Recent Developments in Social Signal Processing

Albert Ali Salah*†, Maja Pantic‡§ and Alessandro Vinciarelli¶,
*Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey
Email: salah@boun.edu.tr
†Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
‡Department of Computing, Imperial College London, London, UK
Email: m.pantic@imperial.ac.uk
§EEMCS, University of Twente Enschede, The Netherlands
¶Department of Computing Science, University of Glasgow, Glasgow, UK
Email: vincia@dcs.gla.ac.uk

***Abstract*—Social signal processing has the ambitious goal of bridging the social intelligence gap between computers and humans. Nowadays, computers are not only the new interaction partners of humans, but also a privileged interaction medium for social exchange between humans. Consequently, enhancing machine abilities to interpret and reproduce social signals is a crucial requirement for improving computer-mediated communication and interaction. Furthermore, automated analysis of such signals creates a host of new applications and improvements to existing applications. The study of social signals benefits a wide range of domains, including human-computer interaction, interaction design, entertainment technology, ambient intelligence, healthcare, and psychology. This paper briefly introduces the field and surveys its latest developments.**

***Index Terms*—Behavioral science, human computer interaction, emotion recognition**

I. Introduction

When science fiction writers imagined a future for smarter computer systems, they often envisioned all-knowing, immensely capable, and purposeful machines. In February 2011, IBM's Watson computer won the *Jeopardy!* contest, displaying how close machines are to being 'all-knowing', thanks to their ability to store and sift through enormous amounts of factual information. While we may philosophically argue that 'knowing' is not applicable here in the strict sense, we cannot deny that Watson surpasses humans in its knowledge of trivia. However, when it comes to social intelligence, computers have a lot of room for improvement. We all use computers, and we know that they are never responsive to our emotions, moods, or to any kind of social context. Even an 'all-knowing' Watson computer would be a very poor dinner companion, as it lacks the skills to decide when to speak, and what to say. The newly emerging field of social signal processing aims at providing computers with the means of analysing and adequately representing human social signals, which will allow them to adapt and properly function in social settings.

Of course, the ultimate aim is not to create a digital dinner companion, although for the lonesome scientist that may be a worthy goal in itself. Social signal processing promises to benefit a host of domains and makes many applications possible. For ambient intelligence, it means environments that are more responsive to the social context [1]. For psychologists, it promises quantitative evaluation tools that can be used in coaching or diagnosis. For entertainment technology, more engaging games can be envisioned. And most importantly, new human-computer interaction challenges can be met by greatly increasing the sensitivity of the computer (or of the robot) to the interacting person's emotional and mental state [2], [3].

Computer-mediated communication is now commonplace for most of us, and while several methods are developed to transmit social signals over its usually simplified channels (like emoticons), the richness of face-to-face communication and of social signals transmitted during real conversations is largely lacking. Even when the medium offers conditions similar to traditional communication media (for instance a phone interface for a travel agency), an automated reply system is at a disadvantage, and may be perceived as cold and uneffective, instead of efficient and capable.

Bridging the social intelligence gap has two main aspects. Consider for a moment the automatic response system. To appear as human-like as possible, it should be able to analyse the incoming signals for their semantic and affective content, but it should also be able to produce appropriate affective responses in turn. The analysis and synthesis aspects allowed social signal processing researchers to create application settings with rigorous evaluation criteria right from the start, and ecological validity concerns introduced additional computational constraints to the already formidable challenge. Many issues are still open in the field, including the inherent uncertainty of machine detectable evidences of human behavior, multiscale temporal dynamics, and the appropriate psychological and cognitive theories that can provide useful concepts and models.

In this paper, we briefly introduce this domain, its main taxonomies and challenges. We refer the reader to previous surveys for an in-depth overview of the domain, as well as for historical insight into its development [4], [5]. We will focus here on the more recent developments, and present a snapshot of the field as it stands now.

II. Definitions and Taxonomies

Poggi and D'Errico define a *social signal* as a communicative or informative signal that, either directly or indirectly, conveys information about social actions, social interactions,

social emotions, social attitudes and social relationships [6].

In [5], a taxonomy is introduced for the analysis of social signals. Verbal signals that are usually direct manifestations of communicative intent are accompanied by *nonverbal behavioral cues* that serve to convey information about emotion, personality, status, dominance, regulation, rapport, etc. in a given social context. These cues reside in different modalities (or *codes*), namely physical appearance, gestures and postures, face and gaze behavior, vocal behavior, and use of space and environment.

Nonverbal behavioral cues are at the core of social signal processing (SSP). They typically describe temporal muscular and physiological changes that occur over short time intervals. Some cues last for milliseconds and are therefore difficult to perceive, but still play a role: An example is the movement of orbicularis oculi muscle on the face, which can be used to distinguish real smiles from posed ones [7]. These cues are perceived by humans during communication (with remarkable accuracy) consciously or unconsciously, and can radically alter the interpretation of the situation: A slight muscle movement or an inflection in the voice may add sarcasm to an otherwise innocent comment.

Paul Ekman and Wallace Friesen, building on their research in the 60s, as well as on earlier work by Efron [8], have classified nonverbal cues according to their origin, usage, and coding [9]. The type of messages conveyed by such cues are:

- *Emblems*: Signs that have direct verbal translation and used consciously, like the gesture of cutting one's throat. They are employed as substitutes for verbal signals, or to emphasize them.
- *Illustrators*: Signs that emphasize speech by providing visualization of spatial and temporal aspects, like illustration of a timeline, pointing, or a sketch of an object drawn in the air.
- *Affect Displays*: Signs that display more intimate and personal states, like emotions shown on the face.
- *Regulators*: Signs that coordinate the timing of other signals during communication. Turn-taking cues and backchannel signals are of this type.
- *Adaptors*: Signs that originate from habits, either in a self-manipulative fashion (like wiping the lips with the tongue) or through manipulation of objects (like twirling a pen).

This categorization does not properly do justice to the rich vocal nonverbal behaviors relevant as social signals, like paralinguistic information and voice quality, although some nonlinguistic vocalizations, silences, and turn-taking patterns would be considered regulators.

A few dimensions can arguably serve as relevant taxonomical distinctions in SSP. The temporal scale of signals is one such dimension, and it can span a range from milliseconds (as exemplified above with a muscle twitch) to minutes, hours, and much longer in case of behavioral habits. Another relevant distinction is individual vs. dyadic vs. group behaviour. Since we are evolutionary bound to produce social signals, we produce them even when we are alone, but dyadic interactions and multi-party interactions involve different phenomena, like cohesion [10] in groups and postural congruence in dyads [11]. A third distinction pertains to the depth of analysis. We can classify a face into smiling/non-smiling classes, but we can go deeper and classify whether a smile is real or posed, or whether it carries hints of sarcasm or pity.

We will review recent work in SSP from two aspects here. In Section III we will use a taxonomy in terms of signal channels, and report developments as per modality (i.e. faces and eyes, vocal behaviour, gestures, interaction geometry, and multimodal cues). There are social signals hidden in other modalities, for instance in appearance features (e.g. somatotypes, make-up and clothing), but we will not describe these here. In Section IV we will take up a complementary application perspective, and list developments in several application areas.

While we have define here the domain of SSP as actual physical behavior of humans (and of synthetic agents that primarily mimick them), human social signals are not restricted to the real world. Movement and interaction patterns of people traced via mobile phones, chatting and micro-blogging behaviour, connection formations over social networking platforms, and behaviors exhibited by avatars in virtual worlds all involve social signals, albeit of a different nature. The study of such virtual social signals goes by the name of *computational social science* [12].

III. Domains of Social Signals Analysis

A. Face and eyes

Faces convey information about gender, age, and emotions of a person, which are valuable sources in social signal processing. Of these, the most important area for SSP is facial expression analysis.

The state-of-the-art in facial expression analysis places emphasis on identifying Facial Actions (FACS), evaluation of expression in natural settings, as opposed to posed expressions, and a more detailed analysis of the temporal evolution of expressions as opposed to analysis from static images. Methods of processing facial affect are extensively reviewed in [13].

As the complexity of the classification problem grows, there is greater need for incorporating domain-specific knowledge into the learning system. In a recent work on facial action detection, Zhu *et al.* achieve this by a smarter training set selection for subsequent learning through a dynamic cascade bidirectional bootstrapping scheme and report some of the best results so far in AU detection on the RU-FACS database [14]. We mention some further developments in facial expression analysis in Section V.

Face analysis is also used for performing mutual gaze following and joint attention actions. Joint attention is the ability of coordination of a common point of reference with the communicating party. This skill is investigated in [15] for interaction with a robot, and in [16] for interaction with a virtual agent. Gaze direction is important in face-mediated affect, as direct gaze communicates threat or friendliness and plays a role in the expression of joy and anger, whereas averted

gaze facilitates avoidance-oriented expressions like fear and sadness [17].

One aspect of faces that received attention recently is the stereotypical judgments people base on face images. The appearance of a face can invoke (in an unjustified way) feelings of trust, warmth, confidence, etc. Alexander Todorov and his colleagues did an experiment in 2004, where they presented subjects with photographs of US general election candidates, for brief periods of time. From the ensuing competence judgments, they were able to predict election results with accuracy close to 70% [18]. A more extensive study on facial appearance and voting was reported in [19]. These findings demonstrate that facial appearance can act as a strong social signal, and even though the stereotypical judgments based on facial appearance do not have an objective basis (i.e. competent people do not necessarily look competent), they can be used to predict people's responses to them.

Automatic analysis of faces in natural settings depends on accurate face registration and facial feature tracking, and these are active research topics.

B. Vocal behaviour

Vocal behaviour concerns linguistic and paralinguistic information that may not be strictly related with the semantic meaning. Nonverbal vocal cues are valuable as they are 'honest' signals [20], [21], difficult to fake and revealing about socially relevant information. Low level speech features like spectral cues and prosody can be used to infer different signals from speech, such as motivation, empathy and dominance [22]. In [23], prosody is used to predict personality traits attributed to speakers. The authors used the well-known Big Five personality model to assess extraversion, agreeableness, conscientiousness, neuroticism and openness [24], and obtained $63-79\%$ recognition rate.

In [25] the authors propose a low-level speech feature that takes into account turn-taking behavior by modeling the length and transitions of speech and silence fragments in dyadic interactions. Probabilistic modeling based on this simple feature can be used to determine the roles assumed by interacting people in a meeting scenario. The work in [26] addresses the problem of discovering nonverbal patterns in a group via probabilistic topic models, and classify group dynamics and leadership within a group.

In [27], 36 statistical functionals are applied to low-level acoustic feature descriptor contours for turn-based emotion analysis. The authors use a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) approach for producing valence-arousal classification frame-by-frame. They also propose fusing the acoustic features with linguistic features (spotted keywords), which slightly improves results obtained with only acoustic features.

C. Gestures

Apart from face and voice cues, gestures in general, and upper-body movements in particular are analyzed for affective content. The approach in [28] stresses the use of head and hand movements for communicating emotions, especially in the context of recent human-centric computing paradigms. They seek parsimonious representations for describing a set of 12 emotion classes (elation, amusement, pride, pleasure, relief, interest, hot anger, fear, despair, cold anger, anxiety, sadness) grouped into high/low arousal and positive/negative valence clusters.

Newer sensors (like MS Kinect) offer ways of tracking the body in real-time via depth cues, which makes gesture and posture related approaches simpler. However, these sensors have restricted ranges, and often require some calibration. Automatic body part segmentation and tracking are still important problems, as offline content also needs to be analysed.

D. Interaction geometry and synchrony

The relative positioning of individuals and their postures during an interaction provides important social cues. These can include widely used social conventions, as well as heavily culture and context dependent situations, including proper manners and relevant social codes. In ordinary conversations, cues like interpersonal distance and posture can be used to derive conclusions about social and emotional relations between people. Recent approaches for the analysis of interaction geometry cues mostly use visual sensors. In [29], infrared cameras are used to track and classify the social situations in a multiperson interaction. It is mentioned that this kind of analysis can be of relevance in mobile service applications. A review of nonverbal signal analysis in small group interactions is given in [30].

Interaction synchrony refers to the temporal alignment of the interacting parties (especially for dyadic interactions) and involves simultaneous movement, similarity of interaction tempo, coordination and smoothness [31]. The amount of perceived synchrony is of importance for various applications like finding dominance in groups, mediated interactions with ECA or robots.

IV. Applications of Social Signal Analysis

Applications of SSP target role recognition, collective action recognition, interest level and dominance detection in interactions, among others [5]. We present a selection here.

A. Analysing interacting humans

A primary usage of SSP is in analysing and evaluating interacting humans for certain aspects. This can serve different purposes. For instance, autism and social anxiety disorders cause problems in emitting and interpreting social signals. Psychologists, when they analyse such individuals, record long sessions of interaction and manually annotate these. Automatic analysis tools can reduce the effort of annotation significantly.

The social interaction of humans is very rich, and recent research focuses on different aspects like dominance, pride, mirroring, agreement, etc. Each of these aspects can have many indicators. For instance in [32], audio-visual cues are used for detecting cases of agreement and disagreement during an interaction. Among agreement cues, the authors list head nods,

listener smiles, eyebrow raises, laughter, sideways leaning and mimicry, whereas disagreement can be signalled by head shakes, ironic smiles, sarcastic cheek creases, nose flares, leg clamps, and many more. Indeed, the authors list almost 40 different ways of signaling disagreement, through head gestures, facial actions, body posture and actions, auditory cues, hand actions, and gaze. For a thorough understanding of a real interaction, these signals need to be captured.

Dominance is one of the signals that received a lot of attention [33], [34], [35]. In [36] movement-based features from body and face, as well as mouth activity were analysed in a classification framework to determine dominance in dyadic conversations. In [34], a small meeting scenario is considered, where audio recordings are also available in addition to camera input. By using visual and audio cues, the authors demonstrate that audio cues are more successful than visual cues for establishing dominance, but the fusion of both improves over audio only.

B. Coaching

The skills of a good tutor incorporates social skills to motivate, challenge, bolster, and intrigue students [37]. One application of SSP is in creating computer systems for tutoring, which require implementing ways to emulate those skills. This means the student should be probed for signals of interest, boredom, curiosity, etc. [38].

Another application is the assessment of the coach (or teacher), as these motivational devices need to be used properly and in a timely manner [39].

Finally, the subject of coaching can be a social signal itself. The integration of real-time social signal processing into expert systems opens up new venues for this mature field. To give an example, Pfister and Robinson describe a classification scheme for real-time speech assessment, evaluated in the context of public speaking skills [40]. In this application nonverbal speech cues are extracted and used for assigning affective labels (absorbed, excited, interested, joyful, opposed, stressed, sure, thinking, unsure) to short speech segments, as well as for assessing the speech in terms of its perceived qualities (clear, competent, credible, dynamic, persuasive, pleasant), resulting in a novel and useful coaching scenario.

C. Social robotics

Building intelligent robots that can participate in conversations in a natural way, or to fulfill certain social roles in everyday environments is a great challenge. Among all computer systems, social robotics suffers most from a lack of social skills, as its aims are much more ambitious compared to other applications. These aims include childcare robots [41], healthcare robots, and service robots for domestic settings.

Apart from the capacity to analyse social signals, additional requirements for a socially responsive robot are primarily the ability to function in noisy environments, to process multimodal and temporal information in real time, and to produce correct signals at the correct time. This is also called "closing the affective loop," and it is an issue for robots and virtual agents alike [42].

We should mention here the Robocup@Home initiative[1], which is an attempt to make robots more sociable by integrating them into real domestic settings and by giving them simple tasks. The appearance and social behavior of the robot is judged by a jury in this challenge, in addition to sensory-motor skills.

D. Interaction with virtual agents

Embodied conversational agents (ECAs) require a number of capabilities that use non-verbal social signals for realistic interaction. These include initialization and termination of conversation sessions, turn-taking, and feedback functions [43]. The agent uses facial movements, head motions, and body movements to give these signals. Parametric models of body and limb motions are derived from actual interactions, and transferred to synthetic characters for realistic body and limb movements [44], [43], [45]. For instance, gesture rate and performance can be adjusted to tune the appeared extroversion of a virtual agent [46].

During an interaction with a virtual agent, nonverbal cues can be very dominant. In particular, the sensitive artificial listener (SAL) technique proposes that it may be possible to give adequate responses based on such nonverbal cues, even if what the other party is saying is not understood [47].

In [48] facial expressions are combined with movement cues obtained from the shoulder area, as well as with audio cues, to predict emotions in the valence-arousal space for an artificial listener, which monitors the interacting human for affective signals to give appropriate responses in real-time. This work also demonstrates that it is useful to learn correlations between valence and arousal.

Online worlds create novel social spaces where virtual avatars interact with each other. Game developers think of ways of enriching the expressiveness of the avatars, as well as natural ways of transmitting desired social signals from the controllers of the avatars to the actual virtual agent. Taken to the extreme, it is possible to induce affective responses in the real users through avatar interaction. An example application is presented in [49], where the users wear haptic interface devices to transmit social signals automatically through their avatars in Second Life. The HaptiHeart device conveys emotion-related heart rates (through speakers positioned on the body), HaptiButterfly creates a fluttering in the stomach via vibration motors, HaptiShiver sends shivers down the spine through a cold airflow, and HaptiTickler induces joy by actually tickling the user in the ribs.

V. Benchmarking and Databases

To assess models and systems of SSP, benchmarking efforts and annotated databases are of prime importance [5]. One of the early efforts in this area is the corpus collected by the Augmented Multi-party Interaction (AMI) project [50]. This database included recordings of a small number of people in a meeting scenario by multiple cameras and microphones, as

[1] http://www.ai.rug.nl/robocupathome/

well as annotations including FeelTrace ratings, posture and attented location information. Recently Aran *et al.* took subsets of this corpus and annotated them with dominance ratings [33]. Other group interaction corpora are listed in [30]. Similarly rich multimodal corpora were collected within the CHIL project for smart room interaction and presentation scenarios, and used in the CLEAR evaluation campaigns [51]. Other major databases with annotations include the SAL artificial listener database [47] and the HUMAINE database, which involved the definition and testing of several labeling schemes on a collage of different databases [52].

The expressions of signals differ in strength, and databases are annotated by observers (or raters), which may introduce a certain amount of subjectivity. This can be mitigated by requiring high inter-observer agreement in annotations, as in the extended Cohn Kanade (CK+) Database for facial expressions and facial action units [53]. This problem is approached from another angle in the Geneva multimodal emotion portrayals (GEMEP) corpus [54], which relies on actor portrayals to render 18 different emotions. Actors have more control over their facial muscles than ordinary people, as a result of the training they receive. Especially for subtle expressions, these portrayals can produce good inter-observer reliability. While such expressions present a simpler problem for computer analysis (clear acquisition conditions, exaggerated expressions), the database constitutes an important benchmark. On the other hand, the usefulness of exaggerated portrayals is sometimes criticised as not being ecologically valid [55].

A recent effort in benchmarking is the Facial Expression Recognition and Analysis Challenge, organized at the FG'2011 Conference [56]. The challenge used a subset of the GEMEP corpus for action unit classification. Many more databases that pertain to affective signals are reviewed in [13].

Finally, we mention two recent databases that are relevant to the present survey in that they point out to two major research fronts.

The first one is the Inter-ACT (INTEracting with Robots - Affect Context Task) corpus, which is an affective and contextually rich multimodal video corpus containing affective expressions of children playing chess with an iCat robot [41]. Since social and ambient robotics is gaining interest, we expect more human-robot interaction benchmark datasets to be available in the future, with particular attention to social aspects.

The second database pertains to action recognition, which is receiving a lot of interest from the multimedia community. Benchmarking efforts (like PASCAL VOC, Trecvid, ImageCLEF) have always been very important in multimedia retrieval. More and more, we see a shift towards the inclusion of semantically loaded and social concepts in such settings. The recent SDHA (Semantic Descriptions of Human Actions) Challenge[2], organized as a satellite event to ICPR'2010, provided three public databases for various action recognition settings [57]. While most actions were of simple nature (e.g. walking, sitting), some of the labelled actions have complicated semantic associations that denote social relations between the interacting parties (e.g., stalking, flirting), which makes these databases of relevance to the SSP community.

VI. Conclusive Remarks

The SSPNet[3] project defines the core questions of social signal processing as follows:

1) Is it possible to detect automatically nonverbal behavioral cues in data captured with sensors like microphones and cameras?
2) Is it possible to automatically infer attitudes from nonverbal behavioral cues detected through such sensors?
3) Is it possible to synthesize nonverbal behavioral cues conveying desired relational attitudes for embodiment of social behaviors in artificial agents, robots or other manufacts?

In attempting to answer these questions, SSP brings together computer science, engineering and social sciences together in a unique way. There are many challenges, especially in creating systems that work on real-world data, and in integrating the numerous findings back into useful applications, but there is also great progress.

Acknowledgment

The research that has led to this work has been supported by the European Communitys Seventh Framework Program (FP7/2007-2013), under grant agreement no. 231287 (SSPNet).

References

[1] E. Aarts and J. Encarnação, Eds., *True Visions: The Emergence of Ambient Intelligence*. Springer-Verlag, 2006.
[2] J. Crowley, J. Coutaz, and F. Bérard, "Perceptual user interfaces: things that see," *Communications of the ACM*, vol. 43, no. 3, pp. 54–64, 2000.
[3] C. Breazeal, "Emotion and sociable humanoid robots," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 119–155, 2003.
[4] A. Pentland, "Social Signal Processing," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 108–111, 2007.
[5] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social Signal Processing: Survey of an emerging domain," *Image and Vision Computing Journal*, vol. 27, no. 12, pp. 1743–1759, 2009.
[6] I. Poggi and F. D'Errico, "Cognitive modelling of human social signals," in *International Workshop on Social Signal Processing*, 2010, pp. 21–26.
[7] H. Dibeklioglu, R. Valenti, A. Salah, and T. Gevers, "Eyes do not lie: spontaneous versus posed smiles," in *ACM International Conference on Multimedia*, 2010, pp. 703–706.
[8] D. Efron, *Gesture and environment.* King's Crown Press, 1941.
[9] P. Ekman and W. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Semiotica*, vol. 1, no. 1, pp. 49–98, 1969.
[10] H. Hung and D. Gatica-Perez, "Estimating cohesion in small groups using audio-visual nonverbal behavior," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 563–575, 2010.
[11] T. Chartrand and J. Bargh, "The chameleon effect: The perception–behavior link and social interaction." *Journal of personality and social psychology*, vol. 76, no. 6, p. 893, 1999.
[12] D. Lazer, A. Pentland, L. Adamic, S. Aral, A. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne, "Computational social science," *Science*, vol. 323, pp. 721–723, 2009.

[2] http://cvrc.ece.utexas.edu/SDHA2010/index.html

[3] http://sspnet.eu/

[13] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[14] Y. Zhu, F. De la Torre, J. Cohn, and Y. Zhang, "Dynamic cascades with bidirectional bootstrapping for spontaneous facial action unit detection," in *IEEE Conference on Affective Computing and Intelligent Interaction*, vol. II, 2009, pp. 1–8.

[15] Z. Yucel and A. Salah, "Resolution of focus of attention using gaze direction estimation and saliency computation," in *IEEE Conference on Affective Computing and Intelligent Interaction*, vol. II, 2009.

[16] C. Peters, S. Asteriadis, and K. Karpouzis, "Investigating shared attention with a virtual agent using a gaze-based interface," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 119–130, 2010.

[17] R. Adams and R. Kleck, "Perceived gaze direction and the processing of facial displays of emotion," *Psychological Science*, vol. 14, no. 6, pp. 644–647, 2003.

[18] A. Todorov, A. Mandisodza, A. Goren, and C. Hall, "Inferences of competence from faces predict election outcomes," *Science*, vol. 308, no. 5728, p. 1623, 2005.

[19] J. Armstrong, K. Green, R. Jones, and M. Wright, "Predicting elections from politicians' faces," *International Journal of Public Opinion Research*, vol. 22, no. 4, p. 511, 2010.

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

[21] B. Lepri, K. Kalimeri, and F. Pianesi, "Honest signals and their contribution to the automatic analysis of personality traits–a comparative study," in *Human Behavior Understanding*. Springer, 2010, vol. LNCS 6219, pp. 140–150.

[22] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social Signal Processing: State-of-the-art and future perspectives of an emerging domain," in *ACM International Conference on Multimedia*, 2008, pp. 1061–1070.

[23] G.Mohammadi, M. Mortillaro, and A. Vinciarelli, "The voice of personality: mapping nonverbal vocal behavior into trait attributions," in *International Workshop on Social Signal Processing*, 2010, pp. 17–20.

[24] J. Wiggins, Ed., *The Five-Factor Model of Personality*. Guilford, 1996.

[25] M. Cristani, A. Pesarin, C. Drioli, A. Tavano, A. Perina, and V. Murino, "Generative modeling and classification of dialogs by a low-level turn-taking feature," *Pattern Recognition*, vol. 4, pp. 1785–1800, 2011.

[26] D. Jayagopi and D. Gatica-Perez, "Mining group nonverbal conversational patterns using probabilistic topic models," *IEEE Transactions on Multimedia*, vol. 12, no. 8, pp. 790–802, 2010.

[27] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 7–19, 2010.

[28] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer, "Towards a minimal representation of affective gestures," *IEEE Transactions on Affective Computing*, in press.

[29] G. Groh, A. Lehmann, J. Reimers, M. Friess, and L. Schwarz, "Detecting social situations from interaction geometry," in *IEEE International Conference on Social Computing*, 2010, pp. 1–8.

[30] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: a review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, 2009.

[31] 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*, vol. 30, no. 3, pp. 115–126, 2006.

[32] K. Bousmalis, M. Mehu, and M. Pantic, "Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools," in *International Conference on Affective Computing and Intelligent Interaction*, vol. II, 2009, pp. 121–129.

[33] O. Aran, H. Hung, and D. Gatica-Perez, "A multimodal corpus for studying dominance in small group conversations," in *LREC workshop on Multimodal Corpora*, 2010.

[34] O. Aran and D. Gatica-Perez, "Fusing Audio-Visual Nonverbal Cues to Detect Dominant People in Group Conversations," in *International Conference on Pattern Recognition*, 2010, pp. 3687–3690.

[35] I. Poggi and F. D'Errico, "Dominance signals in debates," in *Human Behavior Understanding*, vol. LNCS 6219, 2010, pp. 163–174.

[36] S. Escalera, O. Pujol, P. Radeva, J. Vitria, and M. Anguera, "Automatic detection of dominance and expected interest," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, 2010.

[37] M. Lepper, M. Woolverton, D. Mumme, and J. Gurtner, "Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors," *Computers as Cognitive Tools*, pp. 75–105, 1993.

[38] D. Heylen, A. Nijholt, and R. Akker, "Affect in tutoring dialogues," *Applied Artificial Intelligence*, vol. 19, no. 3, pp. 287–311, 2005.

[39] F. D'Errico, G. Leone, and I. Poggi, "Types of help in the teacher's multimodal behavior," in *Human Behavior Understanding*. Springer, 2010, vol. LNCS 6219, pp. 125–139.

[40] T. Pfister and P. Robinson, "Real-time recognition of affective states from non-verbal features and its application for public speaking skill analysis," *IEEE Transactions on Affective Computing*, in press.

[41] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. McOwan, "Inter-ACT: an affective and contextually rich multimodal video corpus for studying interaction with robots," in *ACM International Conference on Multimedia*. ACM, 2010, pp. 1031–1034.

[42] ——, "Affect recognition for interactive companions: challenges and design in real world scenarios," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 89–98, 2010.

[43] J. Cassell, "Embodied conversational interface agents," *Communications of the ACM*, vol. 43, no. 4, pp. 70–78, 2000.

[44] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone, "Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents," in *Annual Conference on Computer Graphics and Interactive Techniques*, 1994, pp. 413–420.

[45] D. Chi, M. Costa, L. Zhao, and N. Badler, "The EMOTE model for effort and shape," in *Annual Conference on Computer Graphics and Interactive Techniques*, 2000, pp. 173–182.

[46] M. Neff, Y. Wang, R. Abbott, and M. Walker, "Evaluating the effect of gesture and language on personality perception in conversational agents," in *Intelligent Virtual Agents*, 2010, pp. 222–235.

[47] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amir, and D. Heylen, "The sensitive artificial listener: an induction technique for generating emotionally coloured conversation," in *Workshop on Corpora for Research on Emotion and Affect*, 2008.

[48] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, in press.

[49] D. Tsetserukou, A. Neviarouskaya, H. Prendinger, M. Ishizuka, and S. Tachi, "iFeel_IM: innovative real-time communication system with rich emotional and haptic channels," in *International Conference on Human Factors in Computing Systems*, 2010, pp. 3031–3036.

[50] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The AMI meeting corpus: A pre-announcement," in *Workshop on Machine Learning in Multimodal Interaction*. Springer Verlag, 2005, pp. 28–39.

[51] A. Waibel and R. Stiefelhagen, Eds., *Computers in the human interaction loop*. Springer, 2009.

[52] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J. Martin, L. Devillers, S. Abrilian, A. Batliner *et al.*, "The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data," in *International Conference on Affective Computing and Intelligent Interaction*, 2007, pp. 488–500.

[53] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2010, pp. 94–101.

[54] T. Bänziger, H. Pirker, and K. Scherer, "GEMEP–GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions," in *Workshop on Corpora for Research on Emotion and Affect*, 2006.

[55] L. Vidrascu and L. Devillers, "Anger detection performances based on prosodic and acoustic cues in several corpora," in *Workshop on Corpora for Research on Emotion and Affect*, 2008, pp. 13–16.

[56] M. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *IEEE AFGR*, 2011.

[57] M. Ryoo, C. Chen, J. Aggarwal, and A. Roy-Chowdhury, "An overview of contest on semantic description of human activities (SDHA) 2010," in *Recognizing Patterns in Signals, Speech, Images and Videos*, D. Ünay, Z. Cataltepe, and S. Aksoy, Eds. Springer Verlag, 2010, pp. 270–285.