. Jovanovic, R. op den Akker, Towards automatic addressee identification in multi-party dialogues, in: Proc. of the SIGDial Workshop on Discourse and Dialogue, Boston, Apr. 2004.
[73] N. Jovanovic, R. op den Akker, A. Nijholt, A corpus for studying addressing behavior in multi-party dialogues, in: Proc. of the SIGDial Workshop on Discourse and Dialogue, Lisbon, Sep. 2005.
[74] N. Jovanovic, R. op den Akker, A. Nijholt, Addressee identification in face-to-face meetings, in: Proc. of the Conf. European Chapter of the Association for Computational Linguistics (EACL), Trento, Apr. 2006.
[75] N. Jovanovic, To whom it may concern: addressing in face-to-face meetings, Doctoral Dissertation, Department of Computer Science, University of Twente, Mar. 2007.
[76] M. Katzenmeier, R. Stiefelhagen, T. Schultz, Identifying the addressee in human–human–robot interactions based on head pose and speech, in: Proc. of the Int. Conf. on Multimodal Interfaces (ICMI), State College, PA, Oct. 2004.
[77] L. Kennedy, D. Ellis, Pitch-based emphasis detection for characterization of meeting recordings, in: Proc. of the ASRU, Virgin Islands, Dec. 2003.
[78] L. Kennedy, D. Ellis, Laughter detection in meetings, in: Proc. of the ICASSP 2004.
[79] T. Kim, A. Chang, L. Holland, A. Pentland, Meeting mediator: enhancing group collaboration with sociometric feedback, in: Proc. of the ACM Conf. on Computer Supported Cooperative Work (CSCW), San Diego, Nov. 2008.
[80] M.L. Knapp, J.A. Hall, Nonverbal Communication in Human Interaction, 6th ed., Wadsworth, Berlin, 2005.
[81] O. Kulyk, C. Wang, J. Terken, Real-time feedback based on nonverbal behaviour to enhance social dynamics in small group meetings, in: Proc. of the Int. Workshop on Machine Learning for Multimodal Interaction (MLMI), Edinburgh, Jul. 2005.
[82] M. Knox, N. Mirghafori, Automatic laughter detection using neural networks, in: Proc. of the Interspeech, Antwerp, 2007.
[83] K. Laskowski, T. Schultz, Detection of laughter-in-interaction in multichannel close-talk microphone recordings of meetings, in: Proc. of the Int. Workshop on Machine Learning for Multimodal Interaction (MLMI), Utrecht, Sep. 2008.
[84] A. Leffler, D.L. Gillespie, J.C. Conaty, The effects of status differentiation on nonverbal behavior, Social Psychology Quarterly 45 (3) (1982) 151–161.
[85] C.B. Lynn Smith-Lovin, Interruptions in group discussions: the effects of gender and group composition, American Sociological Review 54 (3) (1989) 424–435.
[86] A. Madan, Thin slices of interest, Master's Thesis, Massachusetts Institute of Technology, 2005.
[87] A. Madan, R. Caneel, A. Pentland, Voices of attraction, in: Proc. of the Int. Conf. on Augmented Cognition (AC-HCI), Las Vegas, Jul. 2005.
[88] N. Mana, B. Lepri, P. Chippendale, A. Cappelletti, F. Pianesi, P. Svaizer, M. Zancanaro, Multimodal corpus of multi-party meetings for automatic social behavior analysis and personality traits detection, in: Proc. of the Int. Conf. on Multimodal Interfaces (ICMI), Workshop on Tagging, Mining, and Retrieval of Human-Related Activity Information, Nagoya, Nov. 2007.
[89] V. Manusov, M.L. Patterson (Eds.), The SAGE Handbook of Nonverbal Communication, Sage, Beverley Hills, CA, 2006.
[90] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, H. Bourlard, Modeling human interactions in meetings, in: Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Hong-Kong, Apr. 2003.
[91] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, D. Zhang, Automatic analysis of multimodal group actions in meetings, IEEE

Please cite this article in press as: D. Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: A review, Image Vis. Comput. (2009), doi:10.1016/j.imavis.2009.01.004

Transactions on Pattern Analysis and Machine Intelligence 27 (3) (2005) 305–317.
[92] J.E. McGrath, Groups: Interaction and Performance, Prentice-Hall, Englewood Cliffs, NJ, 1984.
[93] D. McNeill (Ed.), Language and Gesture, Cambridge University Press, Cambridge, MA, 2000.
[94] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, A. Stolcke, The meeting project at ICSI, in: Proc. of the Human Language Technology Conf. (HLT), San Diego, CA, March 2001.
[95] N. Oliver, E. Horvitz, A. Garg, Layered representations for learning and inferring office activity from multiple sensory channels, in: Proceedings of the International Conference on Multimodal Interfaces (ICMI'02), October 2002.
[96] K. Otsuka, Y. Takemae, J. Yamato, H. Murase, Probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances, in: Proc. of the Int. Conf. on Multimodal Interfaces (ICMI), Trento, Oct. 2005.
[97] K. Otsuka, J. Yamato, Y. Takemae, H. Murase, Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns, in: Proc. of the ACM CHI Extended Abstract, Montreal, Apr. 2006.
[98] K. Otsuka, J. Yamato, Y. Takemae, H. Murase, Conversation scene analysis with dynamic Bayesian network based on visual head tracking, in: Proc. of the IEEE Int. Conf. on Multimedia (ICME), Toronto, Jul. 2006.
[99] K. Otsuka, J. Yamato, H. Sawada, Automatic inference of cross-modal nonverbal interactions in multiparty conversations, in: Proc. of the Int. Conf. on Multimodal Interfaces (ICMI), Nagoya, Nov. 2007.
[100] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, J. Yamato, A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization, in: Proc. of the Int. Conf. on Multimodal Interfaces (ICMI), Chania, Oct. 2008.
[101] A. Pentland, Socially aware computation and communication, IEEE Computer 38 (Mar.) (2005) 63–70.
[102] A. Pentland, A. Madan, Perception of social interest, in: Proc. of the IEEE Int. Conf. on Computer Vision, Workshop on Modeling People and Human Interaction (ICCV-PHI), Beijing, Oct. 2005.
[103] A. Pentland, Honest Signals: How They Shape Our World, MIT Press, Cambridge, MA, 2008.
[104] S. Petridis, M. Pantic, Audiovisual laughter detection based on temporal features, in: Proc. of the Int. Conf. on Multimodal Interfaces (ICMI), Chania, Oct. 2008.
[105] F. Pianesi, M. Zancanaro, B. Lepri, A. Cappelletti, A multimodal annotated corpus of consensus decision making meetings, Language Resources and Evaluation 41 (3–4) (2007).
[106] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, M. Zancanaro, Multimodal recognition of personality traits in social interactions, in: Proc. of the Int. Conf. on Multimodal Interfaces (ICMI), Chania, Oct. 2008.
[107] L.R. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[108] B. Raducanu, J. Vitria, D. Gatica-Perez, You're fired! Nonverbal role analysis in competitive meetings, in: Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Taiwan, Apr. 2009.
[109] D. Reidsma, D. Heylen, and R. Ordelman, ''Annotating Emotion in Meetings Proceedings, in Proc. LREC, Genoa, May 2006.
[110] S. Reiter, B. Schuller, G. Rigoll, Segmentation and recognition of meeting events using a two-layered HMM and a combined MLP-HMM approach, in: Proc. of the IEEE Int. Conf. on Multimedia (ICME), Toronto, Jul. 2006.
[111] R.J. Rienks, D. Heylen, Automatic dominance detection in meetings using easily detectable features, in: Proc. of the Workshop on Machine Learning for Multimodal Interaction (MLMI), Edinburgh, Jul. 2005.
[112] R. Rienks, D. Zhang, D. Gatica-Perez, W. Post, Detection and application of influence rankings in small-group meetings, in: Proc. of the Int. Conf. on Multimodal Interfaces (ICMI), Banff, Nov. 2006.
[113] R. Rienks, A. Nijholt, D. Heylen, Verbal behavior of the more and the less influential meeting participant, in: Proc. of the Int. Conf. on Multimodal Interfaces (ICMI), Workshop on Tagging, Mining and Retrieval of Human Related Activity Information, Nagoya, Oct. 2007.
[114] J.B. Rotter, Generalized expectancies for internal versus external control of reinforcement, Psychological Monographs 80 (1965).
[115] A.J. Salazar, An analysis of the development evolution of roles in the small group, Small Group Research 27 (4) (1996) 475–503.
[116] M. Schmid Mast, Dominance as expressed and inferred through speaking time: a meta-analysis, Human Communication Research 28 (3) (2002) 420–450.
[117] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, H. Carvey, The ICSI meeting recorder dialog act (MRDA) corpus, in: Proc. of the HLT-NAACL SIGDIAL Workshop, Boston, Apr. 2004.
[118] E. Shriberg, Spontaneous speech: how people really talk and why engineers should care, in: Proc. of the European Conf. on Speech Communication and Technology (Eurospeech), Lisbon, Sep. 2005.
[119] V. Stanford, J. Garofolo, M. Michel, The NIST smart space and meeting room projects: signals, acquisition, annotation, and metrics, in: Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Hong Kong, 2003.
[120] R. Stiefelhagen, J. Yang, A. Waibel, Modeling focus of attention for meeting indexing based on multiple cues, IEEE Transactions on Neural Networks 13 (4) (2002) 928–938.
[121] R. Stiefelhagen, Tracking focus of attention in meetings, in: Proc. of the Int. Conf. on Multimodal Interfaces (ICMI), Pittsburgh, PA, 2002.
[122] W.T. Stoltzman, Toward a social signaling framework: activity and emphasis in speech, Master's Thesis, Massachusetts Institute of Technology, Sep. 2006.
[123] J. Sturm, O. Houben-Van Herwijnen, A. Eyck, J. Terken, Influencing social dynamics in meetings through a peripheral display, in: Proc. of the Int. Conf. on Multimodal Interfaces (ICMI), Nagoya, Oct. 2007.
[124] Y. Takemae, K. Otsuka, N. Mukawa, An analysis of speakers' gaze behavior for automatic addressee identification in multiparty conversation and its application to video editing, in: Proc. of the IEEE Int. Workshop on Robot and Human Interface Communication, Okayama, Sep. 2004.
[125] K. Truong, D. van Leeuwen, Automatic detection of laughter, in: Proc. of the Interspeech, Lisbon, 2005.
[126] K. Truong, D. van Leeuwen, Automatic discrimination between laughter and speech, Speech Communication 49 (2) (2007) 144–158.
[127] K.J. Tusing, J.P. Dillard, The sounds of dominance: vocal precursors of perceived dominance during interpersonal influence, Human Communication Research 26 (1) (2000) 148–171.
[128] A. Vinciarelli, Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling, IEEE Transactions on Multimedia 9 (6) (2007) 1215–1226.
[129] A. Vinciarelli, M. Pantic, H. Bourlard, A. Pentland, Social signal processing: state-of-the-art and future perspectives of an Emergin domain, in: Proc. of the ACM Int. Conf. on Multimedia, Vancouver, Oct. 2008.
[130] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, K. Zechner, Advances in automatic meeting record creation and access, in: Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Salt Lake City, UT, May 2001.
[131] B. Wrede, E. Shriberg, Spotting hotspots in meetings: human judgments and prosodic cues, in: Proc. of the Eurospeech, Geneva, Sep. 2003.
[132] B. Wrede, E. Shriberg, The relationship between dialogue acts and hot spots in meetings, in: Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Virgin Islands, Dec. 2003.
[133] M. Zancanaro, B. Lepri, F. Pianesi, Automatic detection of group functional roles in face to face interactions, in: Proc. of the Int. Conf. on Multimodal Interfaces (ICMI), Banff, Nov. 2006.
[134] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, G. Lathoud, Modeling individual and group actions in meetings: a two-layer HMM framework, in: Proc. of the IEEE CVPR Workshop on Event Mining, Washington, DC, Jun. 2004.
[135] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, G. Lathoud, Multimodal group action clustering in meetings, in: Proc. of the ACM Int. Conf. on Multimedia, Workshop on Video Surveillance and Sensor Networks (ACM MM-VSSN), New York, Oct. 2004.
[136] D. Zhang, D. Gatica-Perez, S. Bengio, Semi-supervised meeting event recognition with adapted HMMs, in: Proc. of the IEEE Int. Conf. on Multimedia (ICME), Amsterdam, Jul. 2005.
[137] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, Modeling individual and group actions in meetings with layered HMMs, IEEE Transactions on Multimedia 8 (3) (2006).

Please cite this article in press as: D. Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: A review, Image Vis. Comput. (2009), doi:10.1016/j.imavis.2009.01.004