CHAPTER 12

Nonverbal Behavior Analysis

DANIEL GATICA-PEREZ, ALESSANDRO VINCIARELLI and JEAN-MARC ODOBEZ

12.1 Introduction: a brief history of nonverbal behavior research in IM2

The last decade marked the emergence of the automated understanding of face-to-face social interaction as a research problem in computing. IM2 was originally focused on meetings (a quintessential form of interaction), and so over the years a body of work directed towards analyzing and inferring a variety of behaviors and interactions resulted from the project.

One key aspect of the IM2 work has been the use of nonverbal communication as measurable evidence of social phenomena. The role of nonverbal behavioral cues (gaze, facial expressions, gestures, vocalizations, postures, etc.) as carriers of socially relevant information has been the subject of research in psychology and communication for decades. Furthermore, computing research has developed a large number of approaches aimed at automatic analysis and synthesis of gaze, facial expressions, gestures, and paralanguage. Over 12 years, IM2 enabled both the development of perceptual technologies (computer vision and signal processing) to extract behavioral cues, and their integration to address questions connected to inference of various social variables in increasingly diverse situations (Gatica-Perez, 2009, Vinciarelli et al., 2009b). Table 12.1 shows a timeline of some of the investigated research lines.

The initial research that linked audio-visual perception and social behavior in IM2 can be traced back to 2002 with the initial use of the Smart Meeting Room as a sensing platform to study small group interaction (refer to Chap. 1). The original contribution was the dual realization that groups could be studied as units (as opposed to considering individuals as the ultimate analysis target), and that characterizing group behavior computationally could be achieved through the integration of machine perception technologies with findings in the small group research literature. Such a concept took shape through the definition of "meeting actions", a categorization system that related speaking turns and visual activity, and a framework for recognition using audio and video observations and Hidden Markov Models (McCowan et al., 2005).

Table 12.1 Timeline of IM2 research on social behavior analysis. Data modalities include A: audio; V: video; AV: audio-video; D: depth; G: gyroscope.

Year	**Topic**	**Modality**
2002	Smart Meeeting Room	AV
2002	Meeting Group Actions	AV
2003	Head Pose Estimation	V
2004	Group Interest	AV
2005	VFOA from head pose	V
2006	Wandering VFOA	V
2006	Dominance	AV
2006	Speaking Style	A
2007	Contextual VFOA	AV
2008	Role Recognition	A
2008	Group Characterization	A
2009	Video Blogging Analysis	AV
2009	Emergent Leadership	AV
2009	Group Cohesion	AV
2010	Personality Perception	A
2011	VFOA in open space	V
2011	VFOA with robots	V
2011	Kinect Gaze Sensing	DV
2011	Conflict Detection	AV
2012	Effectiveness of Delivery	AV
2012	Interpersonal Attraction	AG

Between 2002 and 2005, the work on group behavior analysis expanded into several directions. On one hand, new work on computational methodologies to recognize group activities was pursued through layered dynamical approaches (Zhang et al., 2006), in which individual and group behavior were both recognized in tandem HMM architectures, showing a number of benefits. These approaches were also studied by other researchers (Al-Hames et al., 2005). On the other hand, new concepts related to group behavior started to be studied, namely interest, and addressed via dynamical models and nonverbal features (Gatica-Perez et al., 2005).

Gaze is one a nonverbal social cue that plays a major role in human interaction. Its role in human communication ranges from establishing relationships and expressing intimacy to exercising social control, and its function as a cue to regulate the course of interaction, via turn holding, taking, or yielding, has been established in social psychology (Kendon, 1967, Goodwin and Heritage, 1990). The ability to accurately estimate gaze provides a significant input to social behavior analysis algorithms, and so the extraction of gaze information was as a natural task to be addressed in IM2. From an historical perspective, gaze analysis followed closely the work on recognition of group actions in meetings (McCowan et al., 2005). After some years of development as a standalone component (Ba and Odobez, 2006), the integration of gaze with audio cues into a single model was achieved (Ba and Odobez, 2008), reflecting that speaking activity and gaze are coordinated in human communication processes.

From 2005 onwards, with the availability of the Augmented Multi-Party Interaction (AMI) meeting corpus, the research in group behavior within IM2 started to study aspects of social verticality like dominance. The development of approaches for the recognition of perceived dominance included studies on inferring most and least dominant people (Hung et al., 2007, Jayagopi et al., 2009a), on the relation between dominance and status (Jayagopi et al., 2008), and on the effect of specific behavioral nonverbal features and modalities (Hung et al., 2008, 2011). This work expanded over the years through collaborations with other projects at Idiap (Aran and Gatica-Perez, 2010, Sanchez-Cortes et al., 2012) and with other colleagues (Kalimeri et al., 2012).

In parallel, other developments were pursued. One of them was role recognition in multi-party conversations, using automatically extracted speaking turns with speaker diarization techniques developed by IM2 partners (Vinciarelli, 2007). This research was conducted on radio news, radio talk shows, and AMI meetings, and greatly benefited from the collaboration with IM2 speech researchers.

Between 2008 and 2010, other aspects of group behavior began to be studied, including competitive vs. cooperative group interactions (Jayagopi et al., 2009b) and group cohesion (Hung and Gatica-Perez, 2010). A modeling novelty introduced in this period was the study of data mining approaches to discover conversational pattern structure in group discussions, as opposed to much of the work done previously, which relied on supervised learning techniques (Jayagopi and Gatica-Perez, 2009, 2010). Research beyond the AMI corpus also started through external collaborations (Jayagopi et al., 2012).

The last phase of IM2 brought with it a diversity of topics. In the first place, the study of personality traits as targets for automated analysis using audio features was pursued (Mohammadi and Vinciarelli, 2012). Some first approaches were proposed, promising in particular for traits that listeners attribute to speakers they listen to for the first time. In the second place, the study of conflict in group discussions (i.e., situations that arise whenever two or more parties pursue individual, incompatible goals) used TV political debates as data source and audio cues as input (Kim et al., 2012). Finally, research also expanded beyond face-to-face communication, analyzing the phenomenon of conversational video in social media (i.e., video blogging), and finding various connections between behavioral cues extracted from audio and video, social attention from audiences, and personality traits (Biel and Gatica-Perez, 2011, 2013).

In the rest of this chapter, we review work on three research lines. Section 12.2 discusses modeling of visual focus of attention (Odobez's research group). Section 12.3 discusses social signal processing (Vinciarelli's research group). Section 12.4 discusses behavioral analysis of video blogging (Gatica-Perez's research group). We close the chapter with a few final remarks.

12.2 VFOA recognition for communication analysis in meeting rooms and beyond

Sensing gaze is a very difficult task and has been traditionnaly performed using Human-Computer Interaction (HCI) techniques. Usually, such systems require either invasive head mounted sensors or accurate estimates of the eye features obtained from high-resolution iris images (Morimoto and Mimica, 2005). However, they might be very difficult to set up for several people around a table or in a room and can interfere and affect the naturalness of human activity by restricting for instance head mobility and orientation. As a consequence, in order to leave laboratory experiments and move towards more open situations, researchers have started investigating gaze analysis from head pose. While such an approach throws away the important eye-in-head orientation component of the gaze and thus can not lead to accurate gaze estimates, defined as direction in the 3D space, head pose can still be sufficient to address a discrete version of the gaze recognition task that is often the final interest in human behavior analysis applications: the recognition of the visual focus of attention (VFOA) which answers the question "who looks at whom or what", i.e., which visual target the gaze of a person is oriented at. The use of head pose as base cue for gaze recognition is supported by psychovisual evidence showing that people do exploit head orientation to infer people's gaze (Langton et al., 2000), and by empirical evidence demonstrated in different conversational setting. For instance, Stiefelhagen et al. (2002) showed in a simple meeting setting involving 4 people having short conversations that VFOA recognition rates of 70 to 80% could be achieved.

In the following, we present our main investigations in the gaze analysis domain. Since head pose is a central element of the approach, we first present the main methodology we used to estimate it. Then, we demonstrate how VFOA recognition was achieved in two different situations: in a meeting scenario and in an outdoor setting. We conclude with some recent works on VFOA recognition with robots and full gaze estimation from RGB-D data (i.e., Kinect) and perspectives.

12.2.1 Head pose estimation

Head pose estimation methodologies and accuracies typically depend on the image resolution at hand and view point. High performance can reasonably be achieved with 2D or 3D Active Shape or Appearance models when dealing with high resolution images and near frontal head poses. The task is much more difficult when handling mid-resolution head video sequences (e.g., left image of Figure 12.1) and people with natural head movements, that can be extremely fast and have a significant amount of profile or worse looking down head poses.

To address the latter situations and issues, we proposed a robust tracking method in which head tracking and pose estimation are considered as two coupled problems in a Bayesian probabilistic framework (Ba and Odobez, 2005b). More precisely, the *joint* tracking of the head location (position, scale, in-plane

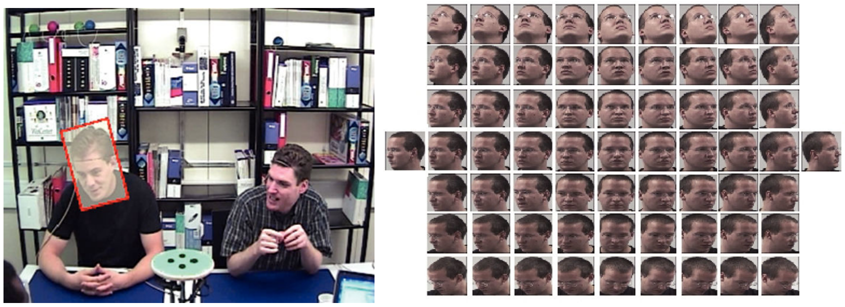

Figure 12.1 Head pose extraction. The head localization (position, scale, in-plane rotation, in the left image) and head pose (discrete index identifying a specific out-of-plane rotation, in the right image) are jointly tracked.

rotation) and pose (represented by a discrete index denoting an element of the out-of-plane head pose set, as shown in Figure 12.1) was conducted. Texture (output of one Gaussian and two Gabor filters) and skin color pose dependent head appearance models were built from a training database, and used to evaluate the likelihood of observed features. An additional pose-independent likelihood model based on background subtraction feature was used to provide better localization information and reduce tracking failures. More details on models and estimation procedure can be found in (Ba and Odobez, 2005b).

Estimated on more than 100 minutes of video recording featuring people involved in natural conversation (Ba and Odobez, 2005a), the algorithm produced an average error of around 12 degrees for the pan and 10 degrees for the tilt. It is important to note that the error was of only 7 degrees when considering video samples with a pose lower than 45 degrees.

12.2.2 VFOA recognition in meetings

From a human interaction viewpoint, meetings constitute a very interesting and quite complex testbed to study communication mechanisms and beyond that, i.e., social constructs. As one key non-verbal behavior involved in this process, VFOA recognition in meetings has been investigated by different researchers using head pose as the main cues (Stiefelhagen et al., 2002, Otsuka et al., 2005, Ba and Odobez, 2006). However, early works like those by Stiefelhagen et al. (2002) and by Otsuka et al. (2005) were only considering small datasets, usually consisting of a few relatively short interactions (from 2 minutes to 8 minutes) in simple settings. With the AMI/IM2 meeting corpus, we obtained a much larger VFOA-annotated corpus (5 hours). It was composed of 12 meetings ranging from 15 to 35 minutes, involving standing people, presentations, and interactions with objects (laptops and remote control mock-ups), all factors which greatly affect people behaviors and their gaze (like bored people looking at table when two other people are discussing for a long time). As a measure of complexity, the VFOA target label set for a given person did not only comprise the three other persons in the meeting like in (Stiefelhagen et al., 2002, Otsuka et al., 2005), but also the slide screen and the table,[1] which accounted for around 50% of the labels. Thus, the task we addressed became much more complex, and required to study and account for more factors than done in the past.

VFOA recognition from head pose alone. Following others, we initially relied on a simple Hidden Markov Model (HMM) applied to each individual person to decode the sequences of head poses in terms of sequence of VFOA labels. There, the HMM dynamic was mainly imposing the continuity of the VFOA sequence label, while the likelihood of the observation (pan and tilt angle of the head pose) for each given VFOA label was modeled with a Gaussian distribution. One important difficulty was the setting of the means of these Gaussians, i.e., the expected head poses for looking at different VFOA target (in our meeting scenario 32 of them should be defined). Previous work often relied on manual setting, potentially followed by adaptation (Otsuka et al., 2005). Using training data is not really an option since VFOA annotation is difficult, time consuming, and data needs to be gathered and annotated for each configuration of the observer, targets and setting (camera position). To address this cumbersome issue, we proposed a methology that exploited results on human gazing behavior and head-eye dynamics involved in saccadic gaze shifts (Langton et al., 2000, Hanes and McCollum, 2006) to define a model that automatically determines which head poses should be associated with looking at a given target. It proved to be as effective as using training data in our set-up.

Contextual multimodal VFOA recognition. Head poses can not replace gaze. They are ambiguous: in realistic scenarios, the same head pose can be used to look at different targets, depending on the situation. To address this issue, researchers have proposed to exploit other cues to remove the ambiguity and favor some VFOA targets over others dependending on the context. Following this approach, we have proposed Dynamic Bayesien model extensions to the HMM to model the interactions between people's VFOA, head poses, as well as contextual cues relating to human communication and group activity. An example is shown in Figure 12.2, and allows to model the following socially grounded aspects of the interplay between these variables. First, the VFOA dynamics of individuals should not be treated separately but jointly, for instance to account for the fact that people tend to share the same focus in conversation or group activity. Secondly, the interaction between speaking patterns (called conversational regimes) and the gaze of people should be taken into account, as people usually express their attention to speakers by turning their head towards the speakers, with such effect being more prominent at dialog transitions when speakers exchange floors. Finally, in many human interaction situations, objects of interest attract people visual attention, thereby overruling the trends for eye gaze behavior observed in 'pure' human-human conversations, a process

[1] This label was also used when people looked at objects on the table like laptops.

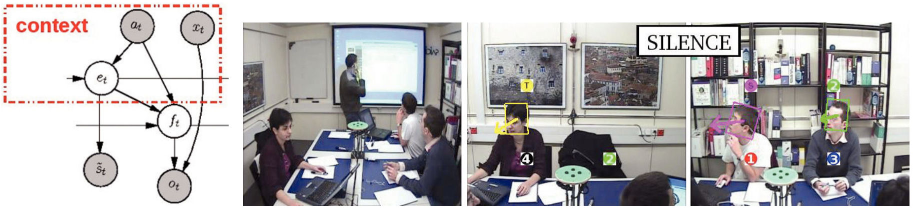

Figure 12.2 Left: One time slice of the Dynamic Bayesian network model for VFOA contextual recognition, expressing (with arrows) the probabilistic relationships between random variables (taken from Ba and Odobez (2011)). Qualitatively, the model expresses that the head pose of all people at time t (variable o_t) is a function of their VFOA f_t and of their location x_t, while their speaking status $\tilde{s}_t$ is determined according to the conversational contextual events e_t that identifies the set of people holding the floor. In addition, the upper part of the graph also represents the dependency of the VFOA states to the context: people VFOA is affected by the conversational event e_t, but this effect is modulated by the task context defined by a presentation activity variable a_t denoting the time since the last displayed slide. Right: sample result of a dynamic meeting. The box surrounding people's head and the arrows denotes the head location and pose estimated using our tracking system. Their color gives the estimated focus of the corresponding person VFOA, which is further stressed by the tag above the person's head. On the body of each person, a tag gives his seating location. The conversational regime (here, silence) is overlaid.

called the *situational attractor* hypothesis. In the case of meeting, this is notably the case of presentations, in which people dominantly look at the screen rather than at the speaker. Our model accounted for this by introducing a presentation status variable representing the time that ellapsed since the last slide change occurred, and which modulated in a timely manner the influence of the conversation on people VFOA.

Results. The results were produced on the 5 hours of meeting data. They showed the benefit of the different model contributions presented above. Using only the head pose, a VFOA frame recognition rate of aroud 40% was achieved. In contrast, the full contextual model resulted in a recognition rate of more than 55%. Importantly, it was shown that unsupervised adaptation of the VFOA target-head pose model parameters during run-time was greatly increasing the performance, a process in which the context played an important role by providing reliable soft labels for adaptation. For instance, during presentation one can reasonably assume people look at the slide and implicitly use this information to learn appropriate parameters.

12.2.3 VFOA recognition for wandering people

Previous models addressed indoor and meeting setups, by recognizing the VFOA of seated people (but potentially looking at standing people). However, one could also be interested in analyzing the VFOA of people wandering

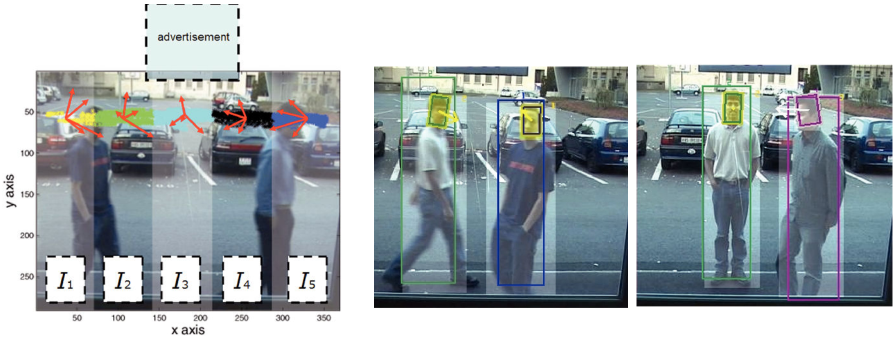

Figure 12.3 Left: wandering focus of attention (WFOA) modeling for the poster setup. The goal is to identify people looking at the advertisement. Head pose pointing vectors (red arrows) associated with looking at the poster are shown in each of the discrete space region. Right: sample results. White and yellow heads indicate a person recognized as looking or not at the poster.

freely in an open space, which could reveal information about space usage and people behaviors. Smith et al. (2008) addressed such a situation, by considering an advertising effectiveness scenario where the goal is to evaluate how many people were exposed to a poster, how many of them looked at it, and for how long. Figure 12.3 illustrates the setup that was selected, but it easy to imagine similar cases, notably to identify people looking at large flat screeny ubiquitous in many environments.

The motion of people adds several difficulties to the VFOA problem. First, robust multi-object tracking need to be solved so as to individuate each people and not count the same person twice as looking at the advertisement. In addition, tracking the head of people (and its orientation) is more difficult due to people motion, and potential occlusions. Smith et al. (2008) addressed this tracking problem using a multi-object state space formulation in a Bayesian framework that jointly estimated the number of people as well as their body location, head location, and head pose. The model considered the prior probabilities of object interactions, for instance to avoid two trackers occupying the same space, as well as prior probabilities of the configuration of body and head locations. Inference was solved through an efficient Reversible-Jump Markov Chain Monte Carlo technique.

Secondly, as the head is moving, interpreting a head pose as looking at the poster requires a position dependent gaze model. The issue was solved using a two-layer regression approach where the location (horizontal position) was discretized into several regions in which a specific mixture of Gaussian head pose model for looking at the poster was learned, as illustrated in the left image of Figure 12.3. Then, recognizing whether a person is looking at the poster or not was conducted through interpolation of these models based on the person position. Despite this simple modeling, the paper demonstrated the validity of the approach and the feasibility of the task on the setup. Images on the right of Figure 12.3 show some result examples.

12.2.4 Some perspectives on VFOA analysis

The VFOA research conducted within the IM2 context generated good interests in the scientific community, and demonstrated the need for developing gaze extraction and analysis tools in numerous application domains. For instance, the work on the wandering VFOA was further studied in the FP7 european project VANAHEIM, in which one of the main research topic dealts with human-centered cue extraction in the surveillance domain. There we showed for instance that head pose estimation in low-resolution images could leverage on several factors, like the coupling between head and body orientation and the online joint adaptation of the body and head pose regressors (Chen and Odobez, 2012).

VFOA for robots. Due to its role as attentional cue and floor control cue, VFOA plays an important role in HCI or HRI. It is one of the research topics addressed in the FP7 European project HUMAVIPS whose main goal is to endow humanoid robots with social interaction capabilities in the presence of multiple people. While VFOA in HRI shares many common elements with the meeting scenario case, it also has its specificities. Perception is done from sensors placed on the robot. Continuous monitoring of people and environment requires the robot to move the head or body and causes visibility interruptions or image blur, making the tracking and head pose extraction much more difficult. As another consequence, VFOA reasoning has to be conducted with only partial information about the environment. Also, since people are free to move, their body orientations, which physically constrain the head motion, vary more than in seated situations. Therefore, it more significantly affect the interpretation of head pose as looking at VFOA target and thus requires to be estimated explicitly or implicitly and added explicitly into the gaze model, as was done by Sheikhi and Odobez (2012). However, although it can improves, it also further complexifies the model and makes robust parameter setting more difficult. On the positive side, since the robot is part of the interaction, all its knowledge about the interaction (when it speaks, who it addresses, when it points or makes references to environmental objects, etc) can be directly exploited as context to reduce ambiguities, in manners similar to what was done in the meeting case (cf Fig. 12.2 left).

Kinect sensing. The advent of cheap depth cameras like Kinect changed the HRI research landscape in recent years. While this is particularly true for articulated human body sensing, it also concerns VFOA analysis. First, using 3D face morphable models, several works like (Funes and Odobez, 2012) demonstrated that very accurate head pose, less than 1 to 2 degree errors, could be obtained in the working conditions of these devices. This leads to both an ease of deployment of gaze analysis systems and to a significant increase of VFOA recognition rates, since VFOA performance and head pose errors were highly correlated (Ba and Odobez, 2009). Secondly, while the depth information can be used to infer head pose, it simultaneously provides a rather accurate eye localization. This effort can be used to crop the eye region in the standard RGB image to infer the true gaze, i.e. the gaze as inferred by eye/pupil direction. By learning an appearance gaze model for the frontal head pose and rendering the eye region for any other pose as if it was seen from this frontal pose, Funes and Odobez (2012) were able to obtain gaze estimates under free head movements with a gaze angular error ranging from 5 to 12 degrees. Examples of results are shown in Figure 12.4. While there is still room for research and improvements, in particular in automatizing individual model calibration steps, there is no doubt that such sensing devices will lead to more widespread exploitation of gaze in the HCI and HRI domains.

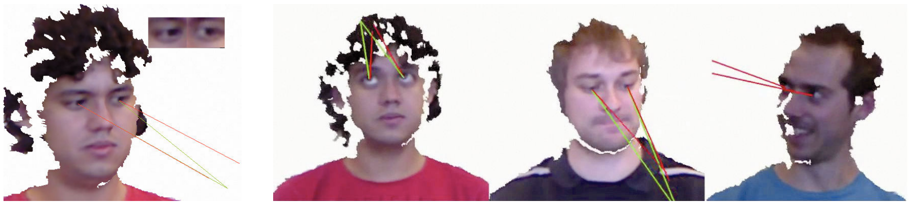

Figure 12.4 Left: 3D rendered mesh from depth and RGB image. Left top: using the 3D tracker output, eye regions from the RGB image can be cropped and rendered as if the head pose was frontal. Right: example of recognized gaze for different individuals under free head movements. Green lines (when available) materialize the ground-truth gaze directions, while the red ones represent estimated gaze.

12.3 Social signal processing

Figure 12.5 shows the number of scientific events (workshops, summer schools, symposia, etc.) that hold the word "*social*" in their title since the beginning of IM2 (source *dbworld*). Socially oriented topics were not popular in computing in the early 2000s, but in the following years they attracted an amount of attention and efforts that still today keeps growing. IM2 contributed to one of the many areas falling behind by the word "social" (social network analysis, social media, socio-technical systems), namely the automatic analysis of human-human communication, realizing the potential of nonverbal communication as an input "feature" for socially intelligent machines.

In this context, IM2 researchers contributed to define the research vision behind Social Signal Processing (SSP), the domain aimed at modeling, analysis and synthesis of nonverbal behavior in social interactions. These efforts resulted in several survey papers (Vinciarelli et al., 2008a,b, 2009b, 2012b), and also in a European Network of Excellence, the SSPNet[2] (Social Signal Processing Network), aimed at fostering an international SSP research community.

The key idea of SSP is that people interpret, typically unconsciously, nonverbal communication in terms of "*social signals*" Pentland (2007), i.e. relational attitudes that people display to one another. Hence, detecting nonverbal

[2] http://www.sspnet.eu

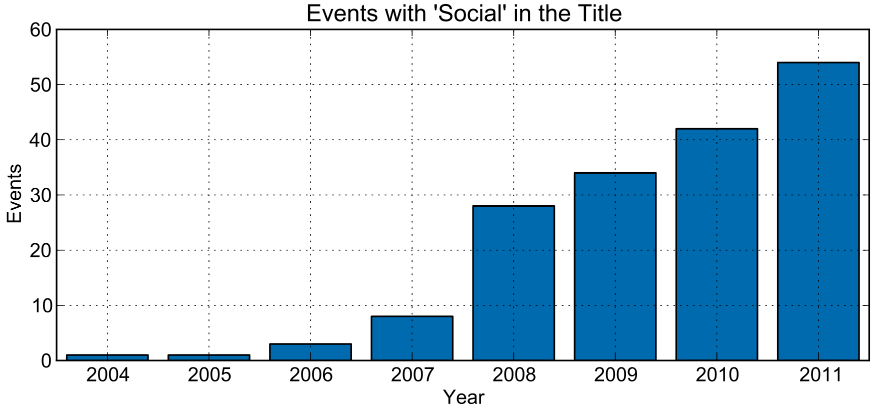

Figure 12.5 Number of events with the word "social" in their title (workshops, conferences, etc.) advertised via "dbworld".

cues means to detect the relational attitudes and, in ultimate analysis, to understand the social phenomena underlying an observed interaction. The rest of this section shows how such a paradigm, repeatedly applied in the framework of IM2, has worked for the analysis of three important phenomena: roles, personality and conflict. All works that will be discussed were carried out in the framework of IM2 and, in the last two cases, in the framework of IM2.SSP, the Individual Project dedicated to Social Signal Processing.

12.3.1 Role recognition

Roles fulfill two major functions in social interactions (Scott and Marshall, 2005): the first is to shape expectations about behavior of both others and ourselves, the second is to make the behavior of interaction participants more predictable, a necessary prerequisite towards smooth social exchanges. As a consequence, roles induce "*characteristic behavior patterns*" (Biddle, 1986), possibly machine detectable, that can be accessed not only via sociological inquiry, but also through technological investigation (Salamin and Vinciarelli, 2012).

IM2 work in this domain focused on the influence that speaker roles have on the organization of speaker turns – who speaks when, how much and with whom – one of the most salient features of every conversation, well known to account for social aspects of underlying interactions (Sacks et al., 1974, Bilmes, 1988). The earliest experiments focused on roles like *Anchorman* or *Weatherman* in radio news. The original aspect of the initial approaches was the extraction of social networks from speaker adjacency information available after speaker diarization (Vinciarelli, 2007). This made it possible to represent speakers portrayed in broadcast material with features typical of social network analysis (e.g., centrality, in- and out-degree, etc.). While being relatively simple, such features proved to be effective in the recognition of broadcast roles, not only when applied to clean data (turn organization extracted manually), but also when applied to noisy ones (turn organization extracted automatically). Overall, the role recognition rate in terms of frame accuracy was higher than 80%.

The results above encouraged the application of similar approaches to more challenging data, including talk-shows (where the roles are the same as those of the news) and meetings (where the roles correspond to different positions in a company). The simple approach proposed by Vinciarelli (2007) was no longer sufficient and new models, the Social Affiliation Networks, had to be introduced. This resulted in frame accuracies higher than 80% for talk-shows and higher than 45% for meetings (Salamin et al., 2009). For this latter scenario, the approach was further strengthened by modeling lexical choices, i.e. by applying text categorization techniques (each role corresponding to a different category) to the automatic transcriptions of what people say (Garg et al., 2008).

Two main problems were left open by the works above. On one hand, each subject was assigned only one role and it was not possible, for the same subject, to play more than one role in the same conversation. On the other hand, roles like those considered above were scenario dependent and did not allow any generalization. Hence, later role recognition approaches introduced sequential models (Conditional Random Fields and Hidden Markov Models) capable of assigning a role to each turn rather than to each subject. In this way, the same subject was allowed to play more than one role per conversation. Furthermore, it was possible to use not only turn organization, but also other behavioral cues such as prosody and dialogue acts. This allowed the performance to be improved (up to 90% frame accuracy on broadcast data) like in Salamin and Vinciarelli (2012), and also to consider sociology inspired roles (Bales, 1950) like "*neutral*", "*attacker*" or "*gatekeeper*" that apply to any possible interaction (Valente and Vinciarelli, 2011, Valente et al., 2011).

The IM2 work on role recognition was the first extensive investigation of the problem, covering several settings and taking into account different role sets. Nowadays, role recognition is a problem addressed by other authors in the literature (Vinciarelli et al., 2012b).

12.3.2 Automatic personality perception

Personality is the latent construct that accounts for "*individuals' characteristic patterns of thought, emotion, and behavior together with the psychological mechanisms - hidden or not - behind those patterns*" (Funder, 2001). Automatic Personality Perception (APP) is the prediction of personality traits that people attribute to others they observe. The SSP paradigm, aimed at inferring socially relevant information from nonverbal cues, appears thus to be particularly suitable to APP.

The first APP effort in IM2.SSP was the collection of a corpus of speech samples annotated in terms of personality. At the time this article is being written, the dataset is one of the largest of its type, in terms of both number of samples (640) and, most importantly, number of subjects (322). Each sample is a 10 seconds long speech segment where only one person talks. The annotators, 11 in total, have listened to all clips of the corpus and, for each of them, they have filled the BFI-10 questionnaire (Rammstedt and John, 2007), a personality assessment instrument that assigns each subject five scores corresponding to the Big Five (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism), the five traits capturing most of the individual differences (Saucier and Goldberg, 1996).

The first experiments were aimed at predicting whether one person was perceived to be "*low*" (below median) or "*high*" (above median) along each of the five traits (Mohammadi and Vinciarelli, 2012). The performance ranged between 60 and 75% depending on the traits and the difference with respect to chance was, in all cases, statistically significant. Both prediction results and measurements about the most influential features were in agreement with the psychological literature: the traits recognized with higher performance were *extroversion* and *conscientiousness*, well known to be the most accessible ones whenever we meet a person for the first time. Furthermore, variability of the main prosodic features (pitch, loudness and tempo) was shown to be the most important factor in personality perception.

Both data and experimental protocol proposed by Mohammadi and Vinciarelli (2012) were adopted in a 2012 international benchmarking campaign[3] (Schuller et al., 2012). The challenge involved 54 participants and the results show that the best performances were obtained by taking into account the gender of the speaker, a characteristic that seems to dominate the attribution of personality traits.

The latest experiments are no longer aimed at predicting the traits, but rather at predicting how different individuals are ranked in terms of personality traits. This makes more sense from a psychological point of view because personality captures differences between people more than individual characteristics. Preliminary results show that ordinal regression approaches based on Gaussian Processes achieve performances statistically significantly higher than chance when ranking up to 6 different persons (Mohammadi et al., 2012). Current work aims at addressing this problem with a fully Bayesian treatment of an ordinal regression model. In this way, the model parameters can show what are the speech features that influence most the way listeners rank voices in the personality space.

12.3.3 Conflict detection

Conflict is an important event in the life of a group, because it can have disruptive effects, including the destruction of the group itself (Levine and Moreland, 1998). In scientific terms, conflict is a mode of interaction that takes place whenever two or more parties pursue individual, incompatible goals in

[3] `http://emotion-research.net/sigs/speech-sig/is12-speaker-trait-challenge`

the same setting (Allwood, 2007). Like any other social phenomenon, conflict leaves traces in nonverbal communication and, hence, it is suitable for the application of SSP approaches.

Like for roles and personality, the first step in IM2 efforts was the collection of appropriate data and the attention focused on Canal9, a corpus of television political debates broadcast in Switzerland (Vinciarelli et al., 2009a). The reason is that debates are built around conflict: if one party gets elected, the other does not, if one party acquires consensus, the other looses it, and so on. Hence, debates are a typical case where several parties pursue incompatible goals in the same setting.

Earliest IM2 works were aimed at the simple detection of conflict, i.e. at giving a binary answer about the presence (or absence) of conflict in a given debate segment. The approaches where developed in collaboration with institutions outside IM2 (University of Verona and Italian Institute of Technology) and were based on the key-concept of "*Steady Conversational Period*" (SCP), originally proposed by Cristani et al. (2011). The idea is that conversations can be represented with a finite set of configurations automatically detectable in the data (e.g., one person is silent and the other talks, nobody talks, everybody talks at the same time, etc.). Conflict tends to be associated to certain SCPs rather than others and this simple consideration allows one to predict correctly up to 80% of the times whether a conversation segment is conflictual or not (Pesarin et al., 2012).

However, while being intuitive and effective , such an approach seems to contradict the idea that conflict is always present in a political debate for the very simple fact that involved parties pursue incompatible goals (see above). The reason is that conflict can have different intensity and the approach proposed by Pesarin et al. (2012) detects intensity peaks or, at least, time intervals when the intensity goes above a certain threshold. For this reason, it was necessary to devise a different way of measuring the conflict and, correspondingly, of collecting and annotating the data (Vinciarelli et al., 2012a).

After a review of the literature on nonverbal correlates of conflict, a questionnaire was setup aimed at matching observable behavior and perceived conflict. Then, each debate of the Canal9 corpus was split into 30 seconds long, non-overlapping segments. Only the segments showing at least two persons were retained resulting into 1430 samples (roughly 12 hours of material). The clips were then distributed, via Mechanical Turk, to roughly 600 annotators that have filled the questionnaire for each of the clips. As a result, each clip was associated to two scores, one accounting for the frequency of certain behavioral cues (fidgeting, loud speaking, interruptions, etc.) and one accounting for the perceived level of conflict. Since the correlation between the two scores was higher than 0.95 (p-value = 10^{-12}), it was possible to establish a clear relationship between the frequency of certain cues and the level of conflict.

At this point, after developing a feature extraction process capable of detecting the cues and measuring their frequency, the application of regression approaches based on Gaussian Processes allowed the prediction of the conflict level. The correlation between predicted level and level assigned by the annotators was close to 0.8. Such an approach provides a finer measurement of confict and, in particular, allows one to use a continuous measurement rather than a discrete decision (Kim et al., 2012).

12.4 Behavioral analysis of video blogging

The availability of large-scale conversational data in social media sites like YouTube, and the connections between these new forms of interaction with the previous work in IM2 were the initial motivations for the work on analysis of video blogs (vlogs) started in 2009. This research applies much of what has been learned in IM2 in terms of methods for automatic perception and interaction modeling and at the same time expands the IM2 vision by weaving it with new trends in human communication and computing.

While nonverbal communication is a classical area (Knapp and Hall, 2005), its study in the context of social media is much more recent, enabled by the global success of networking sites like Facebook. In this context, the work by Biel and Gatica-Perez was the first to examine the vlogging setting, and investigated three issues. First, methods to automatically extract nonverbal cues from real vlogs were assessed. Second, a set of possibilities to characterize social perception of vloggers on social media sites were studied. Finally, potential connections between the two above aspects (i.e., how nonverbal cues explain a number of perceived social variables like social attention and personality trait impressions) were investigated. Each of these issues is discussed in the rest of this section.

12.4.1 Extracting nonverbal communicative cues from vlogs

Vloggers are expressive, communicating a wealth of information through voice, face, and body (see Fig. 12.6). In this video genre, challenges also abound, as a variety of non-conversational content is present (including editing artifacts like openings, closings, and music), video quality varies, the camera might move, the number of people in front of the camera might vary, and so on. A variety of methods to extract some of these features were developed. In a first approach (Biel and Gatica-Perez, 2010b), features from audio were extracted, including speaking-turn related features (speaking time, number of speaking turns) and prosody features (energy, speaking rate). In a subsequent study, visual cues were also studied, including coarse measures of face motion and visual attention, as well as person framing with respect to the camera (Biel and Gatica-Perez, 2010a). Later studies have integrated improved methods to estimate measures of human motion (Biel et al., 2011) as well as facial expressions of emotion (Biel et al., 2012). A number of audio-visual features, inspired by the literature on coordination of looking and speaking in communication, has also been studied (Biel and Gatica-Perez, 2010a, 2011, 2013). While each nonverbal behavior extractor has its own limitations, the current set of features allows for a characterization of vloggers that is amenable for further investigations about social perception in social media.

Figure 12.6 Vlogs represent rich communicative experiences in social media (photo credit: Joan-Isaac Biel).

12.4.2 Characterizing social perception in vlogging

Social media audiences make impressions of the people they watch. These perceptions can be studied in at least two ways. The first one is indirect, as it comes from analyzing the metadata traces that audiences leave in the channels they watch, in the form of views, comments, ratings, etc. The second way of studying social perception is direct, by asking explicit questions to viewers about the impressions they form of vloggers. A highly suitable mechanism to obtain direct data comes from the integration of video - already available on the social media site - with crowdsourcing platforms like Amazon Mechanical Turk.

Both ways of obtaining information about social perception have been studied. Unlike classical work in social psychology, where judgments from observers are typically collected in a laboratory setting, our work has entailed challenges associated to the uncertainty and possibly lower quality of the information collected through social metadata and crowdsourcing. In practice, our studies have found these methodologies to be reliable enough (after applying control mechanisms) to characterize popular and non-popular users (Biel and Gatica-Perez, 2011), as well as a small number of individual variables including Big-Five personality traits, mood, and attractiveness (Biel et al., 2011, Biel and Gatica-Perez, 2012, 2013). These studies were conducted using data from about 450 YouTube vloggers and involved over a hundred workers on Mechanical Turk. Overall, our studies confirmed what other authors have found in psychological research regarding data quality obtained via crowdsourcing (Buhrmester et al., 2011), while adding the new angle of crowdsourcing social impressions from video.

12.4.3 Investigating connections between nonverbal behavior and social perception

The final component of our work relates to the study of possible connections between automatically extracted behavioral cues and social perception constructs. While the psychology literature provides a wealth of results regarding links between nonverbal behavior and impressions of variables like personality in lab settings and everyday life (Knapp and Hall, 2005), fundamentally less is known about how these processes develop in social media, and in online conversational video in particular.

Our work in this direction resulted in two main findings. First, on a study using 2200 vlogs from YouTube, significant correlation was found between specific nonverbal behaviors extracted from audio and video (including speaking time, looking time, and joint looking/speaking features) and the average level of attention that these vlogs received as measured by their log-number of views (Biel and Gatica-Perez, 2010a, 2011). This result suggests that certain nonverbal behaviors are, on average, observed more often in video bloggers who receive larger levels of attention (possibly mediated by a number of variables), but of course no causal links are implied.

Second, in three studies using data from 440 vloggers, it was found that a variety of nonverbal cues have correlation with Big-Five personality impressions obtained via crowdsourcing (Biel et al., 2011, Biel and Gatica-Perez, 2013, Biel et al., 2012). More specifically, Extraversion was the trait that reached the highest agreement across Mechanical Turk workers, the largest cue utilization (i.e., the largest number of behavioral cues with significant effects), and the most accurate predictions. In contrast, the other Big-Five traits (Agreeableness, Conscientiousness, Openness to Experience, and Emotional Stability) were more challenging to predict, and Emotional Stability reached the lowest human agreement and was not captured by the extracted cues, highlighting that overall the problem is challenging.

As a final note, very recent work using manual speech transcriptions of the vlog data has shown that the words spoken in vlogs are useful for the personality prediction task, specifically for some of the traits not captured by the nonverbal cues. Although IM2 research has contributed to transfering speech technologies to the market (Koemei, *www.koemei.com*), one key challenge to the use of spoken words is the fact that automatic speech recognition in unconstrained internet video is still a difficult task.

12.5 Final remarks

This chapter has provided a brief reflection of 12 years of IM2 research on nonverbal behavior analysis, focusing on recent developments along three research lines: modeling of visual focus of attention from visual and multimodal cues (as a valuable cue in itself and as part of conversation); modeling of a number of situations in face-to-face interaction (roles, personality, conflict) from audio cues; and behavioral analysis in online conversational video (vlogging) using multimodal cues. Looking beyond the methods developed and the progress achieved, it is reasonable to expect that the near future will bring additional means to extract behavioral cues (e.g., via more powerful sensors and additional modalities beyond audio and video) and to enable the study of a wider variety of situations involving interacting people or people and machines. A key challenge for the future will be the integration of all this new information spread over data streams, time, and people – in other words, a renewed version of one of the original IM2 objectives.

Acknowledgments

The research summarized here has involved a large number of colleagues at Idiap funded by IM2 and other projects over the last 12 years. In alphabetical order, the list includes O. Aran, S. Ba, S. Bengio, J.-I. Biel, S. Favre, K. Funes, H. Hung, D. Jayagopi, S. Kim, G. Lathoud, I. McCowan, G. Mohammadi, B. Raducanu, H. Salamin, D. Sanchez-Cortes, S. Sheikhi, K. Smith, F. Valente, and D. Zhang. In addition to IM2, the work described here has also been partly funded by other sources, including the EU AMI, AMIDA, SSPNet, HUMAVIPS, and VANAHEIM projects, and the US VACE program.

Bibliography

Al-Hames, M., Dielmann, A., Gatica-Perez, D., Reiter, S., Renals, S., Rigoll, G., and Zhang, D. (2005). Multimodal integration for meeting group action segmentation and recognition. In *in Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinbugh.

Allwood, J. (2007). Cooperation, Competition, Conflict and Communication. *Gothenburg Papers in Theoretical Linguistics*, 94:1–14.

Aran, O. and Gatica-Perez, D. (2010). Fusing audio-visual nonverbal cues to detect dominant people in group conversations. In *in Proc. Int. Conf. on Pattern Recognition (ICPR)*, Istanbul.

Ba, S. and Odobez, J. (2006). A study on visual focus of attention recognition from head pose in a meeting room. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Washington DC.

Ba, S. and Odobez, J.-M. (2008). Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Las-Vegas.

Ba, S. and Odobez, J.-M. (2009). Recognizing human visual focus of attention from head pose in natural meetings. *IEEE Trans. on System, Man and Cybernetics: part B, Cybernetics*, 39(1):16–34.

Ba, S. and Odobez, J.-M. (2011). Multi-person visual focus of attention from head pose and meeting contextual cues. *IEEE Trans. on Patt. Anal. and Machine Intelligence (PAMI)*, 33(1):101–116.

Ba, S. O. and Odobez, J. M. (2005a). Evaluation of head pose tracking algorithm in indoor environments. In *International Conference on Multimedia & Expo, ICME 2005, Amsterdam.*

Ba, S. O. and Odobez, J. M. (2005b). A rao-blackwellized mixed state particle filter for head pose tracking. In *ACM-ICMI Workshop on Multi-modal Multi-party Meeting Processing (MMMP), Trento Italy*, pages 9–16.

Bales, R. (1950). A set of categories for the analysis of small group interaction. *American Sociological Review*, 15(2):257–263.

Biddle, B. (1986). Recent developments in role theory. *Annual Review of Sociology*, 12:67–92.

Biel, J.-I., Aran, O., and Gatica-Perez, D. (2011). You are known by how you vlog: Personality impressions and nonverbal behavior in YouTube. In *Proc. AAAI Int. Conf. of Weblogs and Social Media (ICWSM).*

Biel, J.-I. and Gatica-Perez, D. (2010a). Vlogcast yourself: Nonverbal behavior and attention in social media. In *Proc. Int. Conf. of Multimodal Interfaces (ICMI-MLMI).*

Biel, J.-I. and Gatica-Perez, D. (2010b). Voices of vlogging. In *Proc. AAAI Int. Conf. of Weblogs and Social Media (ICWSM).*

Biel, J.-I. and Gatica-Perez, D. (2011). Vlogsense: Conversational behavior and social attention in youtube. *ACM Transactions on Multimedia Computing, Communications*, 7(1):33:1–33:21.

Biel, J.-I. and Gatica-Perez, D. (2012). The good, the bad, and the angry: Analyzing crowdsourced impressions of vloggers. In *Proc. AAAI Int. Conf. of Weblogs and Social Media (ICWSM).*

Biel, J.-I. and Gatica-Perez, D. (2013). Ieee trans. on multimedia. *The YouTube Lens: Crowdsourced Personality Impressions and Audiovisual Analysis of Vlogs*, 15(1):41–55.

Biel, J.-I., Teijeiro, L., and Gatica-Perez, D. (2012). Facetube: Predicting personality from facial expressions of emotion in online conversational video. In *Proc. ACM Int. Conf. of Multimodal Interaction (ICMI)*.

Bilmes, J. (1988). The concept of preference in conversation analysis. *Language in Society*, 17(2):161–181.

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science*, 6(1):3.

Chen, C. and Odobez, J.-M. (2012). We are not contortionists: coupled adaptive learning for head and body orientation estimation in surveillance video. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Providence*.

Cristani, M., Pesarin, A., Drioli, C., Tavano, A., Perina, A., and Murino, V. (2011). Generative modeling and classification of dialogs by a low-level turn-taking feature. *Pattern Recognition*, 44(8):1785–1800.

Funder, D. (2001). Personality. *Annual Reviews of Psychology*, 52:197–221.

Funes, K. and Odobez, J.-M. (2012). Gaze estimation from multimodal kinect data. In *CVPR Workshop on Face and Gesture and Kinect demonstration competition, Providence*.

Garg, N., Favre, S., Salamin, H., Hakkani-Tür, D., and Vinciarelli, A. (2008). Role recognition for meeting participants: an approach based on lexical information and Social Network Analysis. In *Proceedings of the ACM International Conference on Multimedia*, pages 693–696.

Gatica-Perez, D. (2009). Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing, Special Issue on Human Behavior*, 27(12).

Gatica-Perez, D., McCowan, I., Zhang, D., and Bengio, S. (2005). Detecting group interest-level in meetings. In *in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, USA.

Goodwin, C. and Heritage, J. (1990). Conversation analysis. *Annual Review of Anthropology*, pages 981–987.

Hanes, D. A. and McCollum, G. (2006). Variables contributing to the coordination of rapid eye/head gaze shifts. *Biol. Cybern.*, 94:300–324.

Hung, H. and Gatica-Perez, D. (2010). Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behavior. *IEEE Transactions on Multimedia*, 12(6):563 – 575.

Hung, H., Huang, Y., Friedland, G., and Gatica-Perez, D. (2011). Estimating Dominance In Multi-Party Meetings Using Speaker Diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):847 – 860.

Hung, H., Jayagopi, D., Ba, S., Odobez, J.-M., and Gatica-Perez, D. (2008). Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *Proceedings of the ACM International Conference on Multimodal interfaces (ICMI)*, pages 233 – 236, Chania, Greece.

Hung, H., Jayagopi, D., Yeo, C., Friedland, G., Ba, S., Odobez, J.-M., Ramchandran, K., Mirghafori, N., and Gatica-Perez, D. (2007). Using audio and video features to classify the most dominant person in a group meeting. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 835 – 838, Augsburg, Germany.

Jayagopi, D., Ba, S., Odobez, J., and Gatica-Perez, D. (2008). Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues. In *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI)*, pages 45 – 52, Chania, Greece.

Jayagopi, D. and Gatica-Perez, D. (2009). Discovering group nonverbal conversational patterns with topics. In *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI-MLMI)*, pages 3 – 6, Cambridge, USA.

Jayagopi, D. and Gatica-Perez, D. (2010). Mining group nonverbal conversational patterns using probabilistic topic models. *IEEE Transactions on Multimedia*, 12(8):790 – 802.

Jayagopi, D., Hung, H., Yeo, C., and Gatica-Perez, D. (2009a). Modeling dominance in group conversations using nonverbal activity cues. *Special issue on Multimodal processing in speech-based interactions, IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):501–513.

Jayagopi, D., Kim, T., Pentland, A., and Gatica-Perez, D. (2012). Privacy-sensitive recognition of group conversational context with sociometers. *Springer Multimedia Systems*, 18(1):3 – 14.

Jayagopi, D., Raducanu, B., and Gatica-Perez, D. (2009b). Characterizing Conversational Group Dynamics Using Nonverbal Behaviour. In *Proceedings of the International Conference on Multimedia and Expo (ICME)*, pages 370 – 373, NewYork, US.

Kalimeri, K., Lepri, B., Aran, O., Jayagopi, D., D., G.-P., and Pianesi, F. (2012). Modeling Dominance Effects on Nonverbal Behaviors using Granger Causality. In *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI)*, pages 23 – 26, Santa Monica, USA.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63.

Kim, S., Filippone, M., Valente, F., and Vinciarelli, A. (2012). Predicting the conflict level in television political debates: an approach based on crowdsourcing, nonverbal communication and gaussian processes. In *Proceedings of ACM International Conference on Multimedia.*

Knapp, M. L. and Hall, J. (2005). *Nonverbal communication in human interaction.* Holt, Rinehart and Winston, New York.

Langton, S., Watt, R., and Bruce, V. (2000). Do the eyes have it ? cues to the direction of social attention. *Trends in Cognitive Sciences*, 4(2):50–58.

Levine, J. and Moreland, R. (1998). Small groups. In Gilbert, D. and Lindzey, G., editors, *The handbook of social psychology*, volume 2, pages 415–469. Oxford University Press.

McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., and Zhang, D. (2005). Automatic Analysis of Multimodal Group Actions in Meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):305–317.

Mohammadi, G., Origlia, A., Filippone, M., , and Vinciarelli, A. (2012). From speech to personality: Mapping voice quality and intonation into personality differences. In *Proceedings of ACM International Conference on Multimedia.*

Mohammadi, G. and Vinciarelli, A. (2012). Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing*, 3(3):273–284.

Morimoto, C. and Mimica, M. (2005). Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding*, 98:4–24.

Otsuka, K., Takemae, Y., Yamato, J., and Murase, H. (2005). A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In *Proc. of ICMI*, pages 191–198.

Pentland, A. (2007). Social Signal Processing. *IEEE Signal Processing Magazine*, 24(4):108–111.

Pesarin, A., Cristani, M., Murino, V., and Vinciarelli, A. (2012). Conversation analysis at work: Detection of conflict in competitive discussions through automatic turn-organization analysis. *Cognitive Processing*, 13(2):533–540.

Rammstedt, B. and John, O. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, 41(1):203–212.

Sacks, H., Schegloff, E., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Salamin, H., Favre, S., and Vinciarelli, A. (2009). Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction. *IEEE Transactions on Multimedia*, 11(7):1373–1380.

Salamin, H. and Vinciarelli, A. (2012). Automatic role recognition in multiparty conversations: an approach based on turn organization, prosody and conditional random fields. *IEEE Transactions on Multimedia*, 14(2):338–345.

Sanchez-Cortes, D., Aran, O., Schmid Mast, M., and Gatica-Perez, D. (2012). A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups. *IEEE Transactions on Multimedia*, 14(3-2):816–832.

Saucier, G. and Goldberg, L. (1996). The language of personality: Lexical perspectives on the five-factor model. In Wiggins, J., editor, *The Five-Factor Model of Personality*.

Schuller, B., Steidl, S., Batliner, A., Noeth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., and Weiss, B. (2012). The interspeech 2012 speaker trait challenge. In *Proceedings of Interspeech*.

Scott, J. and Marshall, G., editors (2005). *Dictionary of Sociology*. Oxford University Press.

Sheikhi, S. and Odobez, J.-M. (2012). Investigating the midline effect for visual focus of attention recognition. In *ACM Int Conf. on Multimodal Interaction (ICMI)*.

Smith, K., Ba, S., Gatica-Perez, D., and Odobez, J.-M. (2008). Tracking the visual focus of attention for a varying number of wandering people. *IEEE Trans. on Pattern Analysis and Machine Intelligence,*, 30(7):1212–1229.

Stiefelhagen, R., Yang, J., and Waibel, A. (2002). Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Trans. on Neural Networks*, 13(4):928–938.

Valente, F. and Vinciarelli, A. (2011). Language-independent socio-emotional role recognition in the AMI meetings corpus. In *Proceedings of Interspeech*, pages 3077–3080.

Valente, F., Vinciarelli, A., Yella, S., and Sapru, A. (2011). Understanding social signals in multi-party conversations: Automatic recognition of socio-emotional roles in the AMI meeting corpus. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pages 374–379.

Vinciarelli, A. (2007). Speakers role recognition in multiparty audio recordings using Social Network Analysis and duration distribution modeling. *IEEE Transactions on Multimedia*, 9(6):1215–1226.

Vinciarelli, A., Dielmann, A., Favre, S., and Salamin, H. (2009a). Canal9: a database of political debates for analysis of social interactions. In *Proceedings of the Social Signal Processing Workshop*.

Vinciarelli, A., Kim, S., Valente, F., and Salamin, H. (2012a). Collecting data for socially intelligent surveillance and monitoring approaches: The case of conflict in competitive conversations. In *Proceedings of International Symposium on Communications, Control and Signal Processing*.

Vinciarelli, A., Pantic, M., and Bourlard, H. (2009b). Social Signal Processing: Survey of an emerging domain. *Image and Vision Computing Journal*, 27(12):1743–1759.

Vinciarelli, A., Pantic, M., Bourlard, H., and Pentland, A. (2008a). Social Signal Processing: State-of-the-art and future perspectives of an emerging domain. In *Proceedings of the ACM International Conference on Multimedia*, pages 1061–1070.

Vinciarelli, A., Pantic, M., Bourlard, H., and Pentland, A. (2008b). Social signals, their function, and automatic analysis: A survey. In *Proceedings of the ACM International Conference on Multimodal Interfaces*, pages 61–68.

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., and Schroeder, M. (2012b). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87.

Zhang, D., Gatica-Perez, D., Bengio, S., and McCowan, I. (2006). Modeling Individual and Group Actions in Meetings With Layered HMMs. *IEEE Transactions on Multimedia*, 8(3):509 – 520.