Social Signals, their Function, and Automatic Analysis: A Survey

Alessandro Vinciarelli1,2, Maja Pantic3,4, Hervé Bourlard1,2 and Alex Pentland5
1Idiap research institute, CP592 - 1920 Martigny (CH)
2Ecole Polytechnique Federale de Lausanne - 1015 Lausanne (CH)
3Computing, Imperial College London, 180 Queen’s Gate - London SW7 2AZ (U.K.)
4EEMCS, University of Twente, Drienerlolaan 5 - 7522 NB Enschede (NL)
5The MIT Media Laboratory, 20 Ames St. - Cambridge, MA 01239 (USA)
vincia@idiap.ch, m.pantic@imperial.ac.uk, bourlard@idiap.ch, pentland@media.mit.edu

ABSTRACT
Social Signal Processing (SSP) aims at the analysis of social behaviour in both Human-Human and Human-Computer interactions. SSP revolves around automatic sensing and interpretation of social signals, complex aggregates of nonverbal behaviours through which individuals express their attitudes towards other human (and virtual) participants in the current social context. As such, SSP integrates both engineering (speech analysis, computer vision, etc.) and human sciences (social psychology, anthropology, etc.) as it requires multimodal and multidisciplinary approaches. As of today, SSP is still in its early infancy, but the domain is quickly developing, and a growing number of works is appearing in the literature. This paper provides an introduction to nonverbal behaviour involved in social signals and a survey of the main results obtained so far in SSP. It also outlines possibilities and challenges that SSP is expected to face in the next years if it is to reach its full maturity.

Categories and Subject Descriptors: H.3.1 [Content Analysis and Indexing].

General Terms: Experimentation.

Keywords: Social Signal Processing, Social Behaviour Analysis, Computer Vision, Speech Analysis.

1. INTRODUCTION
Social Signal Processing (SSP) is the new pioneering domain aimed at bringing social intelligence to computers [53]. This is one of the multiple facets of human intelligence and can be thought of as the ability of dealing effectively with social interactions, whether this means to be accepted as leader in a working environment, to be an understanding parent, to be respected in a community, or to capture the attention of the audience. Since humans spend most of their life being involved in social interactions, social intelligence is definitely a key ability that can make the difference between success and failure in life [1].

How can a computer, or a machine in general, develop social intelligence when this seems to pertain to aspects of human psychology that are far from being detectable and accessible by sensors? The answer is in nonverbal communication, the phenomenon that psychologists have been studying for more than one century, and that consists of the wide spectrum of nonverbal behaviours through which humans communicate what cannot be said with words including feelings and attitudes [38][56]. Nonverbal communication can be considered as one of the physical, detectable, and measurable evidences of our inner life, the other being the content of our verbal messages. But unlike the latter, nonverbal communication is typically honest [19][54] and reliable because it is mostly out of the reach of conscious control, thus it leaks information about our actual state, and not about what we want to show as such.

In the case of social interactions, nonverbal communication takes the form of social signals [2][3], complex aggregates of behavioural cues accounting for our attitudes towards other human (and virtual) participants in the current social context. Social signals include phenomena such as attention, empathy, politeness, flirting, and (dis)agreement, and are conveyed through multiple behavioural cues including posture, facial expression, voice quality, gestures, etc. Social signals are what we experience when we watch a television program in a language we do not understand and still we are able to capture most of the social landscape, i.e., what kind of relationships people have (are there hierarchic relationships, Are people happy with one another, and similar).

As an example, consider Figure 1. The information at disposition is limited to two silhouettes, but still people observing the picture assess correctly in 50% of cases the interaction that takes place: the man and the woman form a couple and they are fighting. In the other 50% of the cases, assessors guess correctly at least some elements of the actual situation; they understand that the two persons are members of the same family or close friends (people do not allow mere acquaintances to come so close), and they clearly see that the interaction is intense (the backward posture and the tension of the hand leave few doubts about).

When it comes to automatic sensing and processing of so-
Figure 1: Behavioural cues and social signals. Combinations of multiple behavioural cues (vocal behaviour, posture, mutual gaze, interpersonal distance, etc.) produce social signals (in this case aggressivity or disagreement) that are evident even from static images showing only the silhouettes of the individuals involved in the interaction.

Social signals, the problem is tractable because social signals are based on behavioural cues that are detectable through sensors as simple as cameras and microphones. However, 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 (for an overview of the relevant technologies see [48][49]), the research efforts in machine analysis and synthesis of human social signals are still tentative and pioneering efforts. Yet, the importance of studying social interactions and developing automated assessing of human social behaviour from audiovisual recordings is indisputable. It will result in valuable multimodal tools that could revolutionise basic research in cognitive and social sciences by raising the quality and shortening the time to conduct research that is now lengthy, laborious, and often imprecise. At the same time, such tools form a large step ahead in realising naturalistic, socially-aware computing and interfaces, built for humans, based on models of human behaviour.

This paper provides a survey of the main approaches proposed so far for the analysis social behaviour. Section 2 describes in detail nonverbal communication and its role in social behaviour, Section 3 shows the main results obtained so far in SSP, and Section 4 outlines the main challenges that the researchers in the domain face.

2. NONVERBAL BEHAVIOUR IN SOCIAL INTERACTIONS

While interacting with other people, our attention tends to focus only on verbal messages. Yet, social interactions are rich in nonverbal aspects which influence heavily not only the meaning of the words, but also our perception of social contexts [33]. The correct interpretation of multiple nonverbal behavioural cues we exchange with others, consciously and unconsciously, is the key ability that makes the difference between dealing appropriately and inappropriately with social interactions [1]. Figure 2 shows the three main components of nonverbal communication: nonverbal behavioural cues, i.e., observable changes in facial and body gesture that accompany our communication, codes, i.e., classes of nonverbal behavioural cues related to specific communication means/ modality, and functions, i.e., the goals that nonverbal communication is aimed at.

Behavioural cues accompany any social interaction and, more in general, any moment of life. In this sense, humans cannot “not communicate”. Even when sleeping, humans communicate through their movement or their position [56]. When talking on the phone, people still use gestures to punctuate the discourse even though the listener cannot see them [38]. When they are alone, people still display their emotions. Nonverbal behaviour is one of the most pervasive aspects of human life [31].

2.1 Codes and Behavioural Cues

Codes are classes of nonverbal behaviours sharing a common function or, more often, represent a common communication mean/ modality (e.g., the voice or the face). Five codes (and the related behavioural cues) can be identified:

Physical appearance.
This code includes bodily characteristics such as height, body shape, physiognomy, skin and hair color, as well as artificial characteristics such as clothes, ornaments, make up, and other manufats used to modify/ accentuate the facial/ body aspects.

Gestures and postures.
Gestures are often used consciously, e.g., when waiving a hand in a greeting gesture. However, postures are typically assumed unconsciously and they are one of the most reliable cues about the rapport between people [56].

Face and eyes behaviour.
Facial expressions and gaze behaviour (who looks at whom and how much) are arguably the most reliable cues when
interpreting social signals. Psychological experiments show that rapport judgments based on facial expression only are far more accurate than the judgments based on any other behavioural cue used alone [3].

Vocal behaviour.
All spoken cues that accompany a verbal message such as voice quality (how something is said), linguistic and non-linguistic vocalizations (the former are expressions like “ehm” and “ah-ah”, that are used as words even though they are not real words; the latter are laughter, cries, etc.), silences, and turn-taking patterns (who talks to whom and how smooth is the transition) influence its actual meaning [12].

Space and environment behaviour.
Physical distances between individuals often correspond to their social distances. Anthropologists have shown that people tend to split the space around them into concentric regions where others are allowed depending on social criteria [32]: only family members come closer than 0.5 meters, friends and people we meet frequently (e.g., colleagues) stay between 0.5 and 1.2 meters, formal relationships (e.g., with professionals or hierarchical superiors) take place between 1.2 and 2.0 meters. The figures apply to western cultures, but can change in other contexts.

2.2 The Functions
Combined together, different codes (classes of nonverbal behaviours) send powerful messages that influence specific aspects of social interplay. Table 1 shows that different behavioural cues affect some of the most important social signals, including personality, status, dominance, persuasion, interaction regulation, and rapport. Overall, psychologists have identified six main functions of nonverbal communication.

Forming impressions.
People make judgments about others even before starting an interaction. Behavioural cues like movement and appearance contribute significantly to the impression we form in others because they are noticed first. Overall, the first impression we have about others is completely dominated by nonverbal communication [3].

Although common wisdom suggests that the appearance is not important, psychological observations suggest the reverse. For example, attractiveness elicits desirable social perceptions like high status or good personality even though no objective basis for such an assumption exists (this phenomenon is referred to as “what is beautiful is good” [16]). Tall people are attributed, on average higher social status [26], and the body shapes (round and soft, bony and muscular, or thin and fragile) tend to elicit the attribution of certain personality traits [11].

Managing interaction.
Appropriate nonverbal messages regulate the interaction flow during conversations, showing when it is the right moment to intervene, when a turn is going to its end, etc. [55][73]. Frequent interactions between specific individuals typically lead to coordination and synchronicity of nonverbal messages, so that the resulting interaction is fluent and smooth. The regulation in conversations relies on behaviours (including voice quality and gaze) aimed at maintaining, yielding, denying, or requesting the turn [73]. When the interaction is satisfying, the speaker transitions tend to be smooth and no interruptions or long latency times are observed. When the interactions are not positive, interruptions and other behaviours related to aggressivity and dominance appear more frequently [67]. Note, however, that the amount of overlapping speech accounts for up to 10% of the total time even in normal conversations [64].
Expressing emotion.
Nonverbal communication is the primary mean for expressing emotions [20]. Affective arousal modulates all human communicative signals, but facial expressions and body gestures appear to be the most important in the human judgment of affective behaviour [3][20].

Sending relational messages.
Arguably, this is the most important function in social interactions. Nonverbal messages communicate how people feel about others, what kind of relationship they want to establish, whether they like or dislike others, etc.

In general, nonverbal messages communicate attitudes towards the others without the need of verbalizing them. This is evident in psychological experiments where human observers judge the rapport between people using a single behavioural cue. Usually, independently of the target judged interactions. Nonverbal messages communicate attitudes towards the others without the need of verbalizing them. This is evident in psychological experiments where human observers judge the rapport between people using a single behavioural cue. Usually, independently of the target judged behaviour, the results obtained using the facial expressions alone lead to the best accuracy [3]. Facial expressions, typically represented with the Facial Action Coding System (FACS) [21], express cognitive states like interest and puzzlement [13], social behaviours like accord and rapport [3][13], personality traits like extraversion and temperamen [23], 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) [20].

Relational messages are also conveyed through gestures and postures. Gestures like self-touching and manipulation of small objects, called adaptors, occur typically due to boredom or negative attitudes towards others [38]. Postures are typically assumed unconsciously and they are one of the most reliable cues about the rapport between people [56]. Three main criteria define the social meaning of a posture [60]: inclusion vs. exclusion (facing in the direction opposite to others shows a negative attitude), parallel vs. face-to-face (the choice of face-to-face postures in absence of constraints shows engagement in the interaction), and congruence vs. non-congruence (people having satisfying interactions tend to assume the same posture).

Relational messages are also conveyed by vocal behaviour, which communicates emotions [62], and influences the perception of dominance, extraversion, competence and persusiveness [61]. Linguistic vocalizations include all non-words that are used as if they were actual words, e.g., “ehm”, “ah-ah”, “uhm”, etc. They typically account for embarrassment or difficulty with respect to a social interaction [27], but they are also used when someone else speaks (the back-channel) to show attention, agreement, wonder or contradiction [65]. The non-linguistic vocalizations include nonverbal sounds like laughing, sobbing, crying, whispering, groaning, and similar. These may or may not accompany words, and can be used to reward desirable social behaviour (e.g., through laughter [36]), or to show strong social bonds (e.g., when crying in empathy).

Deceiving and detecting deception.
Nonverbal behaviour plays a major role in the deception process. Deception-related behavioural cues are typically split into strategic and non-strategic, the former are shown consciously to make lies more credible, the latter are out of control and typically leak negative feelings which can then be used to catch liars. Research in psychology actually suggests that the visual channel carrying facial expressions and body gestures is the most important in the human judgment of deceptive behaviour [3][19]. While facial expressions are often strategic behavioural cues (especially the easily activated smiles), body gestures are usually non-strategic cues as people consider them unimportant and do not pay an effort to control [19].

Sending Messages of Power and Persuasion.
Nonverbal behavioural cues are also signs of social control, power, and dominance. Powerful people touch more than they are touched, look at others less than they are looked at, and have control over time and space [4].

3. STATE-OF-THE-ART
The expression Social Signal Processing has been used for the first time in [53] to group under a collective definition several pioneering works of Alex Pentland and his group at MIT. These works aimed at two main applications: the prediction of behavioural outcomes like the result of salary negotiations, hiring interviews, or speed-dating conversations [14], and the analysis of large groups of individuals (around 100 people) through smart cellular phones equipped with proximity detectors and vocal activity analyzers [18][52] (an application called reality mining).

In approximately the same period, few other groups worked on the analysis of social interactions in multimedia recordings targeting three main areas: the analysis of interactions in small groups, the recognition of roles, and the sensing
Table 2: Results obtained by several Social Signal Processing works. For each work, information about the data (kind of interaction, availability, size), the total duration of the recordings, the distinction between real-world and simulated data, and the reported performance, are summarized.

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Data</th>
<th>Time</th>
<th>Source</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>[34]</td>
<td>Meetings from AMI Corpus (34 segments)</td>
<td>3h.00m</td>
<td>simulated</td>
<td>Most dominant person correctly detected in 85% of segments</td>
</tr>
<tr>
<td>[57]</td>
<td>Meetings (8 meetings)</td>
<td>1h.35m</td>
<td>simulated</td>
<td>Most dominant person correctly detected in 75% of meetings</td>
</tr>
<tr>
<td>[58]</td>
<td>Meetings (40 recordings)</td>
<td>20h.00m</td>
<td>simulated</td>
<td>Most dominant person correctly detected in 60% of meetings</td>
</tr>
</tbody>
</table>

**Collective Action Recognition**

| [15] | Meetings (30 recordings, publicly available) | 2h.30m | simulated | Action Error Rate of 12.5% |
| [41] | Meetings (60 recordings, publicly available) | 5h.00m | simulated | Action Error Rate of 8.9% |

**Role Recognition**

| [7] | Meetings (2 recordings, 3 roles) | 0h.45m | simulated | 50.0% of segments (up to 60 seconds long) correctly classified |
| [8] | NIST TREC SDR Corpus (35 recordings, publicly available 3 roles) | 17h.00m | real | 80.0% of the news stories correctly labeled in terms of role |
| [17] | The Survival Corpus (11 recordings, publicly available, 5 roles) | 4h.30m | simulated | 90% of precision in role assignment |
| [24] | AMI Meeting Corpus (138 recordings, publicly available, 4 roles) | 45h.00m | simulated | 67.9% of the data time correctly labeled in terms of role |
| [70] | Radio news bulletins (96 recordings, 6 roles) | 25h.00m | real | 80% of the data time correctly labeled in terms of role |
| [72] | Movies (3 recordings , 4 roles) | 5h.46m | real | 95% of roles correctly assigned |
| [74] | The Survival Corpus (11 recordings, publicly available, 5 roles) | 4h.30m | real | Up to 65% of analysis windows (around 10 seconds long) correctly classified in terms of role |

**Interest Level Detection**

| [25] | Meetings (50 recordings, 3 interest levels) | unknown | simulated | 75% Precision |
| [43] | Children playing with video games (10 recordings, 3 interest levels) | 3h.20m | real | 82% recognition rate |

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 [34][57][58], where multimodal approaches combine several nonverbal features, mainly speaking energy and body movement, to identify at each moment who is the dominant individual. The same kind of features has been applied in [15][41] to recognize the actions performed in meetings like discussions, presentations, etc. The combination of the information extracted from different modalities is performed with different algorithms including Dynamic Bayesian Networks [44] and layered Hidden Markov Models [46].

The recognition of roles has been addressed in two main contexts: broadcast material [8][70][72] and small scale meetings [7][17][74]. The works in [70][72] apply Social Network Analysis [71] to detect the role of people in broadcast news and movies, respectively. The approach in [8] recognizes the roles of speakers in broadcast news using vocal behaviour and lexical features. The roles in meetings are recognized using nonverbal behaviour in the case of [7], while a multimodal approach including both audio and visual features is applied in [17][74]. The interest level has been investigated in [25][43]. The first work applies features extracted from video and audio, while the second is mainly based on pressure sensors detecting the posture.

The reaction of users to social signals exhibited by computers has been investigated in several works showing that computers are social actors, i.e., they elicit the same reactions and perceptions as humans [45]. This happens, e.g., when children tend to imitate the voice quality of cartoon characters appearing on the interface of didactic applications [47], or when beta testers provide higher appreciation scores for interfaces exhibiting some form of mimicry, i.e. of the behaviour imitation typically displayed by humans to mean affiliation and liking [5].

4. CONCLUSIONS AND FUTURE CHALLENGES

Social Signal Processing has the ambitious goal of bringing social intelligence [1][28] to computers. The first results in this research domain have been sufficiently impressive to attract broader scientific [29] and business [9] communities. More importantly, the pioneering efforts in the field have established a viable interface between human sciences and
engineering, which is necessary if socially aware computing is to become our reality - social interactions and behaviours, although complex and rooted in yet to be explored aspects of human psychology, can be analyzed automatically with the help of computers. This "cultural" breakthrough is, in our opinion, the most important result of research in SSP so far. In fact, the pioneering contributions in SSP - [50][51]- have shown that the social signals, typically described as so elusive and subtle that only trained psychologists can recognize them [26], 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 [49]. 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) [10][39], the interpersonal adaptation (mutual accommodation of behavioural patterns between interacting individuals) [30], the interactional synchrony (degree of coordination during interactions) [37], the presence or roles in groups [6][68], the dynamics of conversations [55][73], etc. The collaboration between technology and social sciences demands a mutual effort of the two disciplines. On one hand, engineers must include social sciences in their reflection, while on the other hand, social scientists must formulate their findings in a form useful for engineers and their work on SSP.

The second issue relates to the need of implementing multicue, multi-modal, approaches to SSP. Nonverbal behaviours cannot be read like words in a book [38][56]; 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 [60]; physical distances typically account for social distances, but sometimes they are simply the effect of physical constraints [32]. Moreover, the same signal can correspond to different social behaviour interpretations depending on context and culture [69] (although many advocate that social signals are natural rather than cultural [63]). 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., [50]). This corresponds, from a technological point of view, to the combination of different classifiers that has extensively 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. It is therefore not surprising that some of the most successful works in SSP so far use features extracted from multiple modalities (e.g., [15][34][41]). 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 [22][59].

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., [14][41]), 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) [42]. Also other, less common kinds of data are centered on social interactions, e.g., meeting recordings [40], surveillance material [35], 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 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 [9]. 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 [66]).

Acknowledgments. 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 EU IST Programme project FP6-0027787 (AMIDA) and the EC’s 7th Framework Programme (FP7/2007-2013) under grant agreement no. 211486 (SEMAINE).

5. REFERENCES


