Social Signal Processing: Survey of an Emerging Domain

Alessandro Vinciarelli [a,b] Maja Pantic [c,d] Hervé Bourlard [a,b]

[a] *IDIAP Research Institute - CP592 - 1920 Martigny (Switzerland)*

[b] *Ecole Polytechnique Fédérale de Lausanne (EPFL) - 1015 Lausanne (CH)*

[c] *Imperial College - 180 Queens Gate - London SW7 2AZ (UK)*

[d] *University of Twente - Drienerlolaan 5 -7522 NB Enschede (The Netherlands)*

Abstract

The ability to understand and manage social signals of a person we are communicating with is the core of social intelligence. Social intelligence is a facet of human intelligence that has been argued to be indispensable and perhaps the most important for success in life. This paper argues that next-generation computing needs to include the essence of social intelligence – the ability to recognize human social signals and social behaviours like turn taking, politeness, and disagreement – in order to become more effective and more efficient. Although each one of us understands the importance of social signals in everyday life situations, and in spite of recent advances in machine analysis of relevant behavioural cues like blinks, smiles, crossed arms, laughter, and similar, design and development of automated systems for Social Signal Processing (SSP) are rather difficult. This paper surveys the past efforts in solving these problems by a computer, it summarizes the relevant findings in social psychology, and it proposes a set of recommendations for enabling the development of the next generation of socially-aware computing.

Key words: Social signals, computer vision, speech processing, human behaviour analysis, social interactions.

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. Perceiv-

Email addresses: vincia@idiap.ch (Alessandro Vinciarelli), m.pantic@imperial.ac.uk (Maja Pantic), bourlard@idiap.ch (Hervé Bourlard).

ing, 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* [6][8][19][182] and includes the ability to express and recognise social signals and social behaviours like turn taking, agreement, politeness, and empathy, coupled with the ability to manage them in order to get along well with others while winning their cooperation. Social signals and social behaviours are the expression of ones attitude towards social situation and interplay, and they are manifested through a multiplicity of non-verbal behavioural cues including facial expressions, body postures and gestures, and vocal outbursts like laughter (see Figure 1). Social signals typically last for a short time (milliseconds, like turn taking, to minutes, like mirroring), compared to social behaviours that last longer (seconds, like agreement, to minutes, like politeness, to hours or days, like empathy) and are expressed as temporal patterns of non-verbal behavioural cues. The skills of social intelligence have been argued to be indispensable and perhaps the most important for success in life [66].

When it comes to computers, however, they are socially ignorant [143]. 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) 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 to these social signals in a polite, unintrusive, or persuasive manner, are likely to be perceived as more natural, efficacious, and trustworthy. For example, 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 (see e.g. [141]) 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

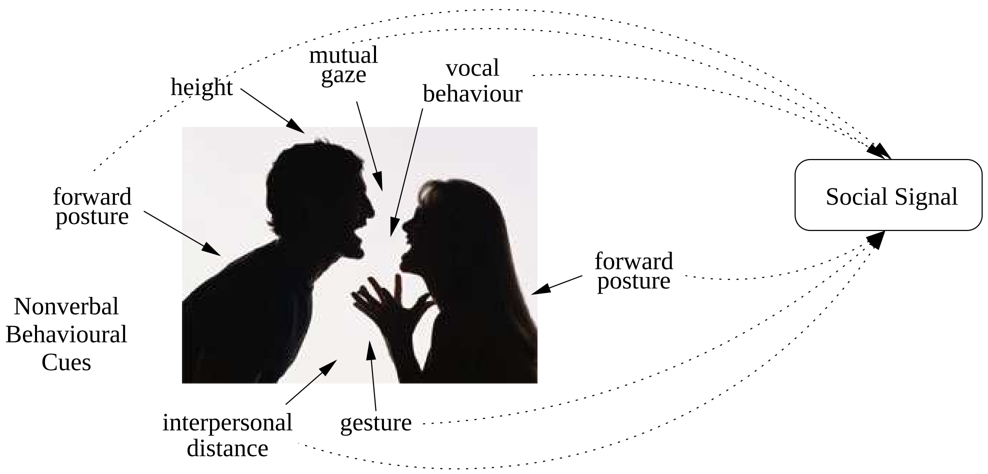


Fig. 1. Behavioural cues and social signals. Multiple behavioural cues (vocal behaviour, posture, mutual gaze, interpersonal distance, etc.) combine to produce a social signal (in this case aggressivity or disagreement) that is evident even if the picture shows only the silhouettes of the individuals involved in the interaction.

technology goes by the general name of Socially-Aware Computing as introduced by Pentland [142][143].

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, and similar [137][138], 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. The importance of studying social interactions and developing automated assessing of human social behaviour from audiovisual recordings is undisputable. 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. At the same time, and as outlined above, such tools form a large step ahead in realising naturalistic, socially-aware computing and interfaces, built for humans, based on models of human behaviour.

Social Signal Processing (SSP) [143][145][202][203] is the new research and technological domain that aims at providing computers with the ability to sense and understand human social signals. 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, see Section 4 for more details), as one of the ten technologies likely to change the world [69], while management experts expect SSP to change organization studies like the microscope has changed medicine few centuries ago [19].

To the best of our knowledge, this is the first attempt to survey the past work done on SSP. The innovative and multidisciplinary character of the research on SSP is the main reason for this state of affairs. For example, in contrast to the research on human affective behaviour analysis that witnessed tremendous progress in the past decade (for exhaustive surveys in the field see, e.g.,[76][140][221]), the research on machine analysis of human social behaviour just started to attract the interest of the research community in computer science. This and the fragmentation of the research over several scientific communities including those in psychology, computer vision, speech and signal processing, make the exercise of surveying the current efforts in machine analysis of human social behaviour difficult.

The paper begins by examining the context in which the research on SSP has arisen and by providing a taxonomy of the target problem domain (Section 2). The paper surveys then the past work done in tackling the problems of machine detection and interpretation of social signals and social behaviours in real-world scenarios (Section 3). Existing research efforts to apply social signal processing to automatic recognition of socially relevant information such as someone's role, dominance, influence, etc., are surveyed next (Section 4). Finally, the paper discusses a number of challenges facing researchers in the field (Section 5). In the authors' opinion, these need to be addressed before the research in the field can enter its next phase – deployment of research findings in real-world applications.

2 Behavioural Cues and Social Signals: A Taxonomy

There is more than words in social interactions [9], whether these take place between humans or between humans and computers [30]. This is well known to social psychologists that have studied nonverbal communication for several decades [96][158]. It is what people experience when they watch a television program in a language they do not understand and still capture a number of important social cues such as differences in status between individuals, overall atmosphere of interactions (e.g., tense vs. relaxed), rapport beteween people (mutual trust vs. mutual distrust), etc.

Nonverbal behaviour is a continuous source of signals which convey information about feelings, mental state, personality, and other traits of people [158]. During social interactions, nonverbal behaviour conveys this information not only for each of the involved individuals, but it also determines the nature and quality of the social relationships they have with others. This happens through a wide spectrum of nonverbal behavioural cues [7][8] that are perceived and displayed mostly unconsciously while producing *social awareness*, i.e. a spontaneous understanding of social situations that does not require attention or reasoning [98].

The term behavioural cue is typically used to describe a set of temporal changes in neuromuscular and physiological activity that last for short intervals of time (milliseconds to minutes) in contrast to *behaviours* (e.g. social behaviours like politeness or empathy) that last on average longer (minutes to hours). As summarised in [47] among the types of messages (communicative intentions) conveyed by behavioural cues are the following:

- *affective/attitudinal/cognitive states* (e.g. fear, joy, stress, disagreement, ambivalence and inattention),
- *emblems* (culture-specific interactive signals like wink or thumbs up),
- *manipulators* (actions used to act on objects in the environment or self-manipulative actions such as lip biting and scratching),
- *illustrators* (actions accompanying speech such as finger pointing and raised eyebrows), and
- *regulators* (conversational mediators such as the exchange of a look, palm pointing, head nods and smiles).

In most cases, behavioural cues accompany verbal communication and, even if they are invisible, i.e., they are sensed and interpreted outside conscious awareness, they have a major impact on the perception of verbal messages and social situations [96]. Early investigations of verbal and nonverbal components in interaction (in particular [113] as cited in [96]) have suggested that the verbal messages account for just 7% of the overall social perception. This conclusion has been later argued because the actual weight of the different messages (i.e. verbal vs non-verbal) depends on the context and on the specific kind of interaction [45]. However, more recent studies still confirm that the nonverbal behaviour plays a major role in shaping the perception of social situations: e.g., judges assessing the rapport between two people are more accurate when they use only the facial expressions than when they use only the verbal messages exchanged [8]. Overall, the nonverbal social signals seem to be the predominant source of information used in understanding social interactions [9].

The rest of this section provides a taxonomy of the SSP problem domain by listing and explaining the most important behavioural cues and their functions in social behaviour. Behavioural cues that we included in this list are those that the research in psychology has recognized as being the most important in human judgments of social behaviour. Table 1 provides a synopsis of those behavioural cues, the social signals they are related to, and the technologies that can be used to sense and analyse them. For more exhaustive explanations of nonverbal behaviours and the related behavioural cues, readers are referred to [7][47][96][158].

	Example Social Behaviours							**Tech.**		
Social Cues	emotion	personality	status	dominance	persuasion	regulation	rapport	speech anlysis	computer vision	biometry
Physical appearance										
height			√	√					√	√
attractiveness		√	√	√	√		√		√	√
body shape		√		√					√	√
Gesture and posture										
hand gestures	√	√			√	√	√		√	√
posture	√	√	√	√	√	√	√		√	√
walking		√	√	√					√	√
Face and eyes behaviour										
facial expressions	√	√	√	√	√	√	√		√	√
gaze behaviour	√	√	√	√	√	√	√		√	
focus of attention	√	√	√	√	√	√	√		√	
Vocal behaviour										
prosody	√	√		√	√		√	√		
turn taking	√	√	√	√		√	√	√		
vocal outbursts	√	√		√	√	√	√	√		
silence	√		√				√	√		
Space and Environment										
distance	√	√	√		√		√		√	
seating arrangement				√	√		√		√	

Table 1
The table shows the behavioural cues associated to some of the most important social behaviours as well as the technologies involved in their automatic detection.

2.1 Physical Appearance

The physical appearance includes natural characteristics such as height, body shape, physiognomy, skin and hair color, as well as artificial characteristics such as clothes, ornaments, make up, and other manufacts used to modify/

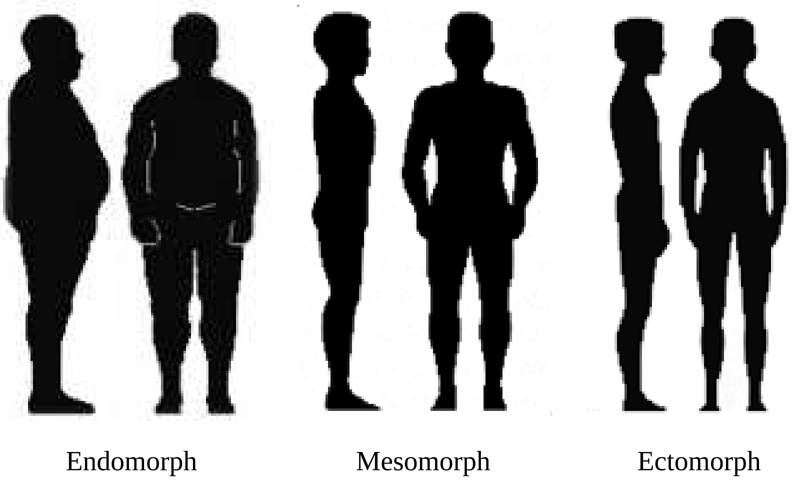


Fig. 2. Somatotypes. The figure shows the three body shapes that tend to elicit the perception of specific personality traits.

accentuate the facial/ body aspects.

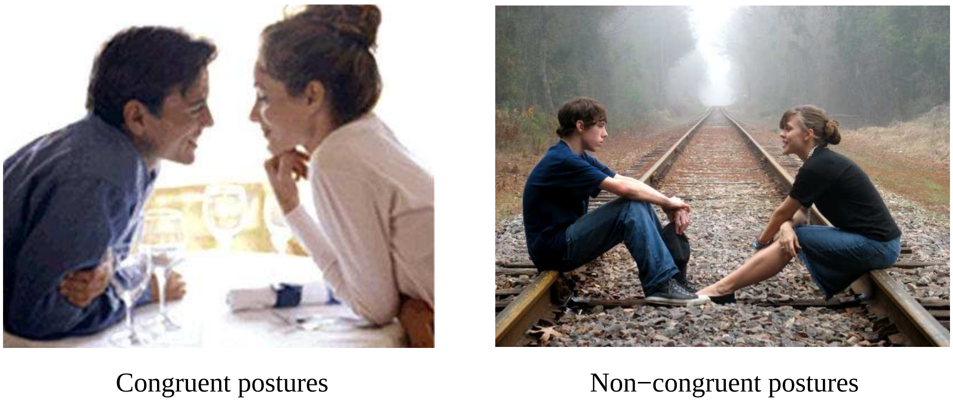
The main social signal associated to physical appearance is the *attractiveness.* Attractiveness produces a positive *halo effect* (a phenomenon also known as ”*what is beautful is good*” [41]). Attractive people are often judged as having high status and good personality even if no objective basis for such judgments exists [70][208]. Attractive people also have higher probability of starting new social relationships with people they do not know [158]. Other physical characteristics are not necessarily related to the attractiveness, but still have a major influence on social perceptions. The most important are height and *somatotype* (see below). Tall individuals tend to be attributed higher social status and, in some cases, they actually hold a higher status. For example, a survey has shown that the average height of the American CEOs of the Fortune 500 companies is around 7.5 cm higher than the average height of the American population. Moreover, 30% of the same CEOs are taller than 190 cm, while only 4% of the rest of the American population lies in the same range of height [63].

Different *somatotypes* (see Figure 2), tend to elicit the attribution of certain personality traits [25]. For example, *endomorphic* individuals (round, fat and soft) tend to be perceived as more talkative and sympathetic, but also more dependent on others. *Mesomorphic* individuals (bony, muscular and athletic) tend to be perceived as more self-reliant, more mature in behaviour and stronger, while *ectomorphic* individuals (tall, thin and fragile) tend to be perceived as more tense, more nervous, more pessimistic and inclined to be difficult. These judgments are typically influenced by stereotypes that do not necessarily correspond to the reality, but still influence significantly the social perceptions [96].

Following the work of Darwin [37], which was the first to describe body expressions associated with emotions in animals and humans, there have been a number of studies on human body postures and gestures communicating emotions. For example the works in [27][198] investigated perception and display of body postures relevant to basic emotions including happiness, sadness, surprise, fear, disgust, and anger, while the studies in [72][152] investigated bodily expressions of felt and recognized basic emotions as visible in specific changes in arm movement, gait parameters, and kinematics. Overall, these studies have shown that both posture and body/ limb motions change with emotion expressed. Basic research also provides evidence that gestures like head inclination, face touching, and shifting posture often accompany social affective states like shame and embarrassment [26][50]. However, as indicated by researchers in the field (e.g. in [112]), as much as 90% of body gestures are associated with speech, representing typical social signals such as illustrators, emblems, and regulators.

In other words, gestures are used in most cases to regulate interactions (e.g., to yield the turn in a conversation), to communicate a specific meaning (e.g., the *thumbs up* gesture to show appreciation), to punctuate a discourse (e.g., to underline an utterance by rising the index finger), to greet (e.g., by waving hands to say goodbye), etc. [123]. However, in some cases gestures are performed unconsciously and they are interesting from an SSP point of view because they account for *honest* information [146], i.e., they leak cues related to the actual attitude of a person with respect to a social context. In particular, *adaptors* express boredom, stress and negative feelings towards others. Adaptors are usually displayed unconsciously and include self-manipulations (e.g., scratching, nose and ear touching, hair twisting), manipulation of small objects (e.g., playing with pens and papers), and self-protection gestures (e.g., folding arms or rythmicly moving legs) [96].

Postures are also typically assumed unconsciously and, arguably, they are the most reliable cues about the actual attitude of people towards social situations [158]. One of the main classifications of postural behaviours proposes three main criteria to assess the social meaning of postures [166]. The first criterion distinguishes between *inclusive* and *non-inclusive* postures and accounts for how much a given posture takes into account the presence of others. For example, facing in the opposite direction with respect to others is a clear sign of non-inclusion. The second criterion is *face-to-face vs. parallel body orientation* and concerns mainly people involved in conversations. Face-to-face interactions are in general more active and engaging (the frontal position addresses the need of continuous mutual monitoring), while people sitting parallel to each other tend to be either buddies or less mutually interested. The



Congruent postures

Non−congruent postures

Fig. 3. Postural congruence. The figure on the left shows how people deeply involved in an interaction tend to assume the same posture. In the other picture, the forward inclination of the person on the right is not reciprocated by the person on the left.

third criterion is *congruence vs. incongruence*: symmetric postures tend to account for a deep psychological involvement (see left picture in Figure 3), while non-symmetric ones correspond to the opposite situation. The postural congruence is an example of a general phenomenon called *chameleon effect* or *mirroring* [22], that consists of the mutual imitation of people as a mean to display affiliation and liking. Postural behaviour includes also walking and movements that convey social information such as status, dominance and affective state [109].

2.3 Face and Eye Behaviour

The human face is involved in an impressive variety of different activities. It houses the majority of our sensory apparatus: eyes, ears, mouth and nose, allowing the bearer to see, hear, taste and smell. Apart from these biological functions, the human face provides a number of signals essential for interpersonal communication in our social life. The face houses the speech production apparatus and is used to identify other members of the species, to regulate the conversation by gazing or nodding, and to interpret what has been said by lip reading. It is our direct and naturally preeminent means of communicating and understanding somebody's affective state and intentions on the basis of the shown facial expression [89]. Personality, attractiveness, age and gender can be also seen from someone's face [8]. Thus the face is a multi-signal sender/receiver capable of tremendous flexibility and specificity. It is therefore not surprising that the experiments (see beginning of Section 2) about the relative weight of the different nonverbal components in shaping social perceptions always show that facial behaviour plays a major role [8][68][113].

Two major approaches to facial behaviour measurement in psychological re-

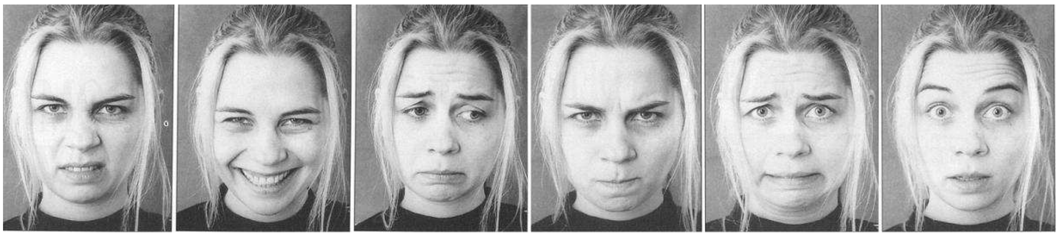


Fig. 4. Basic emotions. Prototypic facial expressions of six basic emotions (disgust, happiness, sadness, anger, fear, and surprise).

search are message and sign judgment [23]. The aim of message judgment is to infer what underlies a displayed facial expression, such as affect or personality, while the aim of sign judgment is to describe the *surface* of the shown behavior, such as facial movement or facial component shape. Thus, a brow furrow can be judged as *anger* in a message-judgment and as a facial movement that lowers and pulls the eyebrows closer together in a sign-judgment approach. While message judgment is all about interpretation, sign judgment attempts to be objective, leaving inference about the conveyed message to higher order decision making.

As indicated in [23], most commonly used facial expression descriptors in message judgment approaches are the six basic emotions (fear, sadness, happiness, anger, disgust, surprise; see Fig. 4), proposed by Ekman and discrete emotion theorists, who suggest that these emotions are universally displayed and recognized from facial expressions [89]. In sign judgment approaches [24], a widely used method for manual labeling of facial actions is the Facial Action Coding System (FACS) [48].

FACS associates facial expression changes with actions of the muscles that produce them. It defines 9 different Action Units (AUs) in the upper face, 18 in the lower face, 11 for head position, 9 for eye position, and 14 additional descriptors for miscellaneous actions. AUs are considered to be the smallest visually discernable facial movements. Using FACS, human coders can manually code nearly any anatomically possible facial expression, decomposing it into the specific AUs that produced the expression. As AUs are independent of interpretation, they can be used for any higher order decision making process including recognition of basic emotions (EMFACS; see [48]), cognitive states like interest and puzzlement [32], psychological states like suicidal depression [50] or pain [212], social behaviours like accord and rapport [8][32], personality traits like extraversion and temperament [50], and social signals like status, trustworthiness, emblems (i.e., culture-specific interactive signals like wink), regulators (i.e., conversational mediators like nod and gaze exchange), and illustrators (i.e., cues accompanying speech like raised eyebrows) [8][46][47]. FACS provides an objective and comprehensive language for describing facial expressions and relating them back to what is known about their meaning from the behavioral science literature. Because it is comprehensive, FACS also allows for the discovery of new patterns related to emotional or situational states. For example, what are the facial behaviors associated with social signals such as empathy, persuasion, and politeness? An example where subjective judgments of expression failed to find relationships which were later found with FACS is the failure of naive subjects to differentiate deception and intoxication from facial display, whereas reliable differences were shown with FACS [165]. Research based upon FACS has also shown that facial actions can show differences between those telling the truth and lying at a much higher accuracy level than naive subjects making subjective judgments of the same faces [56]. Exhaustive overview of studies on facial and gaze behaviour using FACS can be found in [50].

2.4 Vocal Behaviour

The vocal nonverbal behaviour includes all spoken cues that surroud the verbal message and influence its actual meaning. The effect of vocal nonverbal behaviour is particularly evident when the tone of a message is ironic. In this case the face value of the words is changed into its opposite by just using the appropriate vocal intonation. The vocal nonverbal behaviour includes five major components: *voice quality*, *linguistic* and *non-linguistic vocalizations*, *silences*, and *turn-taking patterns*. Each one of them relates to social signals that contribute to different aspects of the social perception of a message.

The *voice quality* corresponds to the prosodic features, i.e., pitch, tempo, and energy (see Section 3.3.4 for more details) and, in perceptual terms, accounts for *how* something is said [31]. The prosody conveys a wide spectrum of socially relevant cues: emotions like anger or fear are often accompanied by energy bursts in voice (shouts) [168], the pitch influences the perception of dominance and extroversion (in general it is a personality marker [167]), the speaking fluency (typically corresponding to high rythm and lack of hesitations) increases the perception of competence and results into higher persuasiveness [167]. The vocalizations include also effects that aim at giving particular value to certain utterances or parts of the discourse, e.g., the pitch accents (sudden increases of the pitch to underline a word) [79], or changes in rhythm and energy aiming at structuring the discourse [80].

The *linguistic vocalizations* (also known as *segregates*) include all the non-words that are typically used as if they were actual words, e.g., "*ehm*","*ah-ah*", "*uhm*", etc. Segregates have two main functions, the first is to replace words that for some reason cannot be found, e.g., when people do not know how to answer a question and simply utter a prolonged "*ehm*". They are often referred to as *disfluencies* and often account for a situation of embarassment or difficulty with respect to a social interaction [64]. The second important function is the so-called *back-channeling*, i.e., the use of segregates to accompany someone else speaking. In this sense they can express attention, agreement, wonder, as well as the attempt of grabbing the floor or contradicting [176].

The *non-linguistic vocalizations*, also known as vocal outbursts, include nonverbal sounds like laughing, sobbing, crying, whispering, groaning, and similar, that may or may not accompany words, and provide some information about the attitude towards social situations. For instance, laughter tends to reward desirable social behaviour [90] and shows affiliation efforts, while crying is often involved in *mirroring* (also known as *chameleon effect* [22]), that is in the mutual imitation of people connected by strong social bonds [91]. Also, research in psychology has shown that listeners tend to be accurate in decoding some basic emotions as well as some non-basic affective and social signals such as distress, anxiety, boredom, and sexual interest from vocal outbursts like laughs, yawns, coughs, and sighs [163].

The silence is often interpreted as simple non-speech, but actually plays a major role in the vocal behaviour [219]. There are three kinds of silence in speech: *hesitation silence*, *psycholinguistic silence*, and *interactive silence* [158]. The first takes place when a speaker has difficults in talking, e.g., because she is expressing a difficult concept or must face a hostile attitude in listeners. Sometimes, hesitation silences give rise to segregates that are used to *fill* the silence space (hence segregates are called sometimes fillers). The psycholinguistic silences take place when the speaker needs time to encode or decode the speech. This kind of silences happen often at the beginning of an intervention because the speaker needs to think about the next words. In this sense, this is often a sign of difficulty and problems in dealing with a conversation. The interactive silences aim at conveying messages about the interactions taking place: silence can be a sign of respect for people we want to listen to, a way of ignoring persons we do not want to answer to, as well as a way to attract the attention to other forms of communication like mutual gaze or facial expressions.

Another important aspect of vocal nonverbal behaviour is turn-taking [154]. This includes two main components: the regulation of the conversations, and the coordination (or the lack of it) during the speaker transitions. The regulation in conversations includes behaviours aimed at maintaining, yielding, denying, or requesting the turn. Both gaze and voice quality (e.g. coughing) are used to signal *transition relevant points* [217]. When it comes to vocal nonverbal cues as conversation regulators, specific pitch and energy patterns show the intention of yielding the turn rather than maintaining the floor. Also, linguistic vocalizations (see above) are often used as a form of back-channeling to request the turn. The second important aspect in turn-taking is the coordination at the speaker transitions [20]. Conversations where the latency times between turns are too long sound typically awkward. The reason is that in flu-

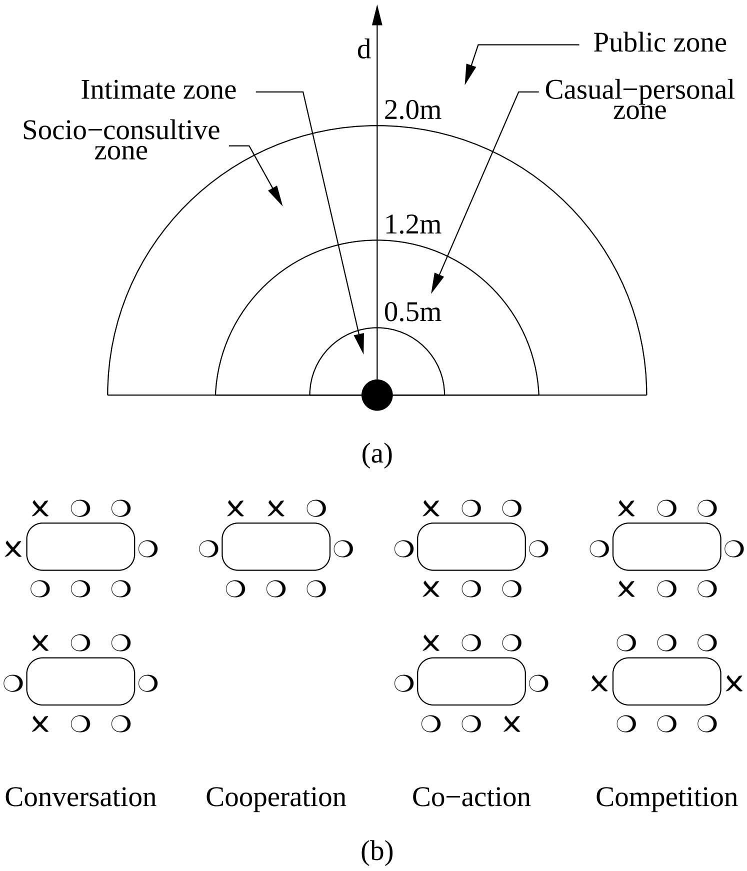


Fig. 5. Space and seating. The upper part of the figure shows the concentric zones around each individual associated to different kinds of rapport (*d* stands for distance). The lower part of the figure shows the preferred seating arrangements for different kinds of social interactions.

ent conversations, the mutual attention reduces the above phenomenon and results into synchronized speaker changes, where the interactants effectively interpret the signals aimed at maintaining or yielding their turns. Overlapping speech is another important phenomenon that accounts for disputes as well as status and dominance displays [180]. Note, however, that the amount of overlapping speech accounts for up to 10% of the total time even in normal conversations [175].

2.5 Space and Environment

The kind and quality of the relationships between individuals influences their interpersonal distance (the physical space between them). One of the most common classifications of mutual distances between individuals suggests the existence of four concentric zones around a person accounting for different kinds of relationships with the others [77]: the *intimate zone*, the *casual-*

personal zone, the *socio-consultive zone* and the *public zone* (see Figure 5a).

The *intimate zone* is the innermost region and it is open only to the closest family members and friends. Its dimension, like in the case of the other zones, depends on the culture and, in the case of western Europe and United States, the intimate zone corresponds to a distance of 0.4-0.5 meters. In some cases, e.g., crowded buses or elevators, the intimate zone must be necessarily opened to strangers. However, whenever there is enough space, people tend to avoid entering the intimate zone of others. The *casual-personal zone* ranges (at least in USA and Western Europe) between 0.5 and 1.2 meters and it typically includes people we are most familiar with (colleagues, friends, etc.). To open such an area to another person in absence of constraints is a major signal of friendship. The *socio-consultive* distance is roughly between 1 and 2 meters (again in USA and Western Europe) and it is the area of formal relationships. Not surprisingly, professionals (lawyers, doctors, etc.) typically receive their clients sitting behind desks that have a profundity of around 1 meter, so that the distance with respect to their clients is in the range corresponding to the socio-consultive zone. The *public zone* is beyond 2 meters distance and it is, in general, outside the reach of interaction potential. In fact, any exchange taking place at such a distance is typically due to the presence of some obstacle, e.g., a large meeting table that requires people to talk at distance.

Social interactions take place in environments that influence behaviours and perceptions of people with their characteristics. One of the most studied environmental variables is the seating arrangement, i.e., the way people take place around a table for different purposes [96][158]. Figure 5b shows the seating positions that people tend to use to perform different kinds of tasks (the circles are the empty seats) [164]. The seating position depends also on the personality of people: dominant and higher status individuals tend to seat at the shorter side of rectangular tables, or in the middle of the longer sides (both positions ensure high visibility and make easier the control of the conversation flow) [106]. Moreover, extrovert people tend to privilege seating arrangements that minimize interpersonal distances, while introvert ones do the opposite [164].

3 The State of the Art

The problem of machine analysis of human social signals includes four subproblem areas (see Figure 5):

(1) recording the scene,
(2) detecting people in it,
(3) extracting audio and/or visual behavioural cues displayed by people de-

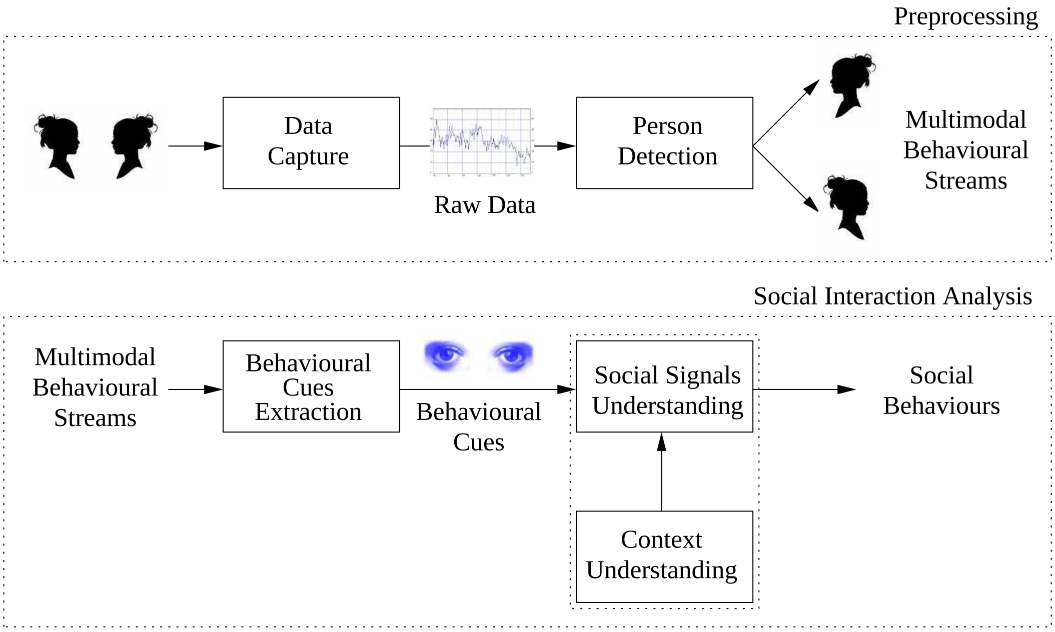

Preprocessing
Data Capture
Raw Data
Person Detection
Multimodal Behavioural Streams
Social Interaction Analysis
Multimodal Behavioural Streams
Behavioural Cues Extraction
Behavioural Cues
Social Signals Understanding
Social Behaviours
Context Understanding

Fig. 6. Machine analysis of social signals and behaviours: a general scheme. The process includes two main stages: The *preprocessing*, takes as input the recordings of social interaction and gives as output the multimodal behavioural streams associated with each person. The *social interaction analysis* maps the multimodal behavioural streams into social signals and social behaviours.

tected in the scene and interpreting this information in terms of social signals conveyed by the observed behavioural cues,

(4) sensing the context in which the scene is recorded and classifying detected social signals into the target social-behaviour-interpretative categories in a context-sensitive manner.

The survey of the past work is divided further into four parts, each of which is dedicated to the efforts in one of the above-listed subproblem areas.

3.1 Data Capture

Data capture refers to using sensors of different kinds to capture and record social interactions taking place in real world scenarios. The choice of the sensors and their arrangement in a specific recording setup determine the rest of the SSP process and limit the spectrum of behavioral cues that can be extracted. For example, no gaze behavior analysis can be performed, if appropriate detectors are not included in the capture system.

The most common sensors are microphones and cameras and they can be arranged in structures of increasing complexity: from a single camera and/

or microphone to capture simple events like oral presentations [201], to fully equipped *smart meeting rooms* where several tens of audio and video channels (including microphone arrays, fisheye cameras, lapel microphones, etc.) are setup and synchronized to capture complex interactions taking place in a group meeting [110][205]. The literature shows also examples of less common sensors such as cellular phones or smart badges equipped with proximity detectors and vocal activity measurement devices [43][144], and systems for the measurement of physiological activity indicators such as blood pressure and skin conductivity [76]. Recent efforts have tried to investigate the neurological basis of social interactions [2] through devices like *functional Magnetic Resonance Imaging* (fMRI) [119], and *Electroencephalography* (EEG) signals [193].

The main challenges in human sensing research domain are *privacy* and *passiveness*. The former involves ethical issues to be addressed when people are recorded during spontaneous social interactions. This subject is outside the scope of this paper, but the *informed consent priciple* [51] should be always respected meaning that human subjects should always be aware of being recorded (e.g., like in broadcast material). Also, the subjects need to authorize explicitly the use and the diffusion of the data and they must have the right of deleting, partially or totally, the recordings where they are portrayed.

The second challenge relates to creating capture systems that are *passive* [125], i.e., unintrusive changing the behaviour of the recorded individuals as little as possible (in principle, the subjects should not even realize that they are recorded). This is a non-trivial problem because passive systems should involve only non-invasive sensors and the output of these is, in general, more difficult to process effectively. On the other hand, data captured by more invasive sensors are easier to process, but at the same time such recording setups tend to change the behaviour of the recorded individuals. Recording human naturalistic behaviour while eliciting specific behaviours and retaining the naturalism/ spontaneity of the behaviour is a very difficult problem tackled recently by several research groups [29][135].

3.2 Person Detection

The sensors used for data capture output signals that can be analyzed automatically to extract the behavioural cues underlying social signals and behaviours. In some cases, the signals corresponding to different individuals are separated at the origin. For example, physiological signals are recorded by invasive devices physically connected (e.g. through electrodes) to each person. Thus, the resulting signals can be attributed withouth ambiguity to a given individual. However, it happens more frequently that the signals contain spurious information (e.g. background noise), or they involve more than one individual. This is the case of the most commonly used sensors, microphones and cameras, and it requires the application of algorithms for *person detection* capable of isolating the signal segments corresponding to a single individual. The rest of this section discusses how this can be done for multiparty audio and video recordings.

3.2.1 Person Detection in Multiparty Audio Recordings

In the case of audio recordings, person detection is called *speaker segmentation* or *speaker diarization* and consists of splitting the speech recordings into intervals corresponding to a single voice, recognizing automatically *who talks when* (see [189] for an extensive survey). The speaker diarization is the most general case and it includes three main stages: the first is the segmentaion of the data into speech and non-speech segments, the second is the detection of the speaker transitions, and the third is the so-called *clustering*, i.e. the grouping of speaker segments corresponding to a single individual (i.e. to a single voice). In some cases (e.g. broadcast data), no silences are expected between one speaker and the following, thus the first step is not necessary. Systems that do not include a speech/ non-speech segmentation are typically referred to as *speaker segmentation* systems.

Speech and non-speech segmentation is typically performed using machine learning algorithms trained over different audio classes represented in the data (non-speech can include music, background noises, silence, etc.). Typically used techniques include Artificial Neural Networks [5], k Nearest Neighbours [107], Gaussian Mixture Models [61], etc. Most commonly used features include the basic information that can be extracted from any signal (e.g. energy and autocorrelation [156]), as well as the features typically extracted for speech recognition like *Mel Frequency Cepstrum Coefficients* (MFCC), *Linear Predictive Coding* (LPC), etc. (see [84] for an extensive survey).

The detection of the speaker transitions is performed by splitting the speech segments into short intervals (e.g. $2-3$ seconds) and by measuring the *difference* (see below) between two consecutive intervals: the highest values of the difference correspond to the speaker changes. The approaches is based on the assumption that the data include at least two speakers. If this is not the case, simple differences in the intonation or the background noise might be detected as speaker changes. The way the difference is estimated allows one to distinguish between the different approaches to the task: in general each interval is modeled using a single Gaussian (preferred to the GMMs because it simplifies the calculations) and the difference is estimated with the symmetric Kullback-Leibler divergence [14]. Alternative approaches [157] use a penalized-likelihood-ratio test to verify whether a single interval is modeled better by a single Gaussian (no speaker change) or by two Gaussians (speaker

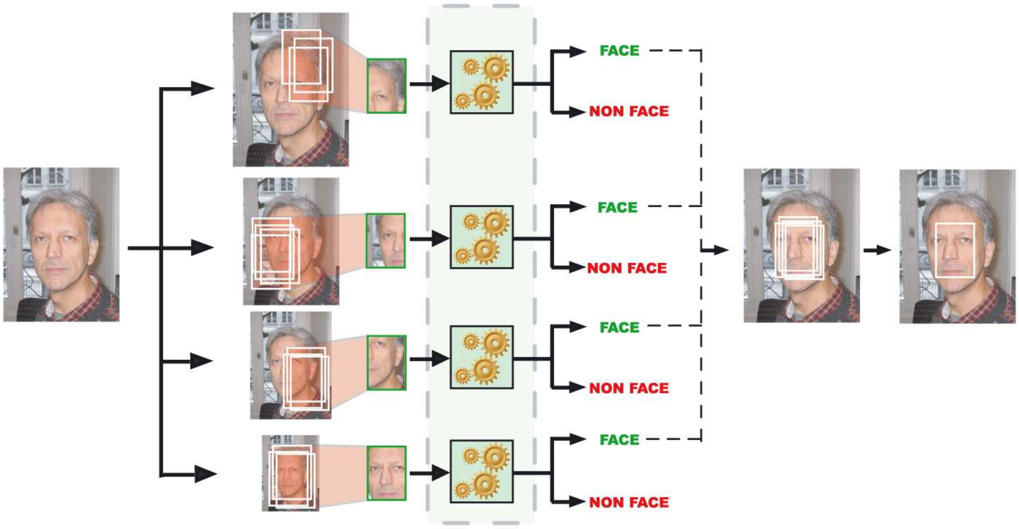


Fig. 7. Face detection. General scheme of an appearance based approach for face detection (picture from "*A tutorial on face detection and recognition*", by S.Marcel, www.idiap.ch/∼marcel).

change).

The last step of both speaker diarization and segmentation is *clustering*, i.e. grouping of the segments corresponding to a single voice into a unique cluster. This is commonly carried out through iterative approaches [14][117][157] where the clusters are initialized using the intervals between the speaker changes detected at the previous step (each interval is converted into a set of feature vectors using common speech processing techniques [84][156]), and then they are iteratively merged based on the similarity of the models used to represent them (single Gaussians or GMMs). The merging process is stopped when a criterion (e.g. the total likelihood of the cluster models starts to decrease) is met.

Most recent approaches tend to integrate three steps above-mentioned into a single framework by using hidden Markov models or dynamic Bayesian networks that align feature vectors extracted at regular time steps (e.g. 30 *ms*) and sequences of states corresponding to speakers in an unsupervised way [4][5].

3.2.2 Person Detection in Multiparty Video Recordings

In the case of video data, the person detection consists in locating faces or full human figures (that must be eventually tracked). Face detection is typically the first step towards facial expression analysis [139] or gaze behavior analysis [199] (see [81][215] for extensive surveys on face detection techniques). The detection of full human figures is more frequent in surveillance systems where the only important information is the movement of people across wide public spaces (e.g. train stations or streets) [62][115]. In the SSP framework, the detection of full human figures can be applied to study social signals related to space and distances (see Section 2.5), but to the best of our knowledge no attempts have been made yet in this direction.

Early approaches to face detection (see e.g. [161][181]) were based on the hypothesis that the presence of a face can be inferred from the pixel values. Thus they apply classifiers like Neural Networks or Support Vector Machines directly over small portions of the video frames (e.g. patches of 20 × 20 pixels) and map them into a face/ non-face classes. The main limitation of such techniques is that it is difficult to train classifiers for a *non-face* class that can include any kind of visual information (see Figure 7). Other approaches (e.g. [82][58]) try to detect human skin areas in images and then use their spatial distribution to identify faces and facial features (eyes, mouth and nose). The skin areas are detected by clustering the pixels in the color space. Alternative approaches (e.g. [101]) detect separately individual face elements (eyes, nose and mouth) and detect a face where such elements have the appropriate relative positions. These approaches are particularly robust to rotations because they depend on the relative position of face elements, rather than on the orientation with respect to a general reference frame in the image.

Another method that can handle out-of-plane head motions is the statistical method for 3D object detection proposed in [169]. Other such methods, which have been recently proposed, include those in [83][207]. Most of these methods emphasize statistical learning techniques and use appearance features. Arguably the most commonly employed face detector in automatic facial expression analysis is the real-time face detector proposed in [204]. This detector consists of a cascade of classifiers trained by AdaBoost. Each classifier employs integral image filters, also called "box filters," which are reminiscent of Haar Basis functions, and can be computed very fast at any location and scale. This is essential to the speed of the detector. For each stage in the cascade, a subset of features is chosen using a feature selection procedure based on AdaBoost. There are several adapted versions of the face detector described in [204] and the one that is often used is that proposed in [52].

The main challenge in detecting human figures is that people wear clothes of different color and appearance, so the pixel values are not a reliable feature for human body detection (see Figure 8). For this reason, some approaches extract features like the histograms of the edge directions (e.g. [34][223]) from local regions of the images (typically arranged in a regular grid), and then make a decision using classifiers like the Support Vector Machines. The same approach can be improved in the case of the videos, by adding motion information

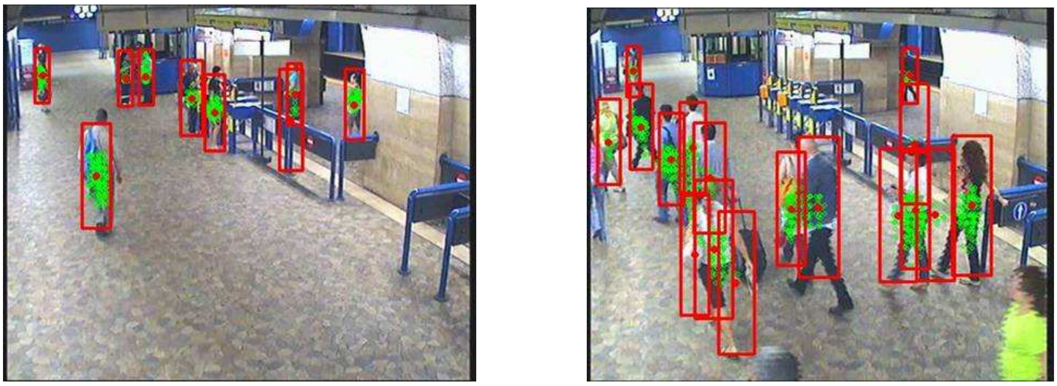


Fig. 8. People detection. Examples of people detection in public spaces (pictures from [216]).

extracted using the optical flow [35]. Other approaches (e.g. [114][194]) try to detect individual body parts and then use general rules of human body anatomy to reason about the body pose (individual body parts have always the same shape and they have the same relative position). For exhaustive survey, see [153].

3.3 Social Signals Detection

Once people in the observed scene are detected, the next step in the SSP process is to extract behavioural cues displayed by these people. Those cues include one or more synchronized audio and/or video signals that convey the information about the behaviour of the person. They are the actual source from which socially-relevant behavioural cues are extracted. The next sections discuss the main approaches to social signals detection from audio and/or visual signals captured while monitoring a person.

3.3.1 Detection of Social Signals from Physical Appearance

To the best of our knowledge, only few works address the problem of analyzing the physical appearance of people. However, these works do not aim to interpret this information in terms of social signals. Some approaches have tried to measure automatically the beauty of faces [1][44][73][75][211]. The work in [1] detects separately the face elements (eyes, lips, etc.) and then maps the ratios between their dimensions and distances into beauty judgments through classifiers trained on images assessed by humans. The work in [44] models the symmetry and the proportions of a face through the geometry of several landmarks (e.g. the corners of the eyes and the tip of the nose), and then applies machine learning techniques to match human judgments. Other techniques (e.g., [131]) use 3D models of human heads and the distance with respect to average faces extracted from large data sets to assess personal beauty. Faces closest to the average seem to be judged as more attractive than others.

Also few works were proposed where the body shape, the color of skin, hair, and clothes are extracted automatically (through a clustering of the pixels in the color space) for identification and tracking purposes [16][36][214]. However these works do not address social signal understanding and are therefore out of the scope of this paper.

3.3.2 Detection of Social Signals from Gesture and Posture

Gesture recognition is an active research domain in computer vision and pattern recognition research communities, but no efforts have been made, so far, to interpret the social information carried by gestural behaviours. In fact, the efforts are directed mostly towards the use of gestures as an alternative to keyboard and mouse to operate computers (e.g., [132][172][213]), or to the automatic reading of sign languages (e.g., [40][97]). Also few efforts have been reported towards human affect recognition from body gestures (for an overview see [76][221]). There are two main challenges in recognizing gestures: detecting the body parts involved in the gesture (in general the hands), and modeling the temporal dynamic of the gesture.

The first problem is addressed by selecting appropriate visual features: these include, e.g., histograms of oriented gradients (e.g., [183][184]), optical flow (e.g., [3][188]), spatio-temporal salient points (e.g., [129]) and space-time volumes (e.g., [67]). The second problem is addressed by using techniques such as Dynamic Time Warping (e.g., [129]), Hidden Markov Models (e.g. [3]), and Conditional Random Fields (e.g., [179]).

Like in the case of gestures, machine recognition of walking style (or gait) has been investigated as well, but only for purposes different from SSP, namely recognition and identification in biometric applications [100][102][206]. The common approach is to segment the silhouette of the human body into individual components (legs, arms, trunk, etc.), and then to represent their geometry during walking through vectors of distances [206], symmetry operators [78], geometric features of body and stride (e.g. distance between head and feets or pelvis) [17], etc.

Also automatic posture recognition has been addressed in few works, mostly aiming at surveillance [57] and activity recognition [206] (See [54][116][153] for extensive overviews of the past work in the field). However, there are few works where the posture is recognized as a social signal, namely to estimate the interest level of children learning to use computers [124], to recognize the affective state of people [38][74] (see [76][221] for exhaustive overview of research efforts in the field), and the influence of culture on affective postures [95].

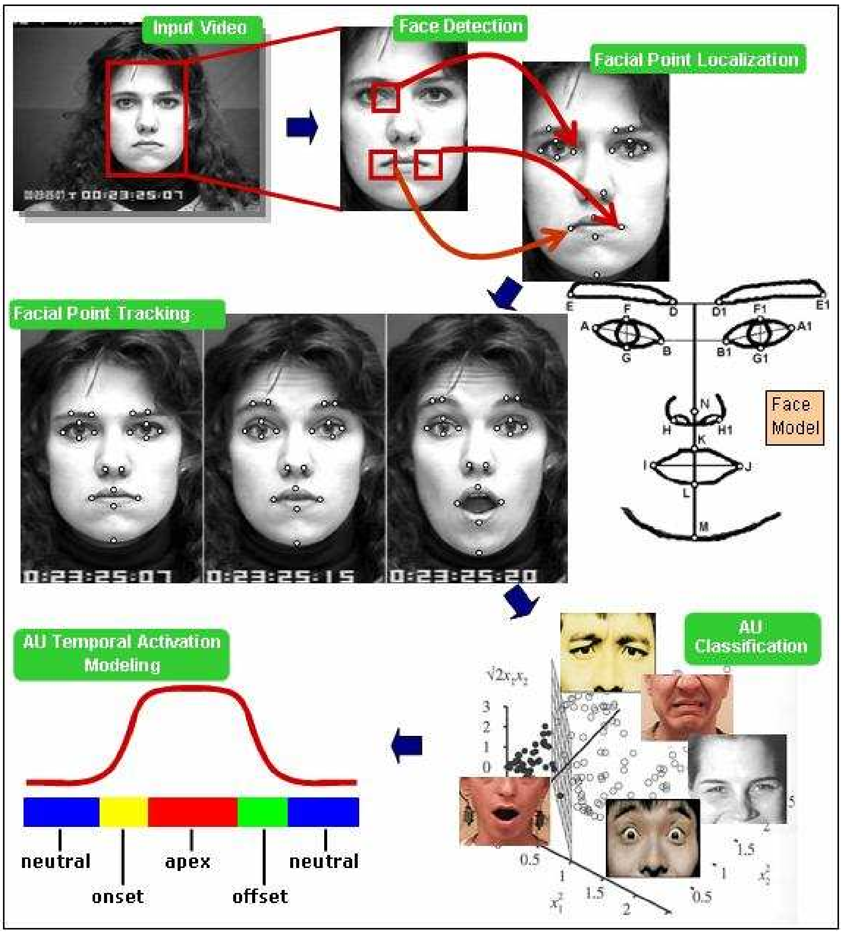


Fig. 9. AU detection. Outline of a geometric-feature-based system for detection of facial AUs and their temporal phases (onset, apex, offset, neutral) proposed in [196].

3.3.3 Detection of Social Signals from Gaze and Face

The problem of machine recognition of human gaze and facial behaviour includes three subproblem areas (see Figure 9): finding faces in the scene, extracting facial features from the detected face region, analyzing the motion of eyes and other facial features and/or the changes in the appearance of facial features, and classifying this information into some facial-behaviour-interpretative categories (e.g., facial muscle actions (AUs), emotions, social behaviours, etc.).

Numerous techniques have been developed for face detection, i.e., identification of all regions in the scene that contain a human face (see Section 3.2). Most of the proposed approaches to facial expression recognition are directed toward static, analytic, 2-D facial feature extraction [135][185]. The usually extracted facial features are either geometric features such as the shapes of the facial components (eyes, mouth, etc.) and the locations of facial fiducial points (corners of the eyes, mouth, etc.) or appearance features representing the texture of the facial skin in specific facial areas including wrinkles, bulges, and furrows. Appearance-based features include learned image filters from Independent Component Analysis (ICA), Principal Component Analysis (PCA), Local Feature Analysis (LFA), Gabor filters, integral image filters (also known as box-filters and Haar-like filters), features based on edge-oriented histograms, and similar [135]. Several efforts have been also reported which use both geometric and appearance features (e.g. [185]). These approaches to automatic facial expression analysis are referred to as hybrid methods. Although it has been reported that methods based on geometric features are often outperformed by those based on appearance features using, e.g., Gabor wavelets or eigenfaces, recent studies show that in some cases geometric features can outperform appearance-based ones [135][136]. Yet, it seems that using both geometric and appearance features might be the best choice in the case of certain facial expressions [136].

Contractions of facial muscles (i.e., AUs explained in section 2.3), which produce facial expressions, induce movements of the facial skin and changes in the location and/or appearance of facial features. Such changes can be detected by analyzing optical flow, facial-point- or facial-component-contour-tracking results, or by using an ensemble of classifiers trained to make decisions about the presence of certain changes based on the passed appearance features. The optical flow approach to describing face motion has the advantage of not requiring a facial feature extraction stage of processing. Dense flow information is available throughout the entire facial area, regardless of the existence of facial components, even in the areas of smooth texture such as the cheeks and the forehead. Because optical flow is the visible result of movement and is expressed in terms of velocity, it can be used to represent directly facial expressions. Many researchers adopted this approach (for overviews, see [135][139][185]). Until recently, standard optical flow techniques were arguably most commonly used for tracking facial characteristic points and contours as well. In order to address the limitations inherent in optical flow techniques such as the accumulation of error and the sensitivity to noise, occlusion, clutter, and changes in illumination, recent efforts in automatic facial expression recognition use sequential state estimation techniques (such as Kalman filter and Particle filter) to track facial feature points in image sequences [135][136][222].

Eventually, dense flow information, tracked movements of facial characteristic points, tracked changes in contours of facial components, and/or extracted appearance features are translated into a description of the displayed facial behaviour. This description (facial expression interpretation) is usually given either in terms of shown affective states (emotions) or in terms of activated facial muscles (AUs) underlying the displayed facial behaviour. Most facial expressions analyzers developed so far target human facial affect analysis and attempt to recognize a small set of prototypic emotional facial expressions like happiness and anger [140][221]. However, several promising prototype systems were reported that can recognize deliberately produced AUs in face images (for overviews, see [135][185]) and even few attempts towards recognition of spontaneously displayed AUs (e.g., [103][108]) and towards automatic discrimination between spontaneous and posed facial behaviour such as smiles [195], frowns [197], and pain [104], have been recently reported as well. Although still tentative, few studies have also been recently reported on separating emotional states from non-emotional states and on recognition of non-basic affective states in visual and audiovisual recordings of spontaneous human behaviour(e.g., for overview see [170][220]). However, although messages conveyed by AUs like winks, blinks, frowns, smiles, gaze exchanges, etc., can be interpreted in terms of social signals like turn taking, mirroring, empathy, engagement, etc., no efforts have been reported so far on automatic recognition of social behaviours in recordings of spontaneous facial behaviour. Hence, while the focus of the research in the field started to shift to automatic (non-basic-) emotion and AU recognition in spontaneous facial expressions (produced in a reflex-like manner), efforts towards automatic analysis of human social behaviour from visual and audiovisual recordings of human spontaneous behaviour are still to be made.

While the older methods for facial behaviour analysis employ simple approaches including expert rules and machine learning methods such as neural networks to classify the relevant information from the input data into some facial-expression-interpretative categories (e.g., basic emotion categories), the more recent (and often more advanced) methods employ probabilistic, statistical, and ensemble learning techniques, which seem to be particularly suitable for automatic facial expression recognition from face image sequences (for comprehensive overviews of the efforts in the field, see [135][221]). Note, however, the present systems for facial expression analysis typically depend on accurate head, face and facial feature tracking as input and are still very limited in performance and robustness.

3.3.4 Detection of Social Signals from Vocal Behaviour

The behavioural cues in speech include voice quality, vocalizations (linguistic and non-linguistic), and silences (see Section 2.4 for details). All of them have been the subject of extensive research in speech, but they have rarely been interpreted in terms of social information, even if they account for roughly 50% of the total time in spontaneous conversations [21]. With few exceptions, the detection of vocal behaviour has aimed at the improvement of Automatic Speech Recognition (ASR) systems, where the vocal non-verbal behaviour represents a form of noise rather than an information.

The voice quality corresponds to the prosody and includes three major aspects, often called the *Big Three*: *pitch*, *tempo* and *energy* [31]. The pitch is the frequency of oscillation of the vocal folds during the emission of voice and it is the characteristic that alone contributes more than anything else to the sound of a voice [120][150]. The measurement of the pitch, often called *fundamental frequency* (or $F0$) because most of the speech energy is concentrated over components corresponding to its integer multiples, can be performed with several standard methods proposed in the literature [84][156]. The pitch is typically extracted as the frequency corresponding to the first peak of the Fourier Transform of short analysis windows (in general 30 ms). Several tools publicly available on the web, e.g. *Wavesurfer*[1] [177] and *Praat*[2] [18], implement algorithms extracting the pitch from speech recordings. The tempo is typically estimated through the speaking rate, i.e. the number of phonetically relevant units, e.g. vowels [149], per second. Other methods are based on measures extracted from the speech signal like the first spectral moment of the energy [121][122] and typically aim at improving speech recognition systems through speaking rate adaptation. The energy, is a property of any digital signal and simply corresponds to the sum of the square values of the samples [156].

No major efforts have been made so far, to the best of our knowledge, to detect the non-linguistic vocalizations (see Section 2.4). The only exceptions are laughter [92][191][192] due to its ubiquitous presence in social interactions, and crying [118][134]. Laugther is detected by applying binary classifiers such as Support Vector Machines to features commonly applied in speech recognition like the *Mel Frequency Cepstral Coefficients* [92], or by modeling *Perceptual Linear Prediction* features with Gaussian Mixture Models and Neural Networks [191][192]. These efforts are based only on audio signals, but few pioneering efforts towards audiovisual recognition of non-linguistic vocal outbursts have been recently reported. A laughter detector which combines the outputs of an audio-based detector that uses MFCC audio features and a visual detector that uses spatial locations of facial feature points is proposed in [86]. They attained 80% average recall rate using 3 sequences of 3 subjects in a person dependent way. In [147] decision level and feature level fusion with audio- and video-only laughter detection are compared. The work uses PLP features and displacements of the tracked facial points as the audio and visual features respectively. Both fusion approaches outperformed single-modal detectors, achieving on average 84% recall in a person-independent test. Extension of this work based on utilisation of temporal features has been reported in [148].

Linguistic vocalizations have been investigated to detect hesitations in spontaneous speech [105][173][174] with the main purpose of improving speech recognition systems. The disfluencies are typically detected by mapping acous-

[1] Publicly available at `http://www.speech.kth.se/wavesurfer/`.

[2] Publicly available at `http://www.praat.org`.

tic observations (e.g. pitch and energy) into classes of interest with classifiers like Neural Networks or Support Vector Machines. The detection of silence is one of the earliest tasks studied in speech analysis and robust algorithms, based on the distribution of the energy, have been developed since the earliest times of digital signal processing [155][156]. Another important aspect of vocal behaviour, i.e. the turn taking, is typically a side-product of the speaker diarization or segmentation step (see Section 3.2).

3.3.5 Detection of Social Signals in Space and Environment

Physical proximity information has been used in *reality mining* applications (see Section 4) as a social cue accounting for the simple presence or absence of interaction between people [43][144]. These works use special cellular phones equipped to sense the presence of similar devices in the vicinity. Automatic detection of seating arrangements has been proposed as a cue for retrieving meeting recordings in [88]. Also, several video-surveillance approaches developed to track people across public spaces can potentially be used for detection of social signals related to the use of the available space (see Section 3.2 for more details).

3.4 Context Sensing and Social Behaviour Understanding

Context plays a crucial role in understanding of human behavioural signals, since they are easily misinterpreted if the information about the situation in which the shown behavioural cues have been displayed is not taken into account. For example, a smile can be a display of politeness (social signal), contentedness (affective cue), joy of seeing a friend (affective cue/ social signal), irony/ irritation (affective cue/ social signal), empathy (emotional response/ social signal), greeting (social signal), to mention just a few possibilities. It is obvious from these examples that in order to determine the communicative intention conveyed by an observed behavioural cue, one must know the context in which the observed signal has been displayed: where the expresser is (outside, inside, in the car, in the kitchen, etc.), what his or her current task is, are other people involved, when the signal has been displayed (i.e., what is the timing of displayed behavioural signals with respect to changes in the environment), and who the expresser is (i.e., it is not probable that each of us will express a particular affective state by modulating the same communicative signals in the same way).

Note, however, 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 recognizing communicative intention including social behaviours, affective 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 [138][137]), 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 a surprise that most of the present approaches to machine analysis of human behaviour are neither context-sensitive nor suitable for handling longer time scales. Hence, the focus of future research efforts in the field should be primarily on tackling the problem of context-constrained analysis of multimodal social signals shown over longer temporal intervals. Here, we would like to stress the importance of two issues: realizing temporal analysis of social signals and achieving temporal multimodal data fusion.

Temporal dynamics of social behavioural cues (i.e., their timing, co-occurrence, speed, etc.) are crucial for the interpretation of the observed social behaviour [8] [50]. However, present methods for human behaviour analysis do not address the *when* context question - dynamics of displayed behavioural signals is usually not taken into account when analyzing the observed behaviour, let alone analysing the timing of displayed behavioural signals with respect to changes in the environment. Exceptions of this rule include few recent studies on modelling semantic and temporal relationships between facial gestures (i.e., AUs, see Section 2.3) forming a facial expression (e.g. [187]), few studies on discrimination between spontaneous and posed facial gestures like brow actions and smiles based on temporal dynamics of target facial gestures, head and shoulder gestures [195][197], and few studies on multimodal analysis of audio and visual dynamic behaviours for emotion recognition [221]. In general, as already mentioned above, present methods cannot handle longer time scales, model grammars of observed persons behaviours, and take temporal and context-dependent evolvement of observations into account for more robust performance. These remain major challenges facing the researchers in the field.

Social signals are spoken and wordless messages like head nods, winks, *uh* and *yeah* utterances, which are sent by means of body gestures and postures, facial expressions and gaze, vocal expressions and speech. Hence, automated analyzers of human social signals and social behaviours should be multimodal, fusing and analyzing verbal and non-verbal interactive signals coming from different modalities (speech, body gestures, facial and vocal expressions). 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 have been proposed that aim at making use of the correlation between audio and visual data streams, and relax the requirement of synchronization of these streams(e.g., [55][220]). However, how to model multimodal fusion on multiple time scales and how to model temporal correlations within and between different modalities is largely unexplored. A much broader focus on the issues relevant to multimodal temporal fusion is needed including the optimal level of integrating these different streams, the optimal function for the integration, and how estimations of reliability of each stream can be included in the inference process. In addition, how to build context-dependent multimodal fusion is another open and highly relevant issue.

4 Main Applications of Social Signal Processing

The expression *Social Signal Processing* has been used for the first time in [145] to group under a collective definition several pioneering works of Alex Pentland and his group at MIT. Some of their works [142][143] extracted automatically the social signals detected in dyadic interactions to predict with an accuracy of more than 70% the outcome of salary negotiations, hiring interviews, and speed-dating conversations [33]. These works are based on vocal social signals including overall *activity* (the total amount of energy in the speech signals), *influence* (the statistical influence of one person on the speaking patterns of the others), *consistency* (stability of the speaking patterns of each person), and *mimicry* (the imitation between people involved in the interactions). Other works used cellular phones equipped with proximity sensors and vocal activity detectors to perform what came to be called *reality mining*, or *social sensing*, i.e., automatic analysis of everyday social interactions in groups of several tens of individuals [43][144]. Individuals are represented through vectors accounting for their proximity with others and for the places they are (home, work, etc.). The application of the Principal Component Analysis to such vectors leads to the so called *eigenbehaviours* [43].

In approximately the same period, few other groups worked on the analysis of social interactions in multimedia recordings targeting three main areas: analysis of interactions in small groups, recognition of roles, and sensing of users interest in computer characters. Results for problems that have been addressed by more than one group are reported in Table 2.

The research on interactions in small groups has focused on the detection of dominant persons and on the recognition of collective actions. The problem of dominance is addressed in [85][160], where multimodal approaches combine several nonverbal features, mainly speaking energy and body movement,

Ref.	Data	Time	Source	Performance
Role Recognition				
[13]	Meetings (2 recordings, 3 roles)	0h.45m	acted	50.0% of segments (up to 60 seconds long) correctly classified
[15]	NIST TREC SDR Corpus (35 recordings, publicly available 3 roles)	17h.00m	spontaneous	80.0% of the news stories correctly labeled in terms of role
[42]	The Survival Corpus (11 recordings, publicly available, 5 roles)	4h.30m	acted	90% of precision in role assignment
[59]	AMI Meeting Corpus (138 recordings, publicly available, 4 roles)	45h.00m	acted	67.9% of the data time correctly labeled in terms of role
[200]	Radio news bulletins (96 recordings, 6 roles)	25h.00m	spontaneous	80% of the data time correctly labeled in terms of role
[210]	Movies (3 recordings , 4 roles)	5h.46m	spontaneous	95% of roles correctly assigned
[218]	The Survival Corpus (11 recordings, publicly available, 5 roles)	4h.30m	spontaneous	Up to 65% of analysis windows (around 10 seconds long) correctly classified in terms of role
Collective Action Recognition				
[39]	Meetings (30 recordings, publicly available)	2h.30m	acted	Action Error Rate of 12.5%
[111]	Meetings (60 recordings, publicly available)	5h.00m	acted	Action Error Rate of 8.9%
Interest Level Detection				
[60]	Meetings (50 recordings, 3 interest levels)	unknown	acted	75% Precision
[124]	Children playing with video games (10 recordings, 3 interest levels)	3h.20m	spontaneous	82% recognition rate

to identify at each moment who is the dominant individual. The same kind of features has been applied in [39][111] to recognize the actions performed in meetings like discussions, presentations, etc. In both above applications,

Ref.	Data	Time	Source	Performance
Dominance Detection				
[85]	Meetings from AMI Corpus (34 segments)	3h.00m	acted	Most dominant person correctly detected in 85% of segments
[159]	Meetings (8 meetings)	1h.35m	acted	Most dominant person correctly detected in 75% of meetings
[160]	Meetings (40 recordings)	20h.00m	acted	Most dominant person correctly detected in 60% of meetings

Table 2
Results obtained by Social Signal Processing works. For each work, information about the data (kind of interaction, availability, size, the total duration of the recordings), whether it is real-world or acted data, and the reported performance are summarized.

the combination of the information extracted from different modalities is performed with algorithms Dynamic Bayesian Networks [126] and layered Hidden Markov Models [130].

The recognition of roles has been addressed in two main contexts: broadcast material [15][53][200][210] and small scale meetings [13][42][59][218]. The works in [53][200][210] apply Social Network Analysis [209] to detect the role of people in broadcast news and movies, respectively. The social networks are extracted automatically using speaker adjacences in [53][200] (people are linked when they are adjacent in the sequence of the speakers), and face recognition [210] (people are linked when their faces appear together in a scene). The approach in [15] recognizes the roles of speakers in broadcast news using vocal behaviour (turn taking patterns and intervention duration) and lexical features. The recognition is performed using boosting techniques. The roles in meetings are recognized with a classifier tree applied to nonverbal behaviour features (overlapping speech, number of interventions, back-channeling, etc.) in the case of [13], while speech and fidgeting activity are fed to a multi-SVM classifier in [42][218]. A technique based on the combination of Social Network Analysis and lexical modeling (Boostexter) is presented in [59].

The reaction of users to social signals exhibited by computer characters has been investigated in several works showing that people tend to behave with Embodied Conversational Agents (ECA) as they behave with other humans. The effectiveness of computers as *social actors*, i.e., entities involved in the same kind of interactions as humans, has been explored in [127][128], where computers have been shown to be attributed a personality and to elicit the same reactions as those elicited by persons. Similar effects have been shown in [28][133], where children interacting with computers have modified their voice to match the speaking characteristics of the animated ECA, showing adaptation patterns typical of human-human interactions [20]. Further evidence of the same phenomenon is available in [10][11], where the interaction between humans and ECA is shown to include the *Chameleon effect* [22], i.e. the mutual imitation of individuals due to reciprocal appreciation or to the influence of one individual on the other.

Psychologists have compared the performance of humans and machines in detecting socially relevant information like gender and movements associated to emotional states [65][151][152]. The results show that machines tend to have a constant performance across a wide range of conditions (different behavioral cues at disposition), while humans have dramatic changes in performance (sometimes dropping at chance level) when certain behavioral cues are no longer at disposition. This seems to suggest that humans do not use the behavioral cues actually at their disposition, but rather rely on task specific behavioral cues without which the tasks cannot be performed effectively [65][151][152]. In contrast, automatic approaches (in particular those based on machine learning) are built to rely on any available behavioral cue and their performance simply depends on how much the available cues are actually correlated with the targeted social information.

5 Conclusions and Future Challenges

Social Signal Processing has the ambitious goal of bringing social intelligence [6][66] in computers. The first results in this research domain have been sufficiently impressive to attract the praise of the technology [69] and business [19] communities. What is more important is that they have established a viable interface between human sciences and engineering - social interactions and behaviours, although complex and rooted in the deepest aspects of human psychology, can be analyzed automatically with the help of computers. This interdisciplinarity is, in our opinion, the most important result of research in SSP so far. In fact, the pioneering contributions in SSP [142][143] have shown that the social signals, typically described as so elusive and subtle that only trained psychologists can recognize them [63], 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.

However, although fundamental, these are only the first steps and the journey towards *artificial social intelligence* and *socially-aware computing* is still long. In the rest of this section we discuss four challenges facing the researchers in the field, for which we believe are the crucial turnover issues that need to be addressed before the research in the field can enter its next phase - the deployment phase.

The first issue relates to *tightening of the collaboration between social scientists and engineers*. The analysis of human behaviour in general, and social behaviour in particular, is an inherently multidisciplinary problem [138][221]. More specifically no automatic analysis of social interactions is possible without taking into account the basic mechanisms governing social behaviours that the psychologists have investigated for decades, such as the *chameleon effect* (mutual imitation of people aimed at showing liking or affiliation) [22][99], the interpersonal adaptation (mutual accommodation of behavioural patterns between interacting individuals) [20][71], the interactional synchrony (degree of coordination during interactions) [93], the presence or roles in groups [12][186], the dynamics of conversations [154][217], etc. The collaboration between technology and social sciences demands a mutual effort of the two disciplines. On one hand, engineers need to include the social sciences in their reflection, while on the other hand, social scientists need to formulate their findings in a form useful for engineers and their work on SSP.

The second issue relates to the need of implementing *multi-cue, multi-modal approaches* to SSP. Nonverbal behaviours cannot be read like words in a book [96][158]; they are not unequivocally associated to a specific meaning and their appearance can depend on factors that have nothing to do with social behaviour. Postures correspond in general to social attitudes, but sometimes they are simply comfortable [166], physical distances typically account for social distances, but sometimes they are simply the effect of physical constraints [77]. Moreover, the same signal can correspond to different social behaviour interpretations depending on context and culture [190] (although many advocate that social signals are natural rather than cultural [171]). In other words, social signals are intrinsically ambiguous and the best way to deal with such problem is to use multiple behavioural cues extracted from multiple modalities. Numerous studies have theoretically and empirically demonstrated the advantage of integration of multiple modalities (at least audio and visual) in human behaviour analysis over single modalities (e.g., [162]). This corresponds, from a technological point of view, to the combination of different classifiers that has extensivley been shown to be more effective than single classifiers, as long as they are sufficiently *diverse*, i.e., account for different aspects of the same problem [94]. It is therefore not surprising that some of the most successful works in SSP so far use features extracted from multiple modalities like in [39][85][111]. Note, however, that the relative contributions of different modalities and the related behavioural cues to affect judgment of displayed behaviour depend on the targeted behavioural category and the context in which the behaviour occurs [49][162].

The third issue relates to *the use of real-world data*. Both psychologists and engineers tend to produce their data in laboratories and artificial settings (see e.g., [33][68][111]), in order to limit parasitic effects and elicit the specific phenomena they want to observe. However, this is likely to simplify excessively the situation and to improve artificially the performance of the automatic approaches. Social interaction is one of the most ubiquitous phenomena in the world - the media (radio and television) show almost exclusively social interactions (debates, movies, talk-shows) [123]. Also other, less common kinds of data are centered on social interactions, e.g., meeting recordings [110], surveillance material [87], and similar. The use of real-world data will allow analysis of interactions that have an actual impact on the life of the participants, thus will show the actual effects of goals and motivations that typically drive human behaviour. This includes also the analysis of *group interactions*, a task difficult from both technological and social point of view because it involves the need of observing multiple people involved in a large number of one-to-one interactions.

The last, but not least, challenging issue relates to the *the identification of applications likely to benefit from SSP*. Applications have the important advantage of linking the effectiveness of detecting social signals to the reality. For example, one of the earliest applications is the prediction of the outcome in transactions recorded at a call center 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 [19]. This can have not only a positive impact on the marketplace, but 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 [178]).

Acknowledgements. The work of Dr. Vinciarelli is supported by the Swiss National Science Foundation through the National Center of Competence in Research on Interactive Multimodal Information Management (IM2). The work of Dr. Pantic is supported in part by the EC's 7^{th} Framework Programme (FP7/2007-2013) under grant agreement no. 211486 (SEMAINE), and the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). The research that has led to this work has been supported in part by the European Community's Seventh Framework Programme (FP7/2007-2013), under grant agreement no. 231287 (SSPNet).

References

[1] P. Aarabi, D. Hughes, K. Mohajer, and M. Emami. The automatic measurement of facial beauty. In *Proceedings of IEEE International*

Conference on Systems, Man, and Cybernetics, pages 2644–2647, 2001.

[2] R. Adolphs. Cognitive neuroscience of human social behaviour. *Nature Reviews Neuroscience*, 4(3):165–178, 2003.

[3] M. Ahmad and S.-W. Lee. Human action recognition using shape and CLG-motion flow from multi-view image sequences. *Pattern Recognition*, 41(7):2237–2252, 2008.

[4] J. Ajmera. *Robust Audio Segmentation.* PhD thesis, École Polytechnique Fédérale de Lausanne (EPFL), 2004.

[5] J. Ajmera, I. McCowan, and H. Bourlard. Speech/music segmentation using entropy and dynamism features in a HMM classification framework. *Speech Communication*, 40(3):351–363, 2003.

[6] K. Albrecht. *Social Intelligence: The new science of success.* John Wiley & Sons Ltd, 2005.

[7] N. Ambady, F. Bernieri, and J. Richeson. Towards a histology of social behavior: judgmental accuracy from thin slices of behavior. In M.P. Zanna, editor, *Advances in Experimental Social Psychology*, pages 201–272. 2000.

[8] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. *Psychological Bulletin*, 111(2):256–274, 1992.

[9] M. Argyle. *The Psychology of Interpersonal Behaviour.* Penguin, 1967.

[10] J.N. Bailenson and N. Yee. Virtual interpersonal touch and digital chameleons. *Journal of Nonverbal Behavior*, 31(4):225–242, 2007.

[11] J.N. Bailenson, N. Yee, K. Patel, and A.C. Beall. Detecting digital chameleons. *Computers in Human Behavior*, 24(1):66–87, 2008.

[12] R.F. Bales. *Interaction Process Analysis: A Method for the Study of Small Groups.* Addison-Wesley, 1950.

[13] S. Banerjee and A.I. Rudnicky. Using simple speech based features to detect the state of a meeting and the roles of the meeting participants. In *Proceedings of International Conference on Spoken Language Processing*, pages 2189–2192, 2004.

[14] C. Barras, X. Zhu, S. Meignier, and J.L. Gauvain. Improving speaker diarization. In *Proceedings of the Rich Transcription Workshop*, 2004.

[15] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker. The rules behind the roles: identifying speaker roles in radio broadcasts. In *Proceedings of American Association of Artificial Intelligence Symposium*, pages 679–684, 2000.

[16] C. Ben Abdelkader and Y. Yacoob. Statistical estimation of human anthropometry from a single uncalibrated image. In K. Franke, S. Petrovic, and A. Abraham, editors, *Computational Forensics.* Springer Verlag, 2009.

[17] A.F. Bobick and A. Johnson. Gait recognition using static activity-specific parameters. In *Proceedings of Computer Vision and Pattern Recognition*, pages 423–430, 2001.

[18] P. Boersma and D. Weenink. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2001.

[19] M. Buchanan. The science of subtle signals. *Strategy+Business*, 48:68–77, 2007.

[20] J.K. Burgoon, L.A. Stern, and L. Dillman. *Interpersonal Adaptation: Dyadic Interaction Patterns*. Cambridge University Press, 1995.

[21] N. Campbell. Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Speech and Language Processing*, 14(4):1171–1178, 2006.

[22] T.L. Chartrand and J.A. Bargh. The chameleon effect: the perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910, 1999.

[23] J.F. Cohn. Foundations of human computing: facial expression and emotion. In *Proceedings of the ACM International Conference on Multimodal Interfaces*, pages 233–238, 2006.

[24] J.F. Cohn and P. Ekman. Measuring facial action by manual coding, facial EMG, and automatic facial image analysis. In J.A. Harrigan, R. Rosenthal, and K.R. Scherer, editors, *Handbook of nonverbal behavior research methods in the affective sciences*, pages 9–64. 2005.

[25] J.B. Cortes and F.M. Gatti. Physique and self-description of temperament. *Journal of Consulting Psychology*, 29(5):432–439, 1965.

[26] M. Costa, W. Dinsbach, A.S.R. Manstead, and P.E.R. Bitti. Social presence, embarrassment, and nonverbal behavior. *Journal of Nonverbal Behavior*, 25(4):225–240, 2001.

[27] M. Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior*, 28(2):117–139, 2004.

[28] R. Coulston, S. Oviatt, and C. Darves. Amplitude convergence in children's conversational speech with animated personas. In *International Conference on Spoken Language Processing*, pages 2689–2692, 2002.

[29] R. Cowie. Building the databases needed to understand rich, spontaneous human behaviour. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[30] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, 2001.

[31] D. Crystal. *Prosodic Systems and Intonation in English*. Cambridge University Press, 1969.

[32] D.W. Cunningham, M. Kleiner, H.H. Bülthoff, and C. Wallraven. The components of conversational facial expressions. *Proceedings of the Symposium on Applied Perception in Graphics and Visualization*, pages 143–150, 2004.

[33] J.R. Curhan and A. Pentland. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology*, 92(3):802–811, 2007.

[34] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[35] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Proceedings of the European Conference on Computer Vision*, pages 428–441, 2006.

[36] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. *International Journal of Computer Vision*, 37(2):175–185, 2000.

[37] C. Darwin. *The Expression of the Emotions in Man and Animals*. J. Murray, 1872.

[38] R. De Silva and N. Bianchi-Berthouze. Modeling human affective postures: an information theoretic characterization of posture features. *Journal of Computational Animation and Virtual World*, 15(3-4):269–276, 2004.

[39] A. Dielmann and S. Renals. Automatic meeting segmentation using dynamic bayesian networks. *IEEE Transactions on Multimedia*, 9(1):25, 2007.

[40] L. Ding and A.M. Martinez. Recovering the linguistic components of the manual signs in american sign language. In *Proceedings of IEEE International Conference on Advanced Video and Signal-based Surveillance*, pages 447–452, 2007.

[41] K. Dion, E. Berscheid, and E. Walster. What is beautiful is good. *Journal of Personality and Social Psychology*, 24(3):285–290, 1972.

[42] W. Dong, B. Lepri, A. Cappelletti, A.S. Pentland, F. Pianesi, and M. Zancanaro. Using the influence model to recognize functional roles in meetings. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 271–278, 2007.

[43] N. Eagle and A. Pentland. Reality mining: sensing complex social signals. *Journal of Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[44] Y. Eisenthal, G. Dror, and E. Ruppin. Facial attractiveness: Beauty and the machine. *Neural Computation*, 18(1):119–142, 2005.

[45] P. Ekman, editor. *Emotion in the human face.* Cambridge University Press, 1982.

[46] P. Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000(1):205–221, 2003.

[47] P. Ekman and W.V. Friesen. The repertoire of nonverbal behavior. *Semiotica*, 1:49–98, 1969.

[48] P. Ekman, W.V. Friesen, and J.C. Hager. *Facial Action Coding System (FACS): Manual.* A Human Face, Salt Lake City (USA), 2002.

[49] P. Ekman, T.S. Huang, T.J. Sejnowski, and J.C. Hager, editors. *Final Report to NSF of the Planning Workshop on Facial Expression Understanding.* Human Interaction Laboratory, University of California, San Francisco, 1993.

[50] P. Ekman and E.L. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS).* Oxford University Press, 2005.

[51] R.R. Faden, T.L. Beauchamp, and N.M.P. King. *A History and Theory of Informed Consent.* Oxford University Press, 1986.

[52] I.R. Fasel, B. Fortenberry, and J.R. Movellan. A generative framework for real time object detection and classification. *Computer Vision and Image Understanding*, 98(1):181–210, 2005.

[53] S. Favre, H. Salamin, J. Dines, and A. Vinciarelli. Role recognition in multiparty recordings using Social Affiliation Networks and discrete distributions. In *Proceedings of the ACM International Conference on Multimodal Interfaces*, pages 29–36, 2008.

[54] D.A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan. Computational studies of human motion part 1: Tracking and motion synthesis. *Foundations and Trends in Computer Graphics and Vision*, 1(2):77–254, 2006.

[55] N. Fragopanagos and J.G. Taylor. Emotion recognition in human–computer interaction. *Neural Networks*, 18(4):389–405, 2005.

[56] M.G. Frank and P. Ekman. Appearing truthful generalizes across different deception situations. *Journal of Personality and Social Psychology*, 86(3):486–495, 2004.

[57] T. Gandhi and M.M. Trivedi. Pedestrian protection systems: Issues, survey, and challenges. *IEEE Transactions On Intelligent Transportation Systems*, 8(3):413–430, 2007.

[58] C. Garcia and G. Tziritas. Face detection using quantized skin color regions merging and wavelet packet analysis. *IEEE Transactions on Multimedia*, 1(3):264–277, 1999.

[59] N. Garg, S. Favre, H. Salamin, D. Hakkani-Tür, and A. Vinciarelli. 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, 2008.

[60] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio. Detecting group interest-level in meetings. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 489–492, 2005.

[61] J.L. Gauvain, L.F. Lamel, and G. Adda. Partitioning and transcription of broadcast news data. In *Proceedings of International Conference on Spoken Language Processing*, pages 1335–1338, 1998.

[62] D.M. Gavrila. Visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

[63] M. Gladwell. *Blink: The Power of Thinking without Thinking.* Little Brown & Company, 2005.

[64] C.R. Glass, T.V. Merluzzi, J.L. Biever, and K.H. Larsen. Cognitive assessment of social anxiety: Development and validation of a self-statement questionnaire. *Cognitive Therapy and Research*, 6(1):37–55, 1982.

[65] J.M. Gold, D. Tadin, S.C. Cook, and R.B. Blake. The efficiency of biological motion perception. *Perception and Psychophysics*, 70(1):88–95, 2008.

[66] D. Goleman. *Social intelligence.* Hutchinson, 2006.

[67] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, 2007.

[68] J.E. Grahe and F.J. Bernieri. The importance of nonverbal cues in judging rapport. *Journal of Nonverbal Behavior*, 23(4):253–269, 1999.

[69] K. Greene. 10 emerging technologies 2008. *MIT Technology Review*, february 2008.

[70] M.R. Greenwald, A.G.and Banaji. Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1):4–27, 1995.

[71] S.W. Gregory, K. Dagan, and S. Webster. Evaluating the relation of vocal accommodation in conversation partners fundamental frequencies to perceptions of communication quality. *Journal of Nonverbal Behavior*, 21(1):23–43, 1997.

[72] M.M. Gross, E.A. Crane, and B.L. Fredrickson. Effect of felt and recognized emotions on body movements during walking. In *Proceedings of the International Conference on The Expression of Emotions in Health and Disease*, 2007.

[73] H. Gunes and M. Piccardi. Assessing facial beauty through proportion analysis by image processing and supervised learning. *International Journal of Human-Computer Studies*, 64(12):1184–1199, 2006.

[74] H. Gunes and M. Piccardi. Bi-modal emotion reognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4):1334–1345, 2007.

[75] H. Gunes, M. Piccardi, and T. Jan. Comparative beauty classification for pre-surgery planning. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pages 2168–2174, 2004.

[76] H. Gunes, M. Piccardi, and M. Pantic. From the lab to the real world: Affect recognition using multiple cues and modalities. In J. Or, editor, *Affective Computing: Focus on Emotion Expression, Synthesis, and Recognition*, pages 185–218. 2008.

[77] E.T. Hall. *The silent language*. Doubleday, 1959.

[78] J.B. Hayfron-Acquah, M.S. Nixon, and J.N. Carter. Automatic gait recognition by symmetry analysis. *Pattern Recognition Letters*, 24(13):2175–2183, 2003.

[79] J. Hirschberg. Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, 63(1-2):305–340, 1993.

[80] J. Hirschberg and B. Grosz. Intonational features of local and global discourse structure. In *Proceedings of the Speech and Natural Language Workshop*, pages 441–446, 1992.

[81] E. Hjelmas and B.K. Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83(3):236–274, 2001.

[82] C.R.L. Hsu, M. Abdel-Mottaleb, and A.K. Jain. Face detection in colour images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):696–706, 2002.

[83] K.S. Huang and M.M. Trivedi. Robust real-time detection, tracking, and pose estimation of faces in video streams. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 965–968, 2004.

[84] X. Huang, A. Acero, and H.W. Hon. *Spoken language processing*. Prentice Hall, 2001.

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

[86] A. Ito, X. Wang, M. Suzuki, and S. Makino. Smile and laughter recognition using speech processing and face recognition from conversation video. In *Proceedings of th International Conference on Cyberworlds*, pages 437–444, 2005.

[87] Y. Ivanov, C. Stauffer, A. Bobick, and W.E.L. Grimson. Video surveillance of interactions. In *Proceedings of the Workshop on Visual Surveillance at Computer Vision and Pattern Recognition*, 1999.

[88] A. Jaimes, K. Omura, T. Nagamine, and K. Hirata. Memory cues for meeting video retrieval. In *Proceedings of Workshop on Continuous Archival and Retrieval of Personal Experiences*, pages 74–85, 2004.

[89] D. Keltner and P. Ekman. Facial expression of emotion. In M. Lewis and J.M. Haviland-Jones, editors, *Handbook of Emotions*, pages 236–249. 2000.

[90] D. Keltner and J. Haidt. Social functions of emotions at four levels of analysis. *Cognition and Emotion*, 13(5):505–521, 1999.

[91] D. Keltner and A.M. Kring. Emotion, social function, and psychopathology. *Review of General Psychology*, 2(3):320–342, 1998.

[92] L. Kennedy and D. Ellis. Laughter detection in meetings. In *Proceedings of the NIST Meeting Recognition Workshop*, 2004.

[93] M. Kimura and I. Daibo. Interactional synchrony in conversations about emotional episodes: A measurement by the between-participants pseudosynchrony experimental paradigm. *Journal of Nonverbal Behavior*, 30(3):115–126, 2006.

[94] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[95] A. Kleinsmith, R. De Silva, and N. Bianchi-Berthouze. Cross-cultural differences in recognizing affect from body posture. *Interacting with Computers*, 18(6):1371–1389, 2006.

[96] M.L. Knapp and J.A. Hall. *Nonverbal Communication in Human Interaction.* Harcourt Brace College Publishers, 1972.

[97] W.W. Kong and S. Ranganath. Automatic hand trajectory segmentation and phoneme transcription for sign language. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[98] Z. Kunda. *Social Cognition.* MIT Press, 1999.

[99] J.L. Lakin, V.E. Jefferis, C.M. Cheng, and T.L. Chartrand. The Chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of Nonverbal Behavior*, 27(3):145–162, 2003.

[100] L. Lee and W.E.L. Grimson. Gait analysis for recognition and classification. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 148–155, 2002.

[101] T.K. Leung, M.C. Burl, and P. Perona. Finding faces in cluttered scenes using random labeled graphmatching. In *Proceedings of the International Conference on Computer Vision*, pages 637–644, 1995.

[102] X. Li, S.J. Maybank, S. Yan, D. Tao, and D. Xu. Gait components and their application to gender recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 38(2):145–155, 2008.

[103] G. Littlewort, M.S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, 2006.

[104] G.C. Littlewort, M.S. Bartlett, and K. Lee. Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 15–21, 2007.

[105] Y. Liu, E. Shriberg, A. Stolcke, and M. Harper. Comparing HMM, maximum entropy, and conditional random fields for disfluency detection. In *Proceeedings of the European Conference on Speech Communication and Technology*, 2005.

[106] D.F. Lott and R. Sommer. Seating arrangements and status. *Journal of Personality and Social Psychology*, 7(1):90–95, 1967.

[107] L. Lu, H.J. Zhang, and H. Jiang. Content analysis for audio classification and segmentation. *IEEE Transactions on Speech and Audio Processing*, 10(7):504–516, 2002.

[108] S. Lucey, A.B. Ashraf, and J. Cohn. Investigating spontaneous facial action recognition through AAM representations of the face. In K. Delac and M. Grgic, editors, *Handbook of Face Recognition*, pages 275–286. I-Tech Education and Publishing, 2007.

[109] L.Z. McArthur and R.M. Baron. Toward an ecological theory of social perception. *Psychological Review*, 90(3):215–238, 1983.

[110] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. Modeling human interaction in meetings. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 748–751, 2003.

[111] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):305–317, 2005.

[112] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought.* University Of Chicago Press, 1996.

[113] A. Mehrabian and S.R. Ferris. Inference of attitude from nonverbal communication in two channels. *Journal of Counseling Psychology*, 31(3):248–252, 1967.

[114] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *Proceedings of the European Conference on Computer Vision*, pages 69–81, 2004.

[115] T.B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.

[116] T.B Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126, 2006.

[117] Y. Moh, P. Nguyen, and J.C. Junqua. Towards domain independent speaker clustering. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 85–88, 2003.

[118] S. Möller and R. Schönweiler. Analysis of infant cries for the early detection of hearing impairment. *Speech Communication*, 28(3):175–193, 1999.

[119] P.R. Montague, G.S. Berns, J.D. Cohen, S.M. McClure, G. Pagnoni, M. Dhamala, M.C. Wiest, I. Karpov, R.D. King, N. Apple, and R.E. Fisher. Hyperscanning: Simultaneous fMRI during linked social interactions. *Neuroimage*, 16(4):1159–1164, 2002.

[120] B.C.J. Moore. *An introduction to the psychology of hearing.* Academic Press, 1982.

[121] N. Morgan, E. Fosler, and N. Mirghafori. Speech recognition using on-line estimation of speaking rate. In *Proceedings of Eurospeech*, pages 2079–2082, 1997.

[122] N. Morgan and E. Fosler-Lussier. Combining multiple estimators of speaking rate. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 729–732, 1998.

[123] D. Morris. *Peoplewatching.* Vintage, 2007.

[124] S. Mota and R.W. Picard. Automated posture analysis for detecting learners interest level. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 49–56, 2003.

[125] S. Mukhopadhyay and B. Smith. Passive capture and structuring of lectures. *Proceedings of the ACM International Conference on Multimedia*, pages 477–487, 1999.

[126] K.P. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning.* PhD thesis, University of California Berkeley, 2002.

[127] C. Nass and K.M. Lee. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, 7(3):171–181, 2001.

[128] C. Nass and J. Steuer. Computers and social actors. *Human Communication Research*, 19(4):504–527, 1993.

[129] I. Oikonomopoulos, I. Patras, and M. Pantic. Spatiotemporal salient points for visual recognition of human actions. *IEEE Transactions On Systems, Man, and Cybernetics - Part B*, 36(3):710–719, 2006.

[130] N. Oliver, A. Garg, and E. Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *Computer Vision and Image Understanding*, 96(2):163–180, 2004.

[131] A. J. O'Toole, T. Price, T. Vetter, J.C. Bartlett, and V. Blanz. 3D shape and 2D surface textures of human faces: the role of averages in attractiveness and age. *Image and Vision Computing*, 18(1):9–19, 1999.

[132] S. Oviatt. User-centered modeling and evaluation of multimodal interfaces. *Proceedings of the IEEE*, 91:1457–1468, 2003.

[133] S. Oviatt, C. Darves, and R. Coulston. Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. *ACM Transactions on Computer-Human Interaction*, 11(3):300–328, 2004.

[134] P. Pal, A.N. Iyer, and R.E. Yantorno. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 721–724, 2006.

[135] M. Pantic and M.S. Bartlett. Machine analysis of facial expressions. In K. Delac and M. Grgic, editors, *Handbook of Face Recognition*, pages 377–416. I-Tech Education and Publishing, 2007.

[136] M. Pantic and I. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, 36(2):433–449, 2006.

[137] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. Human computing and machine understanding of human behavior: A survey. In *Lecture Notes in Artificial Intelligence*, volume 4451, pages 47–71. Springer Verlag, 2007.

[138] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. Human-centred intelligent human-computer interaction (HCI^2): How far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems*, 1(2):168–187, 2008.

[139] M. Pantic and L.J.M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[140] M. Pantic and L.J.M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[141] C. Pelachaud, V. Carofiglio, B. De Carolis, F. de Rosis, and I. Poggi. Embodied contextual agent in information delivering application. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 758–765, 2002.

[142] A. Pentland. Social dynamics: Signals and behavior. In *International Conference on Developmental Learning*, 2004.

[143] A. Pentland. Socially aware computation and communication. *IEEE Computer*, 38(3):33–40, 2005.

[144] A. Pentland. Automatic mapping and modeling of human networks. *Physica A*, 378:59–67, 2007.

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

[146] A. Pentland. *Honest signals: how they shape our world.* MIT Press, 2008.

[147] S. Petridis and M. Pantic. Audiovisual discrimination between laughter and speech. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 5117–5121, 2008.

[148] S. Petridis and M. Pantic. Audiovisual laughter detection based on temporal features. In *Proceedings of IEEE International Conference on Multimodal Interfaces*, pages 37–44, 2008.

[149] T. Pfau and G. Ruske. Estimating the speaking rate by vowel detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 945–948, 1998.

[150] J.O. Pickles. *An introduction to the physiology of hearing.* Academic Press, 1982.

[151] F.E. Pollick, V. Lestou, J. Ryu, and S.B. Cho. Estimating the efficiency of recognizing gender and affect from biological motion. *Vision Research*, 42:2345–2355, 2002.

[152] F.E. Pollick, H.M. Paterson, A. Bruderlin, and A.J. Sanford. Perceiving affect from arm movement. *Cognition*, 82(2):51–61, 2001.

[153] R. Poppe. Vision-based human motion analysis: an overview. *Computer Vision and Image Understanding*, 108:4–18, 2007.

[154] G. Psathas. *Conversation Analysis - The study of talk-in-interaction.* Sage Publications, 1995.

[155] L. Rabiner and M. Sambur. Voiced-unvoiced-silence detection using the Itakura LPC distance measure. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 323–326, 1977.

[156] L.R. Rabiner and R.W. Schafer. *Digital processing of speech signals.* Prentice-Hall Englewood Cliffs, NJ, 1978.

[157] D.A. Reynolds, W. Campbell, T.T. Gleason, C. Quillen, D. Sturim, P. Torres-Carrasquillo, and A. Adami. The 2004 MIT Lincoln laboratory speaker recognition system. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 177–180, 2005.

[158] V.P. Richmond and J.C. McCroskey. *Nonverbal Behaviors in interpersonal relations.* Allyn and Bacon, 1995.

[159] R. Rienks and D. Heylen. Dominance Detection in Meetings Using Easily Obtainable Features. In *Lecture Notes in Computer Science*, volume 3869, pages 76–86. Springer, 2006.

[160] R. Rienks, D. Zhang, and D. Gatica-Perez. Detection and application of influence rankings in small group meetings. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 257–264, 2006.

[161] H.A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[162] J.A. Russell, J.A. Bachorowski, and J.M. Fernandez-Dols. Facial and vocal expressions of emotion. *Annual Reviews in Psychology*, 54(1):329–349, 2003.

[163] J.A. Russell and J.M. Fernandez-Dols, editors. *he Psychology of Facial Expression.* Cambridge University Press, 1997.

[164] N. Russo. Connotation of seating arrangements. *Cornell Journal of Social Relations*, 2(1):37–44, 1967.

[165] M.A. Sayette, D.W. Smith, M.J. Breiner, and G.T. Wilson. The effect of alcohol on emotional response to a social stressor. *Journal of Studies on Alcohol*, 53(6):541–545, 1992.

[166] A.E. Scheflen. The significance of posture in communication systems. *Psychiatry*, 27:316–331, 1964.

[167] K.R. Scherer. *Personality markers in speech.* Cambridge University Press, 1979.

[168] K.R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227–256, 2003.

[169] H. Schneiderman and T. Kanade. A statistical model for 3D object detection applied to faces and cars. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 746–751, 2000.

[170] B. Schuller, R. Müeller, B. Höernler, A. Höethker, H. Konosu, and G. Rigoll. Audiovisual recognition of spontaneous interest within conversations. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 30–37, 2007.

[171] U. Segerstrale and P. Molnar, editors. *Nonverbal communication: where nature meets culture.* Lawrence Erlbaum Associates, 1997.

[172] A. Sepheri, Y. Yacoob, and L. Davis. Employing the hand as an interface device. *Journal of Multimedia*, 1(7):18–29, 2006.

[173] E. Shriberg. Phonetic consequences of speech disfluency. *Proceedings of the International Congress of Phonetic Sciences*, 1:619–622, 1999.

[174] E. Shriberg, R. Bates, and A. Stolcke. A prosody-only decision-tree model for disfluency detection. In *Proceedings of Eurospeech*, pages 2383–2386, 1997.

[175] E. Shriberg, A. Stolcke, and D. Baron. Observations of overlap: findings and implications for automatic processing of multiparty conversation. In *Proceedings of Eurospeech*, pages 1359–1362, 2001.

[176] P.E. Shrout and D.W. Fiske. Nonverbal behaviors and social evaluation. *Journal of Personality*, 49(2):115–128, 1981.

[177] K. Sjölander and J. Beskow. Wavesurfer-an open source speech tool. In *Proceedings of International Conference on Spoken Language Processing*, pages 464–467, 2000.

[178] A.F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, 2006.

[179] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding*, 104(2-3):210–220, 2006.

[180] L. Smith-Lovin and C. Brody. Interruptions in group discussions: the effects of gender and group composition. *American Sociological Review*, 54(3):424–435, 1989.

[181] K.K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.

[182] E. L. Thorndike. Intelligence and its use. *Harper's Magazine*, 140:227–235, 1920.

[183] C. Thurau. Behavior histograms for action recognition and human detection. In *Lecture Notes in Computer Science*, volume 4814, pages 271–284. Springer Verlag, 2007.

[184] C. Thurau and V. Hlavac. Pose primitive based human action recognition in videos or still images. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2008.

[185] Y. Tian, T. Kanade, and J.F. Cohn. Facial expression analysis. In S.Z. Li and A.K. Jain, editors, *Handbook of Face Recognition*, pages 247–276. 2005.

[186] H.L. Tischler. *Introduction to Sociology.* Harcourt Brace College Publishers, 1990.

[187] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1683–1699, 2007.

[188] D. Tran, A. Sorokin, and D.A. Forsyth. Human activity recognition with metric learning. In *Proceedings of the European Conference on Computer Vision*, 2008.

[189] S.E. Tranter and D.A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565, 2006.

[190] H.C. Triandis. *Culture and social behavior.* McGraw-Hill, 1994.

[191] K.P. Truong and D.A. Leeuwen. Automatic detection of laughter. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 485–488, 2005.

[192] K.P. Truong and D.A. van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158, 2007.

[193] L.Q. Uddin, M. Iacoboni, C. Lange, and J.P. Keenan. The self and social cognition: the role of cortical midline structures and mirror neurons. *Trends in Cognitive Sciences*, 11(4):153–157, 2007.

[194] A. Utsumi and N. Tetsutani. Human detection using geometrical pixel value structures. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 34–39, 2002.

[195] M.F. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 38–45, 2007.

[196] M.F. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition Workshop*, pages 149–150, 2006.

[197] M.F. Valstar, M. Pantic, Z. Ambadar, and J.F. Cohn. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 162–170, 2006.

[198] J. Van den Stock, R. Righart, and B. de Gelder. Body expressions influence recognition of emotions in the face and voice. *Emotion*, 7(3):487–494, 2007.

[199] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 301–308, 2001.

[200] A. Vinciarelli. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Transactions on Multimedia*, 9(9):1215–1226, 2007.

[201] A. Vinciarelli and J.-M. Odobez. Application of information retrieval technologies to presentation slides. *IEEE Transactions on Multimedia*, 8(5):981–995, 2006.

[202] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. 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, 2008.

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

[204] P. Viola and M. Jones. Robust real-time face detection. *Computer Vision*, 57(2):137–154, 2004.

[205] A. Waibel, T. Schultz, M. Bett, M. Denecke, R. Malkin, I. Rogina, and R. Stiefelhagen. SMaRT: the Smart Meeting Room task at ISL. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 752–755, 2003.

[206] L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36(3):585–601, 2003.

[207] P. Wang and Q. Ji. Multi-view face detection under complex scene based on combined SVMs. In *Proceedings of International Conference on Pattern Recognition*, pages 179–182, 2004.

[208] R.M. Warner and D.B. Sugarman. Attributions of personality based on physical appearance, speech, and handwriting. *Journal of Personality and Social Psychology*, 50(4):792–799, 1986.

[209] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[210] C.Y. Weng, W.T. Chu, and J.L. Wu. Movie analysis based on roles social network. In *Proceedings of IEEE International Conference on Multimedia and Expo*, pages 1403–1406, 2007.

[211] J. Whitehill and J.R. Movellan. Personalized facial attractiveness prediction. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[212] A.C.C. Williams. Facial expression of pain: An evolutionary account. *Behavioral and Brain Sciences*, 25(4):439–455, 2003.

[213] Y. Wu and T.S. Huang. Vision-based gesture recognition: A review. In *Proceedings of the International Gesture Workshop*, pages 103–109, 1999.

[214] Y. Yacoob and L. Davis. Detection and analysis of hair. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7):1164–1169, 2006.

[215] M.H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[216] J. Yao and J.-M. Odobez. Fast human detection from videos using covariance features. In *Proceedings of European Conference on Computer Vision Visual Surveillance Workshop*, 2008.

[217] G. Yule. *Pragmatics*. Oxford University Press, 1996.

[218] M. Zancanaro, B. Lepri, and F. Pianesi. Automatic detection of group functional roles in face to face interactions. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 28–34, 2006.

[219] B. Zellner. Pauses and the temporal structure of speech. In E. Keller, editor, *Fundamentals of speech synthesis and speech recognition*, pages 41–62. John Wiley & Sons, 1994.

[220] Z. Zeng, Y. Fu, G.I. Roisman, Z. Wen, Y. Hu, and T.S. Huang. Spontaneous emotional facial expression detection. *Journal of Multimedia*, 1(5):1–8, 2006.

[221] Z. Zeng, M. Pantic, G.I. Roisman, and T.H. Huang. A survey of affect recognition methods: audio, visual and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[222] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):699–714, 2005.

[223] Q. Zhu, S. Avidan, M.C. Yeh, and K.T. Cheng. Fast human detection using a cascade of histograms of oriented gradients. In *Proceedings of Computer Vision and Pattern Recognition*, pages 1491–1498, 2006.