Chapter 26
Social Signal Processing: The Research Agenda

Maja Pantic, Roderick Cowie, Francesca D'Errico, Dirk Heylen, Marc Mehu, Catherine Pelachaud, Isabella Poggi, Marc Schroeder, and Alessandro Vinciarelli

Abstract The exploration of how we react to the world and interact with it and each other remains one of the greatest scientific challenges. Latest research trends in cognitive sciences argue that our common view of intelligence is too narrow, ignoring a crucial range of abilities that matter immensely for how people do in life. This range of abilities is called social intelligence and includes the ability to express and recognise social signals produced during social interactions like agreement, politeness, empathy, friendliness, conflict, etc., coupled with the ability to manage them in order to get along well with others while winning their cooperation. Social Signal Processing (SSP) is the new research domain that aims at understanding and modelling social interactions (human-science goals), and at providing computers with similar

M. Pantic (✉)
Computing Dept., Imperial College London, London , UK
e-mail: m.pantic@imperial.ac.uk

M. Pantic · D. Heylen
EEMCS, University of Twente, Enschede, The Netherlands

R. Cowie
Psychology Dept., Queen University Belfast, Belfast, UK

F. D'Errico · I. Poggi
Dept. Of Education, University Roma Tre, Rome, Italy

M. Mehu
Psychology Dept., University of Geneva, Geneva, Switzerland

C. Pelachaud
CNRS, Paris, France

M. Schroeder
DFKI, Saarbrucken, Germany

A. Vinciarelli
Computing Science Dept., University of Glasgow, Glasgow, UK
e-mail: vincia@dcs.gla.acvi.uk
IDIAP Research Institute, Martigny, Switzerland

T.B. Moeslund et al. (eds.), *Visual Analysis of Humans*,
DOI 10.1007/978-0-85729-997-0_26,

abilities in human–computer interaction scenarios (technological goals). SSP is in its infancy, and the journey towards artificial social intelligence and socially aware computing is still long. This research agenda is twofold, a discussion about how the field is understood by people who are currently active in it and a discussion about issues that the researchers in this formative field face.

26.1 Introduction

The exploration of how human beings react to the world and interact with it and each other remains one of the greatest scientific challenges. Perceiving, learning, and adapting to the world are commonly labelled as intelligent behaviour. But what does it mean being intelligent? Is IQ a good measure of human intelligence and the best predictor of somebody's success in life? There is now a growing research in cognitive sciences, which argues that our common view of intelligence is too narrow, ignoring a crucial range of abilities that matter immensely for how people do in life. This range of abilities is called social intelligence [1, 3, 17, 116] and includes the ability to express and recognise social signals like turn taking, agreement, politeness, empathy, friendliness, conflict, etc., coupled with the ability to manage them in order to get along well with others while winning their cooperation. There is no common definition for the concept of social signal (as explained in Sect. 26.2), and the definition that we adopt in this document refers to social signals as to signals produced during social interactions, that either play a part in the information and adjustment of relations and interactions between agents (human and artificial), or provide information about the agents. Social signals are manifested through a multiplicity of non-verbal behavioural cues including facial expressions, body postures and gestures, vocal outbursts like laughter, etc. (Fig. 26.1), which can be automatically analysed by technologies of signal processing (as discussed in Sect. 26.3), or automatically generated by technologies of signal synthesis (as talked about in Sect. 26.4).

When it comes to computers, however, they are socially ignorant [95]. Current computing devices do not account for the fact that human–human communication is always socially situated and that discussions are not just facts but part of a larger social interplay. However, not all computers will need social intelligence and none will need all of the related skills humans have. The current-state-of-the-art categorical computing works well and will always work well for context-independent tasks like making plane reservations and buying and selling stocks. However, this kind of computing is utterly inappropriate for virtual reality applications as well as for interacting with each of the (possibly hundreds) computer systems diffused throughout future smart environments (predicted as the future of computing by several visionaries such as Mark Weiser [128]) and aimed at improving the quality of life by anticipating the users needs. Computer systems and devices capable of sensing agreement, inattention, or dispute, and capable of adapting and responding in real-time to these social signals in a polite, non-intrusive, or persuasive manner, are likely to be perceived as more natural, efficacious, and trustworthy. For example,

Fig. 26.1 Manifestations of social signals include a variety of non-verbal behavioural cues including facial expressions, body postures/gestures, vocal outbursts like laughter, etc.

in education, pupils' social signals inform the teacher of the need to adjust the instructional message. Successful human teachers acknowledge this and work with it; digital conversational embodied agents must begin to do the same by employing tools that can accurately sense and interpret social signals and social context of the pupil, learn successful context-dependent social behaviour, and use a proper socially adept presentation language (e.g., see [93]) to drive the animation of the agent. The research area of machine analysis and employment of human social signals to build more natural, flexible computing technology goes by the general name of Socially Aware Computing as introduced by [94, 95].

Although the importance of social signals in everyday life situations is evident, and in spite of recent advances in machine analysis and synthesis of relevant behavioural cues like gaze exchange, blinks, smiles, head nods, crossed arms, laughter, expressive prosody, and similar [83, 89, 90, 112, 132], the research efforts in machine analysis and synthesis of human social signals like attention, empathy, politeness, flirting, (dis)agreement, etc., are still tentative and pioneering efforts. Nonetheless, the importance of studying social interactions and developing automated systems of social signals analysis from audiovisual recordings is indisputable. It will result in valuable multimodal tools that could revolutionise basic research in cognitive and social sciences by raising the quality and shortening the time to conduct research that is now lengthy, laborious, and often imprecise. The first results in the field attest that social interactions and behaviours, although complex and rooted

in the deepest aspects of human psychology, can be analysed automatically with the help of computers (for extensive overview of the past research in the field of automatic analysis of social signals, see [125]). In fact, the pioneering contributions in Social Signal Processing (SSP) [32, 60, 94] have shown that social signals, typically described as so elusive and subtle that only trained psychologists can recognise them [45], are actually evident and detectable enough to be captured through sensors like microphones and cameras, and interpreted through analysis techniques like machine learning and statistics. At the same time, and as outlined above, tools for social signal synthesis in Human–Computer Interaction (HCI) form a large step ahead in realising naturalistic, socially aware computing and interfaces, built for humans, based on models of human behaviour. For example, combining synthetic speech with laughter influences the perception of social bonds [119]. Similarly, facial expressions influence a human user's evaluation of an Embodied Conversational Agent [105]. Contingency of signals has a key role in creating rapport between human user and virtual agent [46]. Politeness cues [127] and empathic expressions [83] are perceived as more appropriate in interactive scenarios.

SSP [95, 96, 124, 125] is the new research and technological domain that aims at providing computers with the ability to sense and understand human social signals. SSP is in its initial phase and the first step is to define the field and discuss issues facing the researchers in the field. This article attempts to achieve this. In Sect. 26.2, an overview of the relevant terminology defined by the related human-science fields is provided. Next, in the absence of a uniquely accepted definition of social signals, a working definition of social signals is introduced. In Sects. 26.3 and 26.4, challenging issues facing researchers in automatic social signal analysis and synthesis are summarised. Section 26.5 summarises the key goals of the SSP research overall, lists a number of issues that are of importance for the field but are still debated, and discusses the relevant ethical issues. Section 26.6 concludes the paper.

26.2 Social Signals: Terminology, Definition, and Cognitive Modelling

In order to anchor their discipline in the rich conceptual background developed in the behavioural sciences, SSP researchers are faced with the difficult task of defining a theoretical framework within which they will research the phenomena social signals they are eager to automatically detect, interpret, and synthesise. The major issue here is the diversity of conceptual ideas proposed about social signals and behaviour. Disciplines that dealt with the study of human psychological phenomena (mental states and behaviour) developed a myriad of ideas, definitions, and methods for the study of the same subject, human communication. In itself, this may be seen as a strength more than a weakness because having multiple approaches increases the potential for a good understanding of the complexities of human behaviour. However, this diversity may become a problem when people of different traditions come to work together on interdisciplinary research topic (such as SSP).

The increasing specialisation that characterises most scientific disciplines can be a barrier to inter-disciplinarity, for it can hinder communication between scholars. For communication to be successful, one has to use terms that will be understood by various researchers working in the field. In an attempt to achieve this, we describe here the different approaches adopted by the human sciences to study social signals. Section 26.2.1 presents a (non-exhaustive) glossary of concepts generated by different fields (ethology, social psychology, linguistics, semiotics, ...) to study communication. We present commonalities and differences between these approaches, so that scholars who are not familiar with the different disciplines can have a clearer idea of what the different positions are. The goal of this exercise is not to create a common definition for the concept of social signal because it would deny the specificities of each field and would constitute, for some, a loss of conceptual clarity. As SSP is a multi-disciplinary venture, our aim is to avoid the creation of a monolithic view that is unlikely to be adopted by the scientific community at large, or that may block the development of new ideas or research projects. Instead, our goal is to expose the diversity and be aware of it. However, as a definition of the studied phenomena is needed, and in the absence of a uniquely accepted definition of social signals, Sect. 26.2.2 introduces a working definition of social signals.

26.2.1 Terminology

Tables 26.1, 26.2, 26.3 and 26.4 present a (non-exhaustive) glossary of SSP-relevant concepts generated by different fields (ethology, social psychology, linguistics, semiotics, etc.) to study communication.

The main stream of research in animal and human communication acknowledges that signals convey information and/or meaning to a receiver. Although this could be considered as a commonality between all approaches, a critical analysis of research findings and theoretical developments in ethology suggest that the principles that are applied to the study of human language should not necessarily apply to the study of animal signals [102] or to some aspects of human non-verbal behaviour [87]. In other words, the strong semantic component of human language may not necessarily be shared by other channels of communication. For this reason, the ethological definition of signals includes the possibility that they do not carry information.

Different disciplines adopt different ways of defining a signal. For example, ethologists define signals by their properties or nature, and by their function of influencing a perceiver's behaviour or internal state (e.g. [78]). On the other hand, [100] define a social signal by its content (signals that have a social content, like a social attitude, a social emotion, etc.). In social psychology, scholars tend to use the term indicator, sign, signal, and display interchangeably (e.g. [18, 68]), without specifying what they mean by the terms indicator, sign, signal, or display. We assume that this lack of specificity implies that these authors endorse a general dictionary definition of the word, their goal being to study the eliciting circumstances

Table 26.1 Definitions of important ethology concepts for social signal processing

Ethology	
Signal	An act or structure that affects the behaviour (or internal state) of another organism, which evolved because of that effect, and which is effective because the receiver's response has also evolved [78]. A signal may [78] or may not [102] convey reliable information
Cue	A feature of the world, animate or inanimate, that can be used by individuals as a guide to future action [54]
Display	Behaviour pattern that has been modified in the course of evolution to convey information [6]. Displays are usually constituted of several components, like cues and signals
Handicap	A signal whose reliability is ensured because its cost is greater than those required to efficiently convey the information [131]. The signal may be costly to produce, or have costly consequences [122]
Index	A signal whose intensity is causally related with the information that is being signalled and that cannot be faked [77]. Indices are equivalent to performance based signals [38]
Minimal-cost symbol	A signal whose reliability does not depend on its cost (different from a handicap) and which can be made by most members of a population (different from an index) [78]
Icon	A signal which form is similar to its meaning
Symbol	A signal whose form is unrelated to its meaning, e.g. conventional signal [48]

and information content of particular behaviour patterns rather than to develop different concepts for non-verbal communication. Other authors, however, decided to use other terms than signals to describe specific categories of non-verbal behaviour [37].

The variety in definitions may be partly explained by the use of different methodologies and the different empirical questions that have driven research activities in different fields. For example, ethology has always focused on the adaptive significance of behavioural patterns for the organisms displaying them and the selective pressures responsible for the evolution of signals [34, 57]. Psychological science, however, has always placed a greater interest in discovering the significance or meaning of a particular behaviour in the mind of perceivers (for a critic, see [87]). Finally, linguistic has been mostly preoccupied by the role of signals in the regulation of discourse and social interactions among members of conversational groups [31, 52]. The diversity in research methods and theoretical interests led scholars to use different terms to describe the same thing, or the same term to describe different ideas. By no means should this signify that one approach has more authority than the other, or that a research question is more relevant than another. The only drawback is that this state of affair may create confusion in scholars who are interested in social signals but are not familiar with the human and behavioural sciences. We hope that the overview provided here is helpful in that direction.

Table 26.2 Definitions of important psychology concepts for social signal processing

Psychology	
Cue	Stimulus which serves as a sign or signal of something else, the connection having previously been learned [130]
Indicator	No clear definition for non-verbal indicator, seems to be used in a loose fashion to reflect a connection between non-verbal behaviour and some underlying dimension
Signal	No precise definition of signal in social psychology, though some authors seem to imply that signals are intentionally communicative [36]. The category seems to include all non-verbal behaviours or morphological structure that convey information to a receiver [100]
Social signal	Communicative or informative signal that, either directly or indirectly, conveys information about social actions, social interactions, social emotions, social attitudes and social relationships [100]
Sign	Refers to an act that is informative but that was not necessarily produced to communicate information [36]
Emblem	Non-verbal act which has a direct verbal translation that is well-known by all members of a group, class, or culture [33, 37]
Illustrator	Movement directly tied to speech that illustrates what is said verbally [33, 37]
Regulator	Act that maintains and regulates the conversation between two or more individuals [37]
Manipulator	Act that represents adaptive efforts to satisfy bodily needs, actions, to manage emotions, to develop interpersonal contacts, or to learn instrumental activities (see also adaptor in [37])
Emotional expression	Non-verbal act that is specific to a particular emotion [35, 117], or to an underlying emotional dimension [110]
Distal cues	Externalisation of stable traits or transient states, can be motor expression or physical appearance [16, 109]
Proximal percept	Mental representation resulting from the perceptual process of distal cues [16, 109]

Although we can see that research domains mostly differ in the detailed elaborations they made with regards to the nature of signals, their function, and their informative value, a few features and principles used to describe signals are shared among the different fields. First, some acts are considered functionally or intentionally communicative (e.g. signals, emblems, communicative signals); whereas others are simply considered as informative (cues, signs, informative signals), suggesting that information can be derived from them although they have not evolved, or are not intended, for communication1. Most theories also recognise the existence of signals which meaning follows social conventions: symbols, conventional signals, and emblems. The iconic act also seems to meet agreement in the different fields, as it is defined by everyone as an act which meaning is defined by its form. Finally, the importance of multi-modality is also recognised by all fields of research [2, 4, 91]. Commonalities of this sort make collaborations between dis-

Table 26.3 Definitions of important concepts for processes involved in the production and perception of social signals

Processes involved in the production and perception of social signals	
Code	Principle of correspondence between the act and its meaning [37]. The code can be intrinsic, extrinsic, and iconic
Encoding	The process, taking place in the signaller, of relating the distal cue and its meaning. Transfer of information in one domain (e.g. thoughts, stances) to another domain (muscular contraction, blood concentration, ...)
Decoding	The process taking place in the perceiver of relating the proximal percept to a semantic category or some other form of representation
Linguistics and semiotics	
Turn taking	The order in which the participants in a conversation speak one after the other The fulfilment or violation of turn-taking rules in a conversation provides cues about its cooperative or competitive structure [31, 106]
Backchannel	Feedback and comments provided by listeners during face-to-face conversation, through short verbalisations and non-verbal signals, showing how they are engaged in the speakers' dialogue (HUMAINE glossary)

Table 26.4 Definitions of miscellaneous important concepts for social signal processing

Miscellaneous	
Context	All the cues present in the physical and social environment of a perceiver as well as perceiver's characteristics that surrounds the signal
Information (Information theory)	Any physical property of the world that reduces uncertainty in the individual that perceives it [114]
Meaning	The meaning of something is what it expresses or represents (Cambridge Advanced learner's dictionary)
Ground truth	A term, with origins in cartography and aerial imaging, used to describe data that can be taken as definitive, and against which systems can be measured. Its application to emotion is controversial, since it is highly debatable whether emotions as they normally occur are things about which we can have definitive knowledge (HUMAINE glossary)

ciplines possible and create bridges that are necessary for inter-disciplinary research.

26.2.2 Working Definition of Social Signals

As a definition of the studied phenomena is needed, and in the absence of a uniquely accepted definition of social signals, we provide here a working definition of what

'social signals' are. Social signal: Let us first define what a 'signal' is. A signal is a perceivable stimulus PS a behaviour, a morphological trait, a chemical trace produced by an Emitter E. The Emitter E can be an individual or a group of people, a virtual character, an animal, or a machine. The signal is received by some Receiver R, who may interpret the signal and draw some information I from it (the signal's meaning), whether E really intended to convey I or not. Taking this into account, we may define a 'social signal' as follows. A social signal is a signal that provides information about 'social facts', i.e., about social interactions, social emotions, social attitudes, or social relations.

We can further distinguish between informative and communicative signals. A communicative signal is a signal that the Emitter produces in order to convey a particular meaning (see the Speech Acts perspective: [24, 47, 99]), while an informative signal is a signal from which the Receiver draws some meaning even if the Emitter did not intend to convey it (see the Semiotic perspective: [92]). Let us explain these notions by means of an example.

Suppose that during a lunch break there is a group of children talking in a circle, where one of them is slightly outside of the circle. A prediction can be made that the child outside of the circle is at risk of being bullied or being dropped out from the group. The spatial positioning of children is a social signal that conveys information about the social relation between the child in question and the other children, without any of the children being aware that they convey this information. This signal is not a communicative signal, but an informative signal. Furthermore, a distinction can be made between direct and indirect signals. Since social signals are produced (and understood) in context, information coming from the context may combine with the literal meaning of the signals (for a study on the literal meaning of behavioural signals, see [99]) to introduce, through inferential processes, further 'indirect meanings' of the displayed signals that differ from context to context. Let us explain this by means of an example.

Suppose that two people, A and B, sit together and both appear to be sad. This is not a social signal, just the fact that both people express the same emotion. However, if by showing sadness A wants to tighten her bond with B, then her display of sadness is an information signal representing an indirect social signal of her bond to B. Taking these notions into account, we can redefine the definition of 'social signals' as follows. A social signal is a communicative or informative signal that, either directly or indirectly, provides information about 'social facts', that is, about social interactions, social emotions, social evaluations, social attitudes, or social relations.

Hence, we define social signals as communicative and informative signals that concern 'social facts', namely, social interaction, social emotions, social evaluations, social attitudes and social relations. However, there is no strict definition of these notions. In what follows, we propose tentative definitions of these notions. Social interactions: Social interaction is a specific event in which an agent A performs some social actions directed at another agent that is actually or virtually present. Social interactions may be mediated by communicative and informative signals. Typical communicative signals in social interactions are backchannel signals such

as head nods, which inform the recipient that her interaction partner is following and understanding her ([55]; Fig. 26.2).

Social emotions: A clear distinction can be made between individual and social emotions. The latter can be defined as an emotion that an Agent A feels toward and Agent B. Happiness and sadness are typical examples of individual emotions we can be happy or sad on our own; our feelings are not directed to any other person. On the other hand, admiration, envy, and compassion are typical examples of social emotions we have these feelings toward another person. Signals revealing individual emotions of a person and those communicating social emotions both include facial expressions, vocal intonations and outbursts, body gestures and postures, etc. However, if a behavioural cue like a frown is displayed as a consequence of an individual emotion, then this cue is a behavioural signal but not a social signal. It is a social signal only if it displayed in order to communicate a social emotion. In addition, a signal of empathy (e.g., patting a companion on the shoulder to convey that we share his sadness, Fig. 26.2) is a social signal. A typical signal associated with empathy is mimicry. However, mimicry is not always unconscious, which is typical for sincere empathy, but can be deliberately displayed in order to gain acceptance or approval. In the latter case, mimicry does not convey empathy. Studying the role and the effects of both deliberate and unconscious mimicry is a challenge facing the researchers in the field.

Social evaluation: Social evaluation of a person relates to assessing whether and how much the characteristics of this person comply with our standards of beauty, intelligence, strength, justice, altruism, etc. We judge other people because based on our evaluation we decide whether to engage in a social interaction with them, what types of social actions to perform, and what relations to establish with them. Typical signals shown in social evaluation are approval and disapproval, at least when it comes to the evaluator (e.g., Fig. 26.2). As far as the evaluated person is concerned, typical signals involve those conveying desired characteristics such as pride, self-confidence, mental strength, etc., which include raised chin, erected posture, easy and relaxed movements, etc.

Social attitudes: The notion of attitude has been widely investigated in Social Psychology. Social attitude can be defined as the tendency of a person to behave in a certain way toward another person or a group of people. Social attitudes include cognitive elements like beliefs, evaluations, opinions, and social emotions. All these elements determine (and are determined by) preferences and intentions [41].

Agreement and disagreement can be seen as being related to social attitudes. If two persons agree then this means that they have similar opinions, which usually entails an alliance, a commitment to cooperation, and a mutually positive attitude. In contrast, if two persons disagree, this typically implies conflict, non-cooperation, and mutually negative attitude. Typical signals of agreement and disagreement are head nods and head shakes, smile, lip wipe, crossed arms, wagging a hand, etc. [12, 13].

Fig. 26.2 'Social Facts' (from *top left, counter clock wise*): social emotions (compassion and empathy), social attitudes (approval and disapproval), social relations (dominance), and social relations (confederates)

Persuasion is also closely linked to social attitudes; it is a kind of social influence aimed at changing other people's attitudes towards a certain issue, by changing their opinions and evaluations about the target issue, and gaining agreement for the view he or she defends. Typical signals used in persuasion are persuasive words, gestures, gaze patterns, postures, as well as appropriate self-presentation aimed at eliciting the desired social evaluations. Social relations: A social relation is a relation between two (or more) persons in which these persons have common or related goals, that is, in which the pursuit, achievement, or thwarting of a goal of one of these persons determines or is determined in some way by the pursuit, achievement, or thwarting of a goal of the other involved person. Hence, not every relation is a social relation. Two persons sitting next to each other in a bus have a physical proximity relation, but this is not a social relation, although one can arise from it [19, 40]. We can have many different kinds of social relations with other people: dependency, competition, cooperation, love, exploitation, etc.

Exchange Theory [58, 67] has attempted to describe all relations including love and friendship in terms of costs and benefits. According to this theory, a person stays in a relation until it is a satisfying relation. The factors influencing this satisfaction are: rewards (material and symbolic rewards computed in terms of costs and benefits), evaluation of possible alternatives (that affects commitment), and investment

(of time, effort and resources). Several critics have challenged this view as being too close to classical utilitarianism, which does not account for the difference between material and symbolic rewards and rules out altruism [57]. Different typologies of relations have been proposed in terms of criteria like public vs. private, cooperation vs. competition, presence vs. absence of sexual relations, social-emotional support oriented vs. task oriented (e.g., [7]). However, defining the notion of social relation and drawing a typology of social relations, such that they are conceptually sound while being useful for analysis and understanding of social signals, is yet to be attained. Also, assessing how social relations, social attitudes, social emotion, and social interaction overall, affect subsequent social relations is another challenge facing the researchers in the field.

Social relations can be established not only with a single person, but with a group. Within group relations, particular challenges concern the definition and description of mechanisms of power, dominance, and leverage [22, 73]. This relates to: (i) the allocation, change, and enhancement of power relations (e.g., through alliance, influence, and reputation), (ii) the interaction between gender and power relations, and (iii) the nature of leadership and the role of charisma in it. Clearly, all these issues are context and culture dependent. Typical signals revealing social relations include the manner of greeting (saying 'hello' first signals the wish for a positive social relation, saluting signals belong to a specific group like the army), the manner of conversing (e.g., using the word 'professor' signals submission), mirroring (signalling wish to have a positive social relation, or displaying 'typical' group's behaviour), spatial positioning (e.g., making a circle around a certain person distinguishes that person as the leader, touching another person indicates either affective relation or dominance, e.g., Fig. 26.2), etc. For group relationships, the manner of dressing, cutting one's hair, and mirroring, are the typical signals revealing whether a person belongs to a specific group or not. The emblems on the cloths, how elaborate is a hair dress or a crown, and the spatial arrangement of the members of the group are the typical signals revealing the status and the rank (i.e., power relations) of different members of the group.

26.3 Machine Analysis of Social Signals

Non-verbal behaviours like social signals cannot be read like words in a book [69, 103]; they are not always unequivocally associated to a specific meaning (although in general they are; [99]) and their appearance can depend on factors that have nothing to do with social behaviour. For example, some postures correspond to certain social attitudes, but sometimes they are simply comfortable [108]. Similarly, physical distances typically account for social distances, but sometimes they are simply the effect of physical constraints [53]. Moreover, as mentioned above, the same signal can correspond to different social behaviour interpretations depending on context and culture [118], although many advocate that social signals are natural rather than cultural [113]. In other words, social signals are intrinsically ambiguous, high-level semantic events, which typically include interactions with the environment and causal relationships.

Fig. 26.3 Behavioural cues typical of disagreement (*clockwise from top left*): Forefinger raise, forefinger wag, hand wag, and hands scissor [12]. These cues can be recognise with state-of-the-art human–action-recognition techniques like that proposed by [85]

An important distinction between the analysis of high-level semantic events and the analysis of low-level semantic events like the occurrence of an individual behavioural cue like the blink, is the degree to which the context, different modalities, and time, must be explicitly represented and manipulated, ranging from simple spatial reasoning to context-constrained reasoning about multimodal events shown in temporal intervals. However, despite a significant progress in automatic recognition of audiovisual behavioural cues underlying the manifestation of various social signals (e.g., see Fig. 26.3), most of the present approaches to machine analysis of human behaviour are neither multimodal, nor context-sensitive, nor suitable for handling longer time scales [89, 90, 132]. In turn, most of the social signal recognition methods reported so far are single-modal, context-insensitive and unable to handle long-time recordings of the target phenomena [125, 126].

Social interactions: Social interactions have been mostly studied in the context of small group meetings. The early works on automatic analysis of meetings [79] have been mainly aimed at recognising who says what (speaker diarisation and speech recognition) or who does what and when (tracking, movement analysis and action recognition); other aspects of social interactions like interaction cohesion, conversational context, and conversational patters, have not been studied. Arguably

the best-known group doing research towards such a deep analysis of social interactions is that led by Daniel Gatica-Perez. Relevant studies include the overview of the past work on non-verbal analysis of social interactions in small groups [44], automatic recognition of conversational context [64], and interaction cohesion estimation [59].

Social emotions: Whilst the state of the art in machine analysis of basic emotions such as happiness, anger, fear and disgust, is fairly advanced, especially when it comes to analysis of acted displays recorded in constrained lab settings [132], machine analysis of social emotions such as empathy, envy, admiration, etc., is yet to be attempted. Although some of social emotions could be arguably represented in terms of affect dimensions—valence, arousal, expectation, power, and intensity—and pioneering efforts towards automatic dimensional and continuous emotion recognition have been recently proposed ([39, 50, 82] see also [49], for a survey of the past work in the field), a number of crucial issues need to be addressed first if these approaches to automatic dimensional and continuous emotion recognition are to be used with freely moving subjects in real-world scenarios like patient-doctor discussions, talk-shows, job interviews, etc. In particular, published techniques revolve around the emotional expressions of a single subject rather than around the dynamics of the emotional feedback exchange between two subjects, which is the crux in the analysis of any social emotions. Moreover, the state of the art techniques are still unable to handle natural scenarios such as incomplete information due to occlusions, large and sudden changes in head pose, and other temporal dynamics typical of natural facial expressions [132], which must be expected in human–human interaction scenarios in which social emotions occur.

Social evaluations: Only recently, efforts have been reported towards automatic prediction of social evaluations including personality and beauty estimation. Automatic attribution of personality traits, in terms of the 'Big Five' personality model, has been attempted based on non-verbal cues such as prosody [74], proxemics (Zen et al., 2010), position in social networks [86], and fidgeting [97]. Automatic facial attractiveness estimation have been attempted based on the facial shape [51, 66, 111] as well as based on facial appearance information encoded in terms of Gabor filters responses [129]. However, the research in this domain is still in its very first stage and many basic research questions remain unanswered including exactly which features (and modalities) are the most informative for the target problem.

Social attitudes: Similarly to social emotions and social evaluations, automatic assessment of social attitudes has been attempted only recently and there are just a few studies on the topic. These works include studies on automatic assessment of agreement and disagreement in political debates based on non-verbal cues like prosody, head and hand gestures [13, 14], analysis of turn-taking order in conflicts [123], and work on detection of politeness and efficiency in a cooperative social interaction [15].

Social relations: In contrast to other types of social signals, social relations roles, i.e., behavioural patterns associated to expectations of interaction participants [9] have attracted a surge of interest from signal processing research community. A number of relevant works have focused on recognition of roles in constrained settings like news and talks shows [10, 101, 107], while other works have attempted to recognise roles associated to norms expressed as beliefs and preferences like social and functional roles in meetings [98]. The social relation that has been extensively investigated is dominance. Dominance is a personality trait, often intertwined with the social role an individual plays, that makes an individual have a higher influence on the outcomes of a discussion [63]. Typically adopted approaches towards automatic recognition of social relations are based on the analysis of turn-taking structure, i.e. who talks when and how much. This is in line with the findings of Conversation Analysis, showing that regularities in turn-taking account for social phenomena [106]. As turns are organised in sequences, the most effective and most frequently applied techniques are probabilistic models like HMMs (including layered HMMs, Factorial HMMs, etc.), Hidden CRFs, DBNs and similar.

Given the current state of the art in automatic analysis of social signals, the focus of future research efforts in the field should be on addressing various basic research questions and on tackling the problem of context-constrained analysis of multimodal behavioural signals shown in temporal intervals. As suggested by [89, 90], the latter should be treated as one complex problem rather than a number of detached problems in human sensing, context sensing, and human behaviour understanding.

More specifically, there are a number of scientific and technical challenges that we consider essential for advancing the state of the art in machine analysis of human behaviour like social signals. Modalities: Which behavioural channels such as the face, the body and the tone of the voice, are minimally needed for realisation of robust and accurate human behaviour analysis? Does this hold independently of the target communicative intention (e.g., social interactions/emotions/relations) to be recognised? No comprehensive study on the topic is available yet. What we know for sure, however, is that integration of multiple modalities (at least facial and vocal) produces superior results in human behaviour analysis when compared to single-modal approaches. Numerous studies have theoretically and empirically demonstrated this (e.g., see the literature overview by [104], for such studies in psychology, and the literature overview by [132], for such studies in automatic analysis of human behaviour). It is therefore not surprising that some of the most successful works in SSP so far use features extracted from multiple modalities (for an extensive overview of the past works, see [125]). However, other issues listed above are yet to be investigated. Also, note that some studies in the field indicate that the relative contributions of different modalities and the related behavioural cues to judgement of displayed behaviour depend on the targeted behavioural category and the context in which the behaviour occurs [104].

Fusion: How to model temporal multimodal fusion which will take into account temporal correlations within and between different modalities? What is the optimal level of integrating these different streams? Does this depend upon the time scale at

which the fusion is achieved? What is the optimal function for the integration? More specifically, most of the present audiovisual and multimodal systems in the field perform decision-level data fusion (i.e., classifier fusion) in which the input coming from each modality is modelled independently and these single-modal recognition results are combined at the end. Since humans display audio and visual expressions in a complementary and redundant manner, the assumption of conditional independence between audio and visual data streams in decision-level fusion is incorrect and results in the loss of information of mutual correlation between the two modalities. To address this problem, a number of model-level fusion methods were proposed that make use of the correlation between audio and visual data streams, and relax the requirement of synchronisation of these streams [132]. However, how to model multimodal fusion on multiple time scales and how to model temporal correlations within and between different modalities is yet to be explored.

Fusion and context: Do context-dependent fusion of modalities and discordance handling, which are typical for fusion of sensory neurons in humans, pertain in machine context sensing? Note that context-dependent fusion and discordance handling were never attempted within an automated system. Also note that while W4 (where, what, when, who) is dealing only with the apparent perceptual aspect of the context in which the observed human behaviour is shown, human behaviour understanding is about W5+ (where, what, when, who, why, how), where the why and how are directly related to recognising communicative intention including social signals, affect, and cognitive states of the observed person. Hence, SSP is about W5+. However, since the problem of context-sensing is extremely difficult to solve, especially for a general case (i.e., general-purpose W4 technology does not exist yet; [88, 90]), answering the why and how questions in a W4-context-sensitive manner when analysing human behaviour is virtually unexplored area of research. Having said that, it is not surprising that context-dependent fusion is truly a blue-sky research topic.

Technical aspects: Most methods for human sensing, context sensing, and human behaviour understanding work only in (often highly) constrained environments. Noise, fast movements, changes in illumination, etc., cause them to fail. Also, many of the methods in the field do not perform fast enough to support interactivity. Researchers usually choose for more sophisticated processing rather than for real-time processing. The aim of future efforts in the field should be the realisation of more robust, real-time systems, if they are to be deployed in anticipatory interfaces and social-computing technology defused throughout smart environments of the future.

26.4 Machine Synthesis of Social Signals

Automatic synthesis of social signals targets a human observer's or listener's perception of socially relevant information. While it may be true that much of social behaviour goes unnoticed [45], it appears that social signals still have an effect in

terms of unconscious perception [61] without being able to say exactly say why, we either consider a person trustworthy, competent, polite, etc., or not. In automatic behaviour synthesis, the aim is thus to create this perception by timely generating suitable signals and behaviours in a synthetic voice, facial expressions and gestures of an Embodied Conversational Agent (ECA). For a comprehensive overview of works on social signal generation on virtual agents, see [126]. Above we defined social signals as communicative and informative signals that concern 'social facts' including social interaction, social emotions, social evaluations, social attitudes and social relations. The work on synthesis has considered each of these dimensions. We highlight some typical examples.

Social interactions: The prime appearance of virtual humans is as embodied conversational agents, most often referred to as ECAs [20]. The research regarding ECAs is concerned primarily with investigating social interaction in the form of face-to-face conversations that exhibit all the layers of interaction: natural language understanding and generation in combination with non-verbal signals [21], conversation management such as turn-taking and backchannelling [11, 56, 65, 71], and all the other social dimensions that will be mentioned next.

Social emotions: In many scenarios, the recognition and expression of emotions through a virtual humans face [83, 84] and voice [112] or any other form of non-verbal behaviour is very important. Besides the dimension of expression, the synthesis research community has devoted much energy in defining and implementing computational models of behaviours that underlie the decisions of the choice of emotional expression. For an overview see [75].

Social evaluations: The computational models of emotions, based on appraisal models typically contain variables that deal with the evaluation of the human interlocutor and the situation the agent is in. On the other hand, many studies dealing with the evaluation of virtual humans [105] consider the other side of the coin: the question of how the agent is perceived by the human. This can pertain to any of the behaviours exhibited by the agent and any dimension. For instance, [115] consider how different turn-taking strategies evoke different impressions, [29] and [26] consider the effect of wrinkles, just to give two extreme examples of behaviours and dimensions of expression that have been related to social evaluation.

Social attitudes: Several applications of virtual humans aim at changing attitudes of the user. This often takes the form of coaching applications. Bickmore's agent Laura, a fitness instructor, is a prime example [8]. Other relevant work is the treatment of politeness and related expressions (for instance by [28] and [84], Fig. 26.4).

Social relations: The Laura agent was one of the first agents that was extensively studied in a longitudinal study. One of the major research interests in developing the agent for this study was modelling the long-term relations that might develop between the agent and the user over the course of repeated interactions. This involved

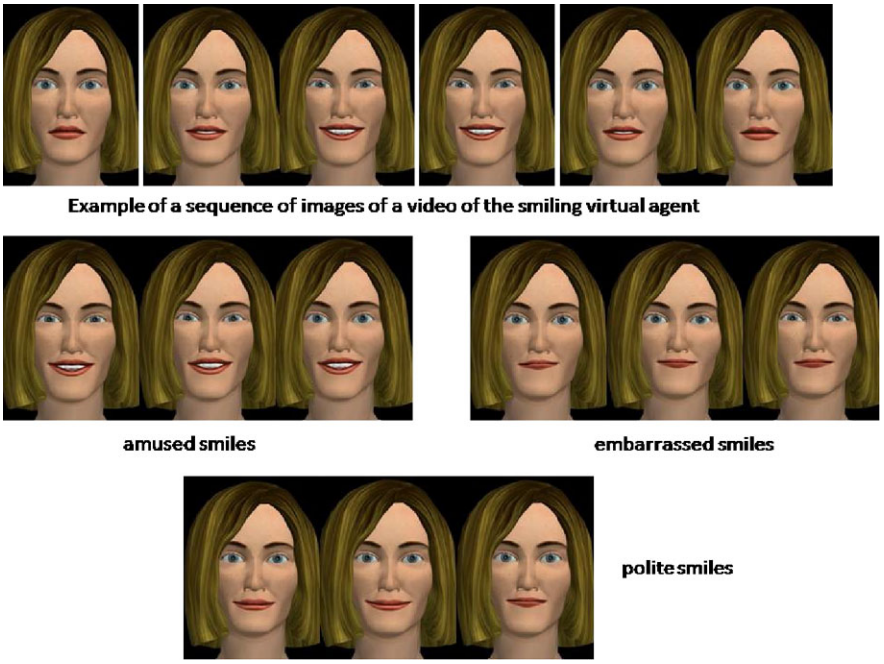

Fig. 26.4 A variety of smiles of a virtual agent (Ochs et al., 2010)

modelling many social psychological theories on relationships formation and friendship. Currently, there is a surge of work on companion agents and robots [23, 70, 72]. However, how to generate suitable behavioural signals is by no means clear, mainly due to the following two reasons. Firstly, too little is known about the types of socially relevant information conveyed in everyday human-to-human interactions, as well as about the signals and behaviours that humans naturally use to convey them. A first step in this direction would be to acknowledge the complexity of the phenomena, as has been done for emotion-related communication [30]. Then, different contexts and effects could be studied based on suitable data, and the findings could be described in terms of explicit markup language [76] or in terms of statistical, data-driven models. Secondly, it is not self-evident that synthetic agents should behave in the same way as humans do, or that they should exhibit faithful copy of human social behaviours. On the contrary, evidence from the cartoon industry [5] suggests that, in order to be believable, cartoon characters need to show strongly exaggerated behaviour. This suggests further that a trade-off between the degree of naturalness and the type of (exaggerated) gestural and vocal expression may be necessary for modelling a believable ECA's behaviour. In addition, a number of aspects of social signals are particularly relevant and challenging when it comes to synthesis of human-like behaviour.

Continuity: Unlike traditional dialogue systems, in which verbal and non-verbal behaviour is exhibited only when the system has the 'turn', socially aware systems

need to be continuous in terms of non-verbal behaviour to be exhibited. In any socially relevant situation, social signals are continuously displayed, and lack of such displays in an automatic conversational system is interpreted as social ignorance [125]. The omission of social signals, typical for today's technology, is a social signal in itself, indicating the lack of social competence. Yet, continuous synthesis of socially appropriate social signals is yet to be attempted. Complex relations between social signals' form and meaning: As explained above, relationships between social signals and their meaning are intrinsically complex. Firstly, the meaning of various signals is often not additive: when signals with meanings x and y are shown at the same time, the meaning of this complex signal may not be derivable from x and y alone. In addition, context plays a crucial role for the choice and interpretation of social signals. For example, environmental aspects such as the level of visibility and noise influence the choice of signals to be shown. On the other hand, societal aspects such as the formality of the situation and previously established roles and relations of the persons involved, and individual aspects such as the personality and affective state influence not only the choice of signals to be shown but the interpretation of the observed signals as well. Hence, context-sensitive synthesis of human behaviour is needed but it still represents an entirely blue-sky research topic. Timing: Social signals are not only characterised by the verbal and non-verbal cues by means of which they are displayed but also by their timing, that is, when they were displayed in relation to the signals displayed by other communicators involved in the interaction. Thus, social signals of an ECA need to be produced in anticipation, synchrony, or response to the actions of the human user with whom the character engages in the social interaction. This requires complex feedback loops between action and perception in real-time systems. This is another entirely unexplored, yet highly relevant, research topic.

Consistency: In general, it appears that human users are very critical when it comes to the consistency of a virtual character [62]. This relates to the challenge of multimodal synchronisation, that is, to timing between facial expression, gesture, and voice conveying a coherent and appropriate message. Research on this aspect is still ongoing there is no consensus on whether multimodal cues need to be fully synchronised, whether the redundancy of information coming from multiple cues is required, or whether it is also possible for one modality to compensate for the lack of expressiveness in other modalities (e.g., [27]). Consistency may also play a role in Mori's notion of an 'uncanny valley' [81]—a robot that looks like a human but does not behave like one is perceived as unfamiliar and 'strange'. Similarly, behaviour that may be consistent with a photo-realistic character may not be perceived as natural for a cartoon-like character, and vice versa. Technical aspects: While it will take decades to fully understand and be able to synthesise various combinations of social signals that are appropriate for different contexts and different ECAs, we expect that it will soon be possible to model some limited but relevant phenomena. One example could be a model of politeness taking into account various modalities that, for a given ECA in a given context, contribute individually and jointly to the perception of a polite or rude behaviour (e.g., see the work by [28]).

There is an obvious relevance for applications: just like their human models, service robots/ECAs should exhibit polite behaviour, whereas rescue robots should be able to insist on security-related requests. However, even when it is clear what signals and behaviours to generate, a practical challenge remains: current technology still lacks flexible models of expressivity and it usually does not operate in real-time. Expressive synthetic speech, for example, is a research topic that despite two decades of active research is still somewhat in its infancy [112]. Existing approaches are either capable of domain-specific natural-sounding vocal expressivity for a small number of possible expressions, or they achieve more flexible control over expressivity but of lower quality. Similarly, fully naturalistic movements of virtual agents can be attained when human movements recorded using motion capture technology are played back [80], but movements generated based on behaviour markup language tend to look less natural [42]. These problems are not specific to synthesis of social signals, and they do not form insurmountable obstacles to research; however, they slow down the research, by making it substantially more time-consuming to create high-quality examples of the targeted expressions. Given the above-mentioned importance of timing, the lack of real-time systems impedes the realisation of timely appropriate social behaviours. Even a slight delay in the analysis and synthesis of signals hinders dynamic adaptation and synchrony that are crucial in social interaction. Furthermore, the technological limitations pose serious difficulties for exploitation of research results in end-user applications, where fast adaptation to new domains is an important requirement. Therefore, enhancing the existing technology remains an important challenge facing the researchers in the field, independently of whether the aim is to develop socially adapt ECAs or robots with no need of social awareness.

26.5 Summary and Additional Issues

Based on the enumeration of goals and challenges facing the researchers in the SSP domain as discussed in the previous chapters, the goals of the SSP research overall can be summarised under three headings: Technological goals, human science goals, and practical impact goals.

Technological goals:

- To develop systems capable of detecting and interpreting behavioural patterns that carry information about human social activity (analysis).
- To develop systems capable of synthesising behavioural patterns that carry socially significant information to humans (synthesis).
- To develop systems capable of spotting patterns of the user's behaviour that carry socially significant information to synthesise appropriate behaviours in an interaction with the user (system responsiveness).
- To develop sophisticated tools for instrumenting human science research.

Human science goals:

- To develop theories regarding the use of social signals during human–human interactions that can inform artificial agent behaviour, and can inform human–computer interactions.
- To contribute to the human science literature by modifying current theories and proposing new theories informed by the computational research in SSP.
- To create databases suitable for the analysis of human–human interactions, and suitable for training synthesis systems.
- To develop representational systems that describe human social behaviour and cognition in ways that are appropriate to technological tasks (such as labelling databases).
- To develop methods of measuring & evaluating social interactions (human/human and human/machine).

Practical impact goals: Application of the research on SSP is not restricted to a narrowly predefined set of issues like the ones listed above. It aims to address practical problems in a range of areas. Natural application areas include artificial agents and companions, human–computer interfaces, ambient intelligence, assisted living, entertainment, education, social skills training, and multimedia indexing. Applications have the important advantage of linking the effectiveness of detection/synthesis of social signals to the reality. For example, one of the earliest applications was the prediction of the outcome in transactions recorded at a call centre, and the results show that the number of successful calls can be increased by around 20% by stopping early the calls that are not promising [17]. Defining a set of promising real-world applications could not only have a positive impact on the eventual deployment of the technology, but could also provide benchmarking procedures for the SSP research, one of the best means to improve the overall quality of a research domain as extensively shown in fields where international evaluations take place every year (e.g., video analysis in TrecVid; Smeaton et al., 2006).

The key challenge: Based on the discussion so far, it should be clear that SSP research meets a specific challenge arising from the nature of the research it requires a strong collaboration between human sciences and technology research. This challenge should not only be achievable, but should be considered paramount to the success of SSP research.

Besides the challenges discussed in the previous sections, there are a number of issues with a significant bearing on the character of the field that are still a matter of debate. Although they have not been decisively resolved, the profile of technological activities in the field implies that it tilts towards a particular kind of balance. Key examples are the following.

- *Should linguistic information be included*? From a human science standpoint, language is the social signal par excellence, and should obviously be included. Technologically, there is an obvious motive to avoid it. To wit, findings in basic research like those reported by [43] and [3] indicate that linguistic messages are rather unreliable means to analyse human behaviour, and it is very difficult to

anticipate a person's word choice and the associated intent in affective and socially situated expressions. In addition, the association between linguistic content and behaviour (e.g. emotion) is language-dependent and generalising from one language to another is very difficult to achieve.

- *Naturalness vs. artificiality*: Research in some related areas (e.g., affective computing) has relied heavily on data from actors or laboratory tasks, because naturalistic data and the related ground truth is too difficult to acquire. In return, some critics imply that only research on totally natural data is of any value. The balance implicit in the SSP research is that naturalness is a matter of degree, especially when it comes to learning the behaviour-synthesis models. Simulation is acceptable and, probably, in some cases practically necessary, so long as the signs in question are actually being used in an appropriate kind of interaction. Although such acted data can be used to learn how to synthesise certain behaviours, deliberately displayed data should be avoided when it comes to training machine learning methods for automatic analysis of social signals. Increasing evidence suggests that deliberate or posed behaviour differs in appearance and timing from that which occurs in daily life [25, 88, 120, 121]. Approaches that have been trained on deliberate and often exaggerated behaviours may fail to generalise up to the complexity of expressive behaviour found in real-world settings.
- *What are the appropriate validity criteria*? Research in computer science, especially in computer vision and pattern recognition, insists that data should be associated with a clear ground truth. In SSP that leads to very difficult demands asking, for instance, what a person really felt or intended in a particular situation. A common alternative is to require high inter-rater agreement. That, too, is problematic, because it is a feature of some social signals that different people 'read' them in different ways. The balance implicit in SSP is that the appropriate test depends on the actual application.

An additional challenging issue that has not been discussed so far relates to the fact that SSP deals with issues that are ethically sensitive. As a result, SSP has a range of ethical obligations. Many are standard, but some are not. Obligations that are shared with many other fields include: avoiding distress, deception and other undesirable effects on participants in studies, maintaining the confidentiality and where appropriate anonymity of participants involved in the research, avoiding the development of systems that could reasonably be regarded as intrusive, and limiting opportunities for abuse of the systems that they develop (e.g., through licensing arrangements). Particular obligations arise from the combination of complexity and sensitivity that is associated with social signals. The general requirement is sensitivity to the ways that social communication can affect people. Applying that to specific cases depends on intellectual awareness of individual issues (personality, age, etc.), of cultural issues (norms, specific signs, etc.), and of general expectations (what is disturbing, humiliating, etc.). Communicating about the area to non-experts raises particular issues. People are prone to systematic misunderstanding of SSP-type systems, so that they rely on them when they ought not to, fear them when they have no need to, and so on. Obligations relevant to offsetting are honesty (i.e., ensuring that what is said about a system is true), modesty (i.e., taking pains to ensure

that its limitations as well as its achievements are understood), and public education (i.e., trying to equip people with the background knowledge to grasp what a particular system might or might not be able to do).

26.6 Conclusion

Social Signal Processing (SSP) [95, 96, 124, 125] is the new research and technological domain that aims at providing computers with the ability to sense and understand human social signals. SSP is in its initial phase and the first step is to define the field and discuss issues facing the researchers in the field, which we attempted to achieve in this article.

Despite being in its initial phase, SSP has already attracted the attention of the technological community: the MIT Technology Review magazine identifies reality mining (one of the main applications of SSP so far), as one of the ten technologies likely to change the world (Greene, 2008), while management experts expect SSP to change organisation studies like the microscope has changed medicine a few centuries ago [17]. What is more important is that the first results in the field attest that social interactions and behaviours, although complex and rooted in the deepest aspects of human psychology, can be analysed and synthesised automatically with the help of computers [125, 126]. However, although fundamental, these are only the first steps, and the journey towards artificial social intelligence and socially aware computing is still long.

Acknowledgements This work has been funded in part by the European Community's 7th Framework Programme [FP7/20072013] under grant agreement no. 231287 (SSPNet).

References

1. Albrecht, K.: Social Intelligence: The New Science of Success. Wiley, New York (2005) [512]
2. Allwood, J.: Cooperation and flexibility in multimodal communication. In: Bunt, H., Beun, R. (eds.) Cooperative Multimodal Communication. Lecture Notes in Computer Science, vol. 2155, pp. 113–124. Springer, Berlin (2001) [517]
3. Ambady, N., Rosenthal, R.: Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychol. Bull. **111**(2), 256–274 (1992) [512,531]
4. Bänziger, T., Scherer, K.: Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus. In: Paiva, A., Prada, R., Picard, R. (eds.) Affective Computing and Intelligent Interaction. Lecture Notes in Computer Science, vol. 4738, pp. 476–487. Springer, Berlin (2007) [517]
5. Bates, J.: The role of emotion in believable agents. Commun. ACM **37**(7), 122–125 (1994) [528]
6. Beer, C.G.: What is a display? Am. Zool. **17**(1), 155–165 (1977) [516]
7. Berscheid, E., Reis, H.T.: Attraction and close relationships. In: Lindzey, G., Gilbert, D.T., Fiske, S.T. (eds.), The Handbook of Social Psychology, pp. 193–281. McGraw-Hill, New York (1997) [522]
8. Bickmore, T.W., Picard, R.W.: Establishing and maintaining long-term human–computer relationships. ACM Trans. Comput.–Hum. Interact. **12**(2), 293–327 (2005) [527]
9. Biddle, B.J.: Recent developments in role theory. Annu. Rev. Sociol. **12**, 67–92 (1986) [525]
10. Bigot, B., Ferrane, I., Pinquier, J., Andre-Obrecht, R.: Detecting individual role using features extracted from speaker diarization results. Multimedia Tools Appl. 1–23 (2011) [525]

11. Bonaiuto, J., Thórisson, K.R.: Towards a neurocognitive model of realtime turntaking in face-to-face dialogue. In: Knoblich, G., Wachsmuth, I., Lenzen, M. (eds.), Embodied Communication in Humans and Machines. Oxford University Press, London (2008) [527]
12. Bousmalis, K., Mehu, M., Pantic, M.: Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. In: Proceedings of the International Conference on Affective Computing and Intelligent Interfaces Workshops, vol. 2 (2009) [520,523]
13. Bousmalis, K., Mehu, M., Pantic, M.: Agreement and disagreement: A survey of nonverbal audiovisual cues and tools. Image Vis. Comput. J. (2012) [520,524]
14. Bousmalis, K., Morency, L., Pantic, M.: Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In: IEEE International Conference on Automatic Face and Gesture Recognition (2011) [524]
15. Brunet, P.M., Charfuelan, M., Cowie, R., Schroeder, M., Donnan, H., Douglas-Cowie, E.: Detecting politeness and efficiency in a cooperative social interaction. In: International Conference on Spoken Language Processing (Interspeech), pp. 2542–2545 (2010) [524]
16. Brunswik, E.: Perception and the Representative Design of Psychological Experiments. University of California Press, Berkeley (1956) [517]
17. Buchanan, M.: The science of subtle signals. Strateg. Bus. **48**, 68–77 (2007) [512,531,533]
18. Burgoon, J.K., Le Poire, B.A.: Nonverbal cues and interpersonal judgments: Participant and observer perceptions of intimacy, dominance, composure, and formality. Commun. Monogr. **66**(2), 105–124 (1999) [515]
19. Byrne, D.: The Attraction Paradigm. Academic Press, New York (1971) [521]
20. Cassell, J., Sullivan, J., Prevost, S., Churchill, E.: Embodied Conversational Agents. MIT Press, Cambridge (2000) [527]
21. Cassell, J., Vilhjálmsson, H.H., Bickmore, T.W.: BEAT: The behavior expression animation toolkit. In: ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'01), pp. 477–486 (2001) [527]
22. Castelfranchi, C.: Social power: A missed point in DAI, MA and HCI. In: Demazeau, Y., Mueller, J.P. (eds.) Decentralized AI, pp. 49–62. North-Holland, Elsevier (1990) [522]
23. Cavazza, M., de la Camara, R.S., Turunen, M.: How was your day?: A companion ECA. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1 – Volume 1. AAMAS '10, pp. 1629–1630. International Foundation for Autonomous Agents and Multiagent Systems, Richland (2010) [528]
24. Cohen, P., Levesque, H.: Performatives in a rationally based speech act theory. In: Annual Meeting of the Association of Computational Linguistics, Pittsburgh, pp. 79–88 (1990) [519]
25. Cohn, J., Schmidt, K.: The timing of facial motion in posed and spontaneous smiles. Int. J. Wavelets Multiresolut. Inf. Process. **2**(2), 121–132 (2004) [532]
26. Courgeon, M., Buisine, S., Martin, J.-C.: Impact of expressive wrinkles on perception of a virtual character's facial expressions of emotions. In: Proceedings of the 9th International Conference on Intelligent Virtual Agents. IVA '09, pp. 201–214. Springer, Berlin (2009) [527]
27. de Gelder, B., Vroomen, J.: The perception of emotions by ear and by eye. Cogn. Emot. **14**(3), 289–311 (2000) [529]
28. de Jong, M., Theune, M., Hofs, D.H.W.: Politeness and alignment in dialogues with a virtual guide. In: International Conference on Autonomous Agents and Multiagent Systems, pp. 207–214 (2008) [527,529]
29. de Melo, C., Gratch, J.: Expression of emotions using wrinkles, blushing, sweating and tears. In: International Conference on Intelligent Virtual Agents (2009) [527]
30. Douglas-Cowie, E., Devillers, L., Martin, J.C., Cowie, R., Savvidou, S., Abrilian, S., Cox, C.: Multimodal databases of everyday emotion: Facing up to complexity. In: International Conference on Spoken Language Processing (Interspeech), pp. 813–816 (2005) [528]
31. Duncan, S.: Some signals and rules for taking speaking turns in conversations. J. Pers. Soc. Psychol. **23**(2), 283–292 (1972) [516,518]
32. Eagle, N., Pentland, A.: Reality mining: sensing complex social signals. J. Pers. Ubiquitous Comput. **10**(4), 255–268 (2006) [514]

33. Efron, D.: Gesture and Environment. King's Crown Press, New York (1941) [517]
34. Eibl-Eibesfeldt, I.: Human Ethology. Aldine De Gruyter, New York (1989) [516]
35. Ekman, P.: Are there basic emotions? Psychol. Rev. **99**(3), 550–553 (1992) [517]
36. Ekman, P.: Should we call it expression or communication? Innov. Soc. Sci. Res. **10**(4), 333–344 (1997) [517]
37. Ekman, P., Friesen, W.: The repertoire of nonverbal behavior: Categories, origins, usage and coding. Semiotica **1**(1), 49–98 (1969) [516-518]
38. Enquist, M.: Communication during aggressive interactions with particular reference to variation in choice of behaviour. Anim. Behav. **33**(4), 1152–1161 (1985) [516]
39. Eyben, F., Wollmer, M., Valstar, M.F., Gunes, H., Schuller, B., Pantic, M.: String-based audiovisual fusion of behavioural events for the assessment of dimensional affect. In: IEEE International Conference on Automatic Face and Gesture Recognition (FG'11) (2011) [524]
40. Festinger, L., Schachter, S., Back, K.: Social Pressures in Informal Groups: A Study of Human Factors in Housing. Stanford University Press, Palo Alto (1950) [521]
41. Fishbein, M., Ajzen, I.: Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research. Addison-Wesley, Reading (1975) [520]
42. Foster, M.E.: Comparing rule-based and data-driven selection of facial displays. In: Proceedings of the Workshop on Embodied Language Processing, pp. 1–8 (2007) [530]
43. Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T.: The vocabulary problem in human-system communication. Commun. ACM **30**(11), 964–971 (1987) [531]
44. Gatica-Perez, D.: Automatic nonverbal analysis of social interaction in small groups: a review. Image Vis. Comput. **27**(12), 1775–1787 (2009) [524]
45. Gladwell, M.: Blink: The Power of Thinking Without Thinking. Little, Brown and Co., New York (2005) [514,526]
46. Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R.: Creating rapport with virtual agents. In: International Conference on Intelligent Virtual Agents, pp. 125–138 (2007) [514]
47. Grice, H.P.: Meaning. Philosoph. Rev. **66**, 377–388 (1957) [519]
48. Guilford, T., Dawkins, M.S.: What are conventional signals? Anim. Behav. **49**, 1689–1695 (1995) [516]
49. Gunes, H., Pantic, M.: Automatic, dimensional and continuous emotion recognition. Int. J. Synthet. Emot. **1**(1), 68–99 (2010) [524]
50. Gunes, H., Pantic, M.: Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In: International Conference on Intelligent Virtual Agents (2010) [524]
51. Gunes, H., Piccardi, M.: Assessing facial beauty through proportion analysis by image processing and supervised learning. Int. J. Human–Comput. Stud. **64**, 1184–1199 (2006) [524]
52. Hadar, U., Steiner, T., Rose, F.C.: Head movement during listening turns in conversation. J. Nonverbal Behav. **9**(4), 214–228 (1985) [516]
53. Hall, E.T.: The Silent Language. Doubleday, New York (1959) [522]
54. Hasson, O.: Cheating signals. J. Theor. Biol. **167**, 223–238 (1994) [516]
55. Heylen, D.: Challenges ahead: Head movements and other social acts in conversations. In: International Conference on Intelligent Virtual Agents (2005) [520]
56. Heylen, D., Bevacqua, E., Pelachaud, C., Poggi, I., Gratch, J.: Generating Listener Behaviour. Springer, Berlin (2011) [527]
57. Hinde, R.: The concept of function. In: Baerends, G., Manning, A. (eds.), Function and Evolution in Behaviour, pp. 3–15. Clarendon Press, Oxford (1975) [516,522]
58. Homans, G.C.: Social Behavior: Its Elementary Forms. Harcourt Brace, Orlando (1961) [521]
59. Hung, H., Gatica-Perez, D.: Estimating cohesion in small groups using audio-visual nonverbal behavior. IEEE Trans. Multimedia, Special Issue on Multimodal Affective Interaction **12**(6), 563–575 (2010) [524]
60. Hung, H., Jayagopi, D., Yeo, C., Friedland, G., Ba, S., Odobez, J.M., Ramchandran, K., Mirghafori, N., Gatica-Perez, D.: Using audio and video features to classify the most dominant person in a group meeting. In: International Conference Multimedia (2007) [514]

61. Hyman, S.E.: A new image for fear and emotion. Nature **393**, 417–418 (1998) [527]
62. Isbister, K., Nass, C.: Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. Int. J. Human–Comput. Stud. **53**, 251–267 (2000) [529]
63. Jayagopi, D., Hung, H., Yeo, C., Gatica-Perez, D.: Modeling dominance in group conversations from non-verbal activity cues. IEEE Trans. Audio, Speech Language Process. **17**(3), 501–513 (2009) [525]
64. Jayagopi, D., Kim, T., Pentland, A., Gatica-Perez, D.: Recognizing conversational context in group interaction using privacy-sensitive mobile sensors. In: ACM International Conference on Mobile and Ubiquitous Multimedia (2010) [524]
65. Jonsdottir, G.R., Thorisson, K.R., Nivel, E.: Learning smooth, human-like turntaking in realtime dialogue. In: Proceedings of the 8th international conference on Intelligent Virtual Agents, pp. 162–175. Springer, Berlin (2008) [527]
66. Kagian, A., Dror, G., Leyvand, T., Meilijson, I., Cohen-Or, D., Ruppin, E.: A machine learning predictor of facial attractiveness revealing human-like psychophysical biases. Vis. Res. **48**, 235–243 (2008) [524]
67. Kelley, H.H., Thibaut, J.: Interpersonal Relations: A Theory of Interdependence. Wiley, New York (1978) [521]
68. Keltner, D.: Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement and shame. J. Pers. Soc. Psychol. **68**(3), 441–454 (1995) [515]
69. Knapp, M.L., Hall, J.A.: Nonverbal Communication in Human Interaction. Harcourt Brace, New York (1972) [522]
70. Koay, K.L., Syrdal, D.S., Walters, M.L., Dautenhahn, K.: Five weeks in the robot house. In: International Conference on Advances in Computer–Human Interactions (2009) [528]
71. Kopp, S., Stocksmeier, T., Gibbon, D.: Incremental multimodal feedback for conversational agents. In: International Conference on Intelligent Virtual Agents (2007) [527]
72. Leite, I., Mascarenhas, S., Pereira, A., Martinho, C., Prada, R., Paiva, A.: Why can't we be friends? – an empathic game companion for long-term interaction. In: International Conference on Intelligent Virtual Agents (2010) [528]
73. Lewis, R.L.: Beyond dominance: the importance of leverage. Q. Rev. Biol. **77**(2), 149–164 (2002) [522]
74. Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using linguistic cues for the automatic recognition of personality in conversation and text. J. Artif. Intell. Res. **30**, 457–500 (2007) [524]
75. Marsella, S., Gratch, J., Petta, P.: Computational Models of Emotions. Oxford University Press, Oxford (2010) [527]
76. Martin, J., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M., Pelachaud, C.: Levels of representation in the annotation of emotion for the specification of expressivity in ECAs. In: International Conference on Intelligent Virtual Agents (2005) [528]
77. Maynard-Smith, J., Harper, D.G.: Animal signals: Models and terminology. J. Theor. Biol. **177**, 305–311 (1995) [516]
78. Maynard-Smith, J., Harper, D.G.: Animal Signals. Oxford University Press, Oxford (2003) [515,516]
79. McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., Zhang, D.: Automatic analysis of multimodal group actions in meetings. IEEE Trans. Pattern Anal. Mach. Intell. **27**(3), 305–317 (2005) [523]
80. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. **104**, 90–126 (2006) [530]
81. Mori, M.: The uncanny valley. Energy **7**, 33–35 (1970) [529]
82. Nicolaou, M., Gunes, H., Pantic, M.: Output-associative RVM regression for dimensional and continuous emotion prediction. In: IEEE International Conference on Automatic Face and Gesture Recognition (2011) [524]
83. Och, M., Niewiadomski, R., Pelachaud, C.: Expressions of empathy in ECAs. In: International Conference on Intelligent Virtual Agents (2008) [513,514,527]

84. Ochs, M., Niewiadomski, R., Pelachaud, C.: How a virtual agent should smile? morphological and dynamic characteristics of virtual agent's smiles. In: International Conference on Intelligent Virtual Agents (IVA'10) (2010) [527]
85. Oikonomopoulos, A., Patras, I., Pantic, M.: Discriminative space–time voting for joint recognition and localization of actions. In: International ACM Conference on Multimedia, Workshops (ACM-MM-W'10) (2010) [523]
86. Olguin, D., Gloor, P., Pentland, A.: Capturing individual and group behavior with wearable sensor. In: AAAI Spring Symposium (2009) [524]
87. Owren, M.J., Bachorowski, J.A.: Reconsidering the evolution of nonlinguistic communication: The case of laughter. J. Nonverbal Behav. **27**(3), 183–200 (2003) [515,516]
88. Pantic, M.: Machine analysis of facial behaviour: Naturalistic and dynamic behaviour. Philos. Trans. R. Soc. Lond. B, Biol. Sci. **364**, 3505–3513 (2009) [526,532]
89. Pantic, M., Pentland, A., Nijholt, A., Huang, T.: Human computing and machine understanding of human behavior: A survey. LNAI **4451**, 47–71 (2007) [513,523,525]
90. Pantic, M., Pentland, A., Nijholt, A., Huang, T.: Human-centred intelligent human–computer interaction (HCI2): How far are we from attaining it? Int. J. Auton. Adapt. Commun. Syst. (IJAACS) **1**(2), 168–187 (2008) [513,523,525,526]
91. Partan, S.R., Marter, P.: Communication goes multimodal. Science **283**(5406), 1272–1273 (1999) [517]
92. Peirce, C.C.: Collected Chapters. Cambridge University Press, Cambridge (1931–1935) [519]
93. Pelachaud, C., Carofiglio, V., Carolis, B.D., de Rosis, F., Poggi, I.: Embodied contextual agent in information delivering application. In: International Conference on Autonomous Agents and Multiagent Systems, pp. 758–765 (2002) [513]
94. Pentland, A.: Social dynamics: Signals and behavior. In: International Conference Developmental Learning (2004) [513,514]
95. Pentland, A.: Socially aware computation and communication. IEEE Comput. **38**(3), 33–40 (2005) [512-514,533]
96. Pentland, A.: Social signal processing. IEEE Signal Process. Mag. **24**(4), 108–111 (2007) [514,533]
97. Pianesi, F., Mana, N., Cappelletti, A.: Multimodal recognition of personality traits in social interactions. In: International Conference on Multimodal Interfaces, pp. 53–60 (2008) [524]
98. Pianesi, F., Zancanaro, M., Not, E., Leonardi, C., Falcon, V., Lepri, B.: Multimodal support to group dynamics. Pers. Ubiquitous Comput. **12**(3), 181–195 (2008) [525]
99. Poggi, I.: Mind, Hands, Face and Body: Goal and Belief View of Multimodal Communication. Weidler, Berlin (2007) [519,522]
100. Poggi, I., D'Errico, F.: Cognitive modelling of human social signals. In: Social Signal Processing Workshop, in Conjunction with International Conference on Multimedia (2010) [515,517]
101. Raducanu, B., Gatica-Perez, D.: Inferring competitive role patterns in reality TV show through nonverbal analysis. Multimedia Tools Appl. (2010) [525]
102. Rendall, D., Owren, M.J., Ryan, M.J.: What do animal signals mean? Anim. Behav. **78**(2), 233–240 (2009) [515]
103. Richmond, V.P., McCroskey, J.C.: Nonverbal Behaviors in Interpersonal Relations. Allyn & Bacon, Needham Heights (1995) [522]
104. Russell, J.A., Bachorowski, J.A., Fernandez-Dols, J.M.: Facial and vocal expressions of emotion. Annu. Rev. Psychol. **54**(1), 329–349 (2003) [525]
105. Ruttkay, Z., Pelachaud, C.: From Brows to Trust: Evaluating Embodied Conversational Agents. Kluwer Academic, Norwell (2004) [514,527]
106. Sacks, H., Schegloff, E.A., Jefferson, G.: A simplest systematics for the organization of turn taking for conversation. Language **50**(4), 696–735 (1974) [518,525]
107. Salamin, H., Favre, S., Vinciarelli, A.: Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction. IEEE Trans. Multimedia **11**(7), 1373–1380 (2009) [525]

108. Scheflen, A.E.: The significance of posture in communication systems. Psychiatry **27**, 316–331 (1964) [522]
109. Scherer, K.R.: Personality inference from voice quality: The loud voice of extroversion. Eur. J. Soc. Psychol. **8**(4), 467–487 (1978) [517]
110. Scherer, K.R.: What does facial expression express? In: Strongman, K.T. (ed.) International Review of Studies of Emotion, vol. 2, pp. 139–165. Wiley, New York (1992) [517]
111. Schmid, K., Marx, D., Samal, A.: Computation of face attractiveness index based on neo-classic canons, symmetry and golden ratio. Pattern Recogn. **41**, 2710–2717 (2008) [524]
112. Schröder, M.: Expressive Speech Synthesis: Past, Present, and Possible Futures. In: Tao, J., Tan, T. (eds.) Affective Information Processing, pp. 111–126. Springer, Berlin? (2009) [513, 527,530]
113. Segerstrale, U., Molnar, P.: Nonverbal Communication: Where Nature Meets Culture. Lawrence Erlbaum Associates, Lawrence (1997) [522]
114. Shannon, C.E., Weaver, W.: The Mathematical Theory of Information. University of Illinois Press, Champaign (1949) [518]
115. ter Maat, M., Heylen, D.: Turn management or impressions management? In: International Conference on Intelligent Virtual Agents, pp. 467–473 (2009) [527]
116. Thorndike, E.L.: Intelligence and its use. Harper's Mag. **140**, 227–235 (1920) [512]
117. Tomkins, S.S.: Consiousness, Imagery and Affect vol. 1. Springer, Berlin (1962) [517]
118. Triandis, H.C.: Culture and Social Behavior. McGraw-Hill, New York (1994) [522]
119. Trouvain, J., Schröder, M.: How (not) to add laughter to synthetic speech. Lect. Notes Comput. Sci. **3068**, 229–232 (2004) [514]
120. Valstar, M.F., Gunes, H., Pantic, M.: How to distinguish posed from spontaneous smiles using geometric features. In: International Conference Multimodal Interfaces, pp. 38–45 (2007) [532]
121. Valstar, M.F., Pantic, M., Ambadar, Z., Cohn, J.F.: Spontaneous vs. posed facial behaviour: Automatic analysis of brow actions. In: International Conference Multimodal Interfaces, pp. 162–170 (2006) [532]
122. Verhencamp, S.L.: Handicap, Index, and Conventional Signal Elements of Bird Song. In: Edpmark, Y., Amundsen, T., Rosenqvist, G. (eds.) Animal Signals: Signalling and Signal Design in Animal Communication, pp. 277–300. Tapir Academic Press, Trondheim (2000) [516]
123. Vinciarelli, A.: Capturing order in social interactions. IEEE Signal Process. Mag. **26**(5), 133–137 (2009) [524]
124. Vinciarelli, A., Pantic, M., Bourlard, H., Pentland, A.: Social signal processing: State-of-the-art and future perspectives of an emerging domain. In: International Conference Multimedia, pp. 1061–1070 (2008) [514,533]
125. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: Survey of an emerging domain. Image Vis. Comput. **27**(12), 1743–1759 (2009) [514,523,525,529,533]
126. Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., Schröder, M.: Bridging the gap between social animal and unsocial machine: A survey of social signal processing. IEEE Trans. Affect. Comput. (2012, in press) [523,527]
127. Wang, N., Johnson, W.L., Rizzo, P., Shaw, E., Mayer, R.E.: Experimental evaluation of polite interaction tactics for pedagogical agents. In: International Conference Intelligent User Interfaces, pp. 12–19 (2005) [514]
128. Weiser, M.: The computer for the 21st century. Sci. Am. Special Issue on Communications, Computers, and Networks **265**(3), 95–104 (1991) [512]
129. Whitehill, J., Movellan, J.: Personalized facial attractiveness prediction. In: IEEE International Conference on Automatic Face and Gesture Recognition (2008) [524]
130. Woodworth, R.S.: Dynamics of Behavior. Holt, New York (1961) [517]
131. Zahavi, A.: Mate selection: selection for a handicap. J. Theor. Biol. **53**, 205–214 (1975) [516]
132. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.H.: A survey of affect recognition methods: Audio, visual and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2009) [513,523-526]