Image and Vision Computing 31 (2013) 203–221

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools ☆

Konstantinos Bousmalis [a,*], Marc Mehu [b], Maja Pantic [a,c]

[a] Department of Computing, Imperial College London, London, SW7 2AZ, UK
[b] Swiss Center for Affective Sciences, University of Geneva, Switzerland
[c] EEMCS, University of Twente, Enschede, The Netherlands

ARTICLE INFO

Article history:
Received 31 October 2011
Received in revised form 20 June 2012
Accepted 3 July 2012

Keywords:
Agreement
Disagreement
Nonverbal behaviour
Social signal processing

ABSTRACT

While detecting and interpreting temporal patterns of nonverbal behavioural cues in a given context is a natural and often unconscious process for humans, it remains a rather difficult task for computer systems. Nevertheless, it is an important one to achieve if the goal is to realise a naturalistic communication between humans and machines. Machines that are able to sense social attitudes like agreement and disagreement and respond to them in a meaningful way are likely to be welcomed by users due to the more natural, efficient and human-centred interaction they are bound to experience. This paper surveys the nonverbal behavioural cues that could be present during displays of agreement and disagreement; discusses a number of methods that could be used or adapted to detect these suggested cues; lists some publicly available databases these tools could be trained on for the analysis of spontaneous, audiovisual instances of agreement and disagreement, it examines the few existing attempts at agreement and disagreement classification, and finally discusses the challenges in automatically detecting agreement and disagreement.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Agreements and disagreements occur daily in human–human interaction, and are inevitable in a variety of everyday situations. These could be as simple as finding a location to dine and as complex as discussing notoriously controversial topics, like politics or religion. Agreement and disagreement are frequently expressed verbally, but the nonverbal behavioural cues that occur during these expressions play a crucial role in their interpretation [1]. This is naturally the case for agreement and disagreement as well as most facets of social attitudes like for instance politeness, flirting, or dominance [2].

A social attitude can be defined as the tendency of a person to behave in a certain way toward another person or a group of people. Social attitudes include cognitive elements like beliefs, evaluations, opinions, and social emotions [3]. Agreement and disagreement can be seen as social attitudes: if two people agree then this means that they have similar opinions, which usually entails an alliance, a commitment to cooperation, and a mutually positive attitude. In contrast, if two people disagree, this typically implies conflict, non-cooperation, and mutually negative attitude. Machine analysis of nonverbal behavioural cues (e.g., blinks, smiles, head nods, folded arms, etc.), has recently been the focus of intensive research, as surveyed by Pantic et al. in [4,5]. Similarly, significant advances have been made in the area of affect recognition (for exhaustive surveys, see [6,7]). However, research efforts on the machine analysis of social attitudes and social signals, signals that provide information about "social facts" are still at a rather early stage [4,2,3]. A few attempts have been made into, for example, recognising roles in multiparty meetings [8], identifying the political stance a participant holds in a debate [9], analysing social attitudes like interest [10–13], agreement and disagreement (see Section 6) among others.

Despite the commonly held knowledge that agreement and disagreement are expressed nonverbally with head nods and head shakes, respectively, there is little evidence that these attitudes are associated with specific behavioural cues. Like the expression of emotions [1] and most interpersonal attitudes [2], the communication of agreement and disagreement is likely to be of a multimodal nature. The issue is twofold: information about agreement and disagreement could be "encoded" in the different components of the multimodal signal and the perception of any of these components in isolation (words, facial action units, gestures, head movement, fundamental frequency, etc.) allows the retrieval of the meaning. This reasoning follows the principle of robustness [4], whereby the same information is believed to be encoded in separate components to increase efficiency of transmission if one of these components fails to operate appropriately [5].

Alternatively, the information about agreement and disagreement may be encoded in one component of the signal, the other components being devoted to other functions, for instance making the signal more efficient at influencing perceivers or transferring additional information.

☆ This paper has been recommended for acceptance by Hatice Gunes and Bjoern Schuller.

* Corresponding author.
E-mail addresses: k.bousmalis@imperial.ac.uk (K. Bousmalis), marc.mehu@unige.ch (M. Mehu), m.pantic@imperial.ac.uk (M. Pantic).

0262-8856/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.imavis.2012.07.003

This view implies that multimodal signals convey multiple messages [5]. To our knowledge, there is no data available that allows a direct test of these two explanations. However, recent theoretical developments suggest that multimodal signals are particularly efficient at solving the robustness problem while at the same time increasing information flow via an optimal set of correlations between the different components and communication channels [4]. The bottom-up approach advocated in this paper follows from the reasoning that strong correlations between agreement/disagreement and nonverbal cues could make it possible to detect these attitudes on the basis of associated clusters of audio-visual cues.

There is no overview available, to the best of our knowledge, of the nonverbal behavioural cues exhibited during agreement and disagreement, and any relevant literature in social psychology is at best scarce. This work[1] attempts to fill this gap and to be the first step towards our eventual objective: creating a system that can automatically detect these relevant behavioural cues, and detect agreement or disagreement based on both their morphology (i.e., presence and intensity) and temporal dynamics (i.e., timing, frequency and duration).

Note that we are interested only in those cues that can be detected using a monocular audiovisual data capturing system. The main reason for this choice is the fact that the average user has a monocular camera connected to their computer system and hence, any output from this research will be directly applicable in standard user applications, without the need for additional and expensive equipment (such as biosensors, thermal cameras, etc.). Furthermore, it will be possible to directly apply the research findings for automatically analysing and detecting agreement and disagreement in television data, e.g., for the summarization of televised political debates.

This paper attempts to organize a diverse, multi-disciplinary and complex literature that spans Social Psychology, Computer Vision, Machine Learning, and Social Signal Processing. The selection of papers discussed in this work often relied on established specialized surveys (e.g. pose estimation, social signal processing, human motion analysis). When that is the case, the reader is always referred to said survey for more information. Sections 2 and 3 involved a wide search for any available papers in psychology that referred to agreement and disagreement. Section 5 involved an extensive search for each cue in the Computer Vision and Machine Learning literatures.

In Section 2 we present the definitions of agreement and disagreement that we will be using, and discuss a typology of (dis)agreement expressions that should be used for the purpose of their automatic analysis, as well as the possibility of treating (dis)agreement in a dimensional approach. Social psychology literature regarding (dis) agreement only provides information about the morphology of behavioural cues in relation to agreement and disagreement, and has not yet presented concrete conclusions about the relevant importance and dynamics of such cues in relation to these social attitudes. We discuss these nonverbal behavioural cues that are relevant to detect agreement and disagreement in Section 3. This discussion serves as a starting point for the researcher who wants to utilise existing computation models, or to build new ones, towards the analysis of multicue dynamics in expressions of (dis)agreement, as well as the challenging task of the detection of such expressions. In Section 4 we present a list of databases rich in spontaneous (dis)agreement episodes, which could be used as a source of data for training such computational models. A number of tools that can be either adapted or used as-is to detect the (dis)agreement-relevant cues in such data are presented in Section 5. In Section 6, we discuss the progress made towards the automatic detection of (dis)agreement, which is so far limited to recognition of the attitudes in pre-segmented episodes. Finally, in Section 7 we discuss the challenges towards the automatic detection of (dis)agreement.

[1] This paper is an extension of [14], which was published in ACII'09. This paper is more complete in many ways: it presents a more thorough list of relevant cues; it discusses these cues in much more detail, and under the prism of culture; it presentsa lot more tools that can be used for the automatic detection of cues; the discussion on databases is a lot more detailed; there is a separate section for systems that have dealt with agreement and disagreement; and a separate section that deals with challenges.

2. Agreement and disagreement

Distinguishing between different kinds of agreement and disagreement is difficult, mainly because of the lack of a widely accepted definition of (dis)agreement [1]. However, the definition of (dis)agreement for the purpose of automatic detection needs to be a simple, yet concrete one. Poggi et al. [15] define agreement as the belief one holds that one is having the *same opinion* as one's interlocutor(s). We adopt this definition and similarly, we define disagreement as the belief one holds that one is having the *opposite opinion* as one's interlocutor(s). The communication of either belief via a speech act and/or nonverbal behavioural act is what we will assume in this work to be *an expression of agreement and disagreement* respectively. We emphasise, at this point, that we only consider agreement and disagreement that involve congruency or contradiction of *opinions* and not, for example, goals or emotions. Within this definition, it is important to keep in mind that agreement and disagreement can be an initial state in an interaction, or the product of a change of opinion due to this interaction, such as the case of being persuaded by another interlocutor.

With the task of automatically detecting expressions of agreement and disagreement in mind, we distinguish among three ways one could express these social attitudes with:

- *Direct Speaker's (Dis)Agreement*: A speaker uses specific words that convey direct (dis)agreement, e.g., "I (dis)agree with what you have just said".
- *Indirect Speaker's (Dis)Agreement*: A speaker does not explicitly state his or her (dis)agreement, but expresses an opinion that is congruent (agreement) or contradictory (disagreement) to an opinion that was expressed earlier in the conversation.
- *Nonverbal Listener's (Dis)Agreement*: A listener expresses nonverbally her (dis)agreement to an opinion that was just expressed. This could be via auditory cues like "mm hmm" or visual cues like a head nod or a smile [16]. (For a full list of the nonverbal cues that can be displayed during (dis)agreement, see Tables 1 and 2.) In this last case the expression is more ambiguous, since the meaning of nonverbal behavioural cues is not as specific as that of words, yet it is equally, if not more, important and should not be ignored. Ekman [17] talked about listeners' expressions of agreement and disagreement, mentioning that they are different from the speaker's expressions. Argyle [18] specifically discussed the fact that speakers attend to listeners for nonverbal signals that not only serve as feedback to the process of the conversation, but also as an expression of the listeners' opinion. Seiter et al. [19–21] have specifically discussed the importance of listeners' expressions of disagreement particularly in the context of televised political debates.

Table 1
Cues of agreement. For relevant descriptions of AUs, see FACS [36].

Cue	Kind	References
Head nod	Head gesture	[26,16,27,18,28–31]
Listener smile/lip corner pull (AU12, AU13)	Facial action	[18,32,28]
AU1 + AU2 + head nod	Facial action, head gesture	[17,33]
AU1 + AU2 + smile (AU12, AU13)	Facial action	[17,33]
AU1 + AU2 + agreement word	Facial action, verbal cue	[17,33]
Sideways leaning	Body posture	[27,34,18]
Laughter	Audiovisual cue	[1]
Mimicry	Second-order vocal and/or gestural cue	[18,27,35]

Table 2
Cues for disagreement. For relevant descriptions of AUs, see FACS [36].

Cue	Kind	References
Head shake	Head gesture	[29,18,27,19,17,20,28]
Head roll	Head gesture	[17,28]
Cut off	Head gesture	[30]
Lip bite (AU32) + head shake	Facial action, head gesture	[28]
Ironic smile/smirking [AU12 L/R (+AU14)]	Facial action	[17,37,18,19]
Eyebrow raise (AU1 + AU2) + … …[AU10 and/or AU15 and/or AU17 and/or AU43]	Facial action	[17]
"Mock astonishment" [AU1 + AU2 + (AU5 and/or AU26)]	Facial action	[17]
Barely noticeable lip-clenching (AU23, AU24)	Facial action	[30]
Cheek crease (AU14)	Facial action	[28]
Lowered eyebrow/frowning (AU4)	Facial action	[20,30,33,18,19]
Lip pucker (AU18)	Facial action	[30]
Slightly parted lips (AU25)	Facial action	[30]
Mouth movement (preparatory for speech) (AU25/AU26)	Facial action	[17]
Nose flare (AU38)	Facial action	[28]
Nose twist (AU9 L/R and/or AU10 L/R and/or AU11 L/R)	Facial action	[38,28]
Tongue show (AU19)	Facial action	[30]
Suddenly narrowed/slitted eyes (fast AU7)	Facial action	[30]
Eye roll	Facial action/gaze	[29,19–21]
Gaze aversion	Gaze	[16]
Clenched fist	Hand action	[28,30]
Forefinger raise	Hand action	[28]
Forefinger wag	Hand action	[28]
Hand chop	Hand action	[28]
Hand cross	Hand action	[28]
Hand wag	Hand action	[28,39,30]
Hands scissor	Hand action	[28]
Arm folding	Body posture	[34,28,30]
Large body shift	Body action	[30]
Leg clamp	Body posture	[28]
Head/chin support on hand	Body/head posture	[34,28,30]
Neck clamp	Hand/head action	[28]
Head scratch	Head/hand action	[28]
Self-manipulation	Hand/facial action	[30,28]
Feet pointing away	Feet posture	[30]
Sighing	Auditory cue	[21]
Throat clearing	Auditory cue	[30]
Delays: delayed turn initiation, pauses, filled pauses	Second-order auditory cue	[22,40,23,41,1,42]
Utterance length	Second-order auditory cue	[1,41]
Interruption	Second-order auditory cue	[42,9]

In addition to these expressions, disagreement may be viewed as a dispreferred activity, and a weak agreement could actually be a preface to an act of disagreement [22,23]. Moreover, agreement and disagreement could both be manifested at different levels, such as 'enthusiastic', 'reluctant', or 'unwilling' [24,15]. These should be kept in mind while trying to automatically detect (dis)agreement, and should be topics of further research by both social psychologists and computer scientists. A researcher of this topic, should try, at these early stages of research on automatic (dis)agreement analysis, to collect as homogenous data as possible, without sacrificing their spontaneity. As such factors make the problem of (dis)agreement analysis truly complex, in this work we will not consider different levels of (dis)agreement as a dispreferred activity.

Apart from discussing (dis)agreement in terms of different levels, i.e. a categorical approach of describing these social signals, (dis) agreement could be described in a dimensional approach [25]. In this approach, social signals are not independent from one another; rather, they are related in a systematic manner. Related work on describing emotions in a dimensional approach suggests that the majority of variability is covered by two dimensions: valence and arousal. The valence dimension refers to how positive or negative the emotion is, and ranges from unpleasant to pleasant feelings. The arousal dimension refers to how excited or apathetic the subject experiencing an affective state is, and ranges from boredom to excitement. It is possible to describe different levels of agreement and disagreement in terms of valence and arousal, however there is no such study yet that could provide such mappings. Another way of approaching this would be to treat (dis)agreement as one dimension of its own, as a continuous signal that ranges from strong disagreement to strong agreement. Section 7 outlines some of the challenges in the automatic detection of agreement and disagreement when it is approached in a dimensional and not a categorical way.

3. Cues of agreement and disagreement

We summarize, in this section, cues that could prove helpful in detecting (dis)agreement in natural encounters. Tables 1 and 2 list all cues that could be present during an agreement and a disagreement act based on the Social Psychology literature. It will become evident in this section, that as mentioned earlier, the combination and temporal dynamics of such cues will most likely be the key to identifying episodes of (dis)agreement, since most cues individually might have various, often opposite, interpretations. Rating studies based on data extracted from databases rich in spontaneous episodes of (dis) agreement, such as the ones presented in Section 4, are currently being conducted in order to establish the discriminative power of such combinations and their temporal characteristics, as well as their impact on human judgements.

One will also notice that our focus is primarily on *nonverbal* audiovisual cues, excluding, but not ignoring linguistic cues. There are a number of reasons for this choice. Most of the nonverbal cues we present are either universally performed or at least universally comprehended, without the necessity for common language. In many cases, as was suggested by Givens [30], there is an evolutionary and neurological explanation for associating the presented cues to (dis)agreement, such as the nod, the shake, the lip pucker, and throat clearing, among others. Analysis of lexical cues would probably help with (dis)agreement detection and a developer of such a system might want to include such cues, but these would largely depend on culture, language and even dialect, and would not prove particularly helpful with nonverbal listener's (dis)agreement, where we would expect to find richer nonverbal expression [16]. Cunningham et al. [24] also found that nonverbal cues, like rigid head movements were sufficient in human recognition of posed expressions of (dis) agreement, with non-rigid facial actions playing a lesser, but still significant role. Hence, at this stage, lexical cues are deemed out of scope for this work. However, the interested reader is encouraged to read the relevant work by Shriberg and colleagues, such as [41,40] (see also Section 6).

3.1. Backchannel cues

Ekman [17] specifically states that although emotional expressions during conversations are a reaction to the "affective content", they can also relate to the feelings regarding the nature and progress of the discussion. During a natural conversation, the participants, when in a listener's role, tend to continually give feedback as a means to facilitate floor-appointment, confirm their involvement, or even assess the quality of the conversation itself. The cues used for such feedback are called backchannel cues.

Brunner [32] specifies that there are three levels of meaning a feedback backchannel could have, with the higher level implying and containing the lower ones. These are: Level 1—Involvement, Level 2—Level of understanding, Level 3—Actual response, e.g., (dis)agreement.

Therefore, if a feedback backchannel communicates (dis)agreement, it also communicates a high level of understanding, and of course active involvement in the conversation. Argyle [18] supported this view by stating that backchannel signals may indicate attention and understanding, provide feedback like agreement, or be a part of mimicry, which in turn could itself signify agreement.

Therefore, agreement and disagreement could be conveyed using backchannel signals and it could be argued that most of the implicit nonverbal cues of (dis)agreement we will examine are of this sort. This means that their polysemic nature should be taken into account in the process of automatic analysis towards the detection of (dis) agreement. For example, smiles, nods and shakes, some of the most important (dis)agreement cues, are also some of the most common backchannels [32,18,16]. However, it is important to keep in mind that their expression during a conversation does not necessarily signal (dis)agreement, but could instead simply convey the level of involvement and/or understanding.

3.2. Facial actions

Action Units (AUs) are atomic facial signals, the smallest visually discernible facial movements. FACS [36], a widely used method for manual labelling of facial actions, defines 9 upper face AUs, 18 lower face AUs, and 5 miscellaneous AUs. FACS also provides the rules for segmentation of AUs' temporal phases (onset, apex, and offset) in a face image sequence. Using FACS, human coders can manually annotate virtually all visible facial display, decomposing it into the AUs and their temporal segments that produced the display. As AUs simply describe facial muscle movements, they are independent of interpretation, and hence they can be used to describe expressions of different attitudes.

For example, *Listener Smiles* are rather indicative of agreement, but, as backchannels, they could have different meanings [32,38]. Brunner [32] argues that smiles act on the third backchannel level, i.e., they provide a positive response to what is being said, they provide acknowledgment of understanding, and keep the listener involved in the conversation.

Ironic smiles are a result of a conflict between two sets of muscles and therefore are not as naturally occurring as benign smiles and can be used to display disagreement [17,37,18,19]. Similar to the ironic smile is the *Cheek Crease*, during which a lip corner is pulled back strongly, deliberately distorting a smile to convey sarcasm [28]. These cues seem to be present in expressions of both posed and spontaneous disagreement [28,21].

Equally important for (dis)agreement detection are the eyebrows. Although Ekman [17] does not specifically mention that eyebrow actions are backchannel signals, he distinguishes between emotional expressions and conversational actions as the two types of facial social signals, and specifically discusses how the eyebrows can play a part in a number of different displays that can serve as a communicative or expressive function. Since *Eyebrow Raise*—AU1 + AU2 and *Frowning*—AU4 are the easiest eyebrow movements to perform, they are often used in combinations with other cues to convey different meanings. They will also exhibit different temporal dynamics (onset, duration, offset) between, for example, their usage as part of emphatic actions and as part of an emotional expression. Ekman separates speaker and listener eyebrow movement, noting that AU1 + AU2 and AU4 are some of the most frequent facial batonic and emphatic actions used by a speaker.[2]

[2] Referring to western cultures.

Eyebrow actions seem to be more directly relevant to disagreement when performed by the listener. Both AU1 + AU2 and AU4 can be used to show disagreement, doubt, uncertainty, or figurative lack of understanding and disbelief at what is being or has been said [33,17,19,20,30]. In this case, AU1 + AU2 will be used in conjunction with other actions: lowering of the lip corners (AU15), relaxation of the upper eyelids (AU43), raising the chin (AU17), raising the upper lip (AU10), and/or head rolling (see Section 2). In the case of a "mock astonishment", AU1 + AU2 is combined with the raising of the upper eyelids (AU5) and a jaw drop (AU26) with abrupt onsets and long durations [17]. AU1 + AU2 could be used in a conversation to convey disagreement with what is being said by the current speaker, but it could also be used to convey agreement when combined with other agreement cues and specifically a smile (AU12), a head nod, or an agreement utterance (e.g., "um-humm') [33,43]. Finally, AU1 + AU2 could serve as an interrogative function, even if the verbal content is not a question per se [17]. Chovil [38] discusses that AU4 could also occur in a number of different functions, including anger, frustration, puzzlement, difficulty, etc. Since the expressions of many of these are believed to accompany displays of disagreement [44,42,19], AU4 could easily be identified as a cue for disagreement.

Another cue that could prove useful in spotting disagreement is listener's *Speech Preparatory Movement*—AU25/AU26 which might signal that the listener wants to respond to what is being said, but presumably might not do so out of politeness [17,22,23]. This is in agreement with Givens [30] arguing that a sudden appearance of *Slightly Parted Lips*—AU25 is a strong signal of nonverbal listener's disagreement. However, the intention movement of speech preparation may be interpreted differently according to people's status/role.

Givens also considers a *Lip Pucker*—AU18 to be an unconscious and first sign of disagreement, as quarrelsome words start forming in the brain. Nose wrinkling or *Nose Twist*, as referred to by Morris [28], could also be used to convey disagreement; Chovil [38] specifically states it may be used by a listener to reject a proposal by the current speaker.

The *Tongue Show*—AU19 can serve as a "sign of unspoken disagreement, disbelief, disliking, displeasure, or uncertainty" [30], even if a co-occurring verbal remark signifies agreement. The *Nose Flare*—AU38, a result of the contraction of the muscles on either side of the nose, which is often accompanied by a sharp intake of air is one more possible cue of disagreement, as is the *Lip Bite* when accompanied by a head shake [30,28].

3.3. Head gestures

Head gestures are crucial cues conveying various social signals, yet they have not been studied as much as facial actions. In agreement and disagreement, specifically, head nods and shakes respectively seem to be the most prevalent and straightforward cues, with nods intuitively conveying affirmation and shakes negation.

Although a *Head Nod* is believed to be a nearly-universal indication of agreement, it could, as a backchannel signal, serve a variety of meanings and functions depending on the number of cycles, amplitude and duration, the co-occurring cues, and the context of the interaction [26,28,27,45,31]. For example, nods usually have an affirmative meaning if they contain a high amplitude and number of cycles, whereas smaller, one-way nods usually serve as signals of involvement in the conversation [16,18]. Nods could also be used as means to allow the current speaker to continue talking, as feedback to the speaker, or even as an attempt to obtain the floor, especially if they are rapid [33,46,18].

Poggi et al. [31] extensively analysed 50 nod cases as they naturally occurred in the course of political debates from the Canal9 Database of Political Debates [47], and suggested a typology of nods, which seems to share many commonalities with Ekman's [17] analysis of the eyebrows. *Speaker nods* seem to have a batonic/emphatic character when they occur in synchrony with speech or with hand gestures, whereas

listener's nods include all levels of backchannel signals as outlined in Section 2. Moreover, speaker's nods can serve as an interrogative function, when combined with AU1 + AU2 or tilted head. *Listener nods* are more likely to occur before the speaker's utterance completion, and combined with the cultural-dependent backchannel vocalisations, like 'mhm' in North America, they can be a third-level backchannel and signify agreement [16]. Poggi et al. [31] overall suggest 12 distinct types of listener's nods, most of which are agreement-related — approval, permission, submission, thanks, agreement, confirmation etc. The exception is the *ironic agreement* type attributed to a nod combined with an asymmetric smile (AU12 L/R), which by itself is considered a cue of disagreement as noted in Section 1. The possibility of a negative meaning conveyed by this otherwise strong indicator of affirmation and agreement is also confirmed by Rosenfeld et al. [48,16], who also suggest that when a listener's nod occurs in the middle of a speaker's sentence it could also signify impatience.

Along with the head nod, the head shake is considered one of the most well recognizable head gestures [28,49]. A *Head Shake* could specifically mean the refusal or reluctance to believe what is being said [26,39,17], but Kendon shows in his extensive study of the gesture [49] that, like head nods, there are more than one meanings that could be attributed to a head shake, depending, as in the case of nods, on the amplitude, number of cycles and duration. Hence, although head shakes can have a dissenting meaning, they could also be part of a speech dysfluency, a question, or laughter among other displays [18,27,49,30]. In many cultures, including Bulgaria, Greece and Turkey, throwing the head back, the direct opposite of the nod, is the dissenting head gesture, which according to Jakobson [39] may communicate disagreement.

Other important head gestures for disagreement detection could be the *Head Roll*, the action of repeatedly tilting the head left and right expressing doubt. Although Morris [28] specifies that this gesture, although universally understood, is not performed by many western cultures, Ekman [17] seems to have noticed the combination of the head roll with AU1 + AU2 to signify disbelief and incredulousness with the speaker's words in Northern Americans and therefore, at least in many of the European countries.

Head movements associated with disagreement may originate in actions that drive the head away from an undesired source. These movements have later generalized to expressions of negativity. Such a movement is the *Cut Off*, which is described by Givens [30] as a form of listener gaze aversion in which the head is abruptly turned away to one side and may indicate uncertainty or disagreement with what is being said. As with all other cues discussed in this section, it is the combination and temporal interaction of such gaze aversion movements with other cues that will probably allow the automatic identification of disagreement from e.g. disinterest or uncertainty, in this case.

3.4. Hand and body gestures

Although facial actions and head gestures will most likely be more useful and generic as they tend to appear more frequently in existing databases, there are a number of other cues that are not less important and could be helpful to the detection of agreement and disagreement. For instance, Givens mentions that the *Adam's-apple Jump* is an "unconscious sign of emotional anxiety, embarrassment, or stress", that could be caused to a listener due to strong disagreement with a speaker [30].

The *Forefinger* and *Hand Wag*, during which an erect forefinger or a hand with the palm outwards, respectively, is wagged from side-to-side has a dissenting meaning [39,28]. The hand wag is local to Central Europe and the equivalent to the *Forefinger Raise* in Eastern Europe, the movement of the raised index finger perpendicular to the line of the shoulders, often pivoted on the elbow [39,28]. The *Hand Cross* is simply a two-handed version of the hand wag, i.e. both hands, palms outwards, are wagged from side-to-side. The *Hand Chop* is the action during which a hand imitates an axe, and the *Hand Scissor* is the action during which the hands imitate the blades of a pair of scissors, with the hands starting crossed and suddenly separating as they move outwards. Morris [28] mentions that both are often used unconsciously during a heated discussion. extbfArm Folding is widely known as signifying a defensive attitude and could also signify disagreement, for example, in situations where one is being verbally attacked in a heated argument [34,30,28]. The *Leg Clamp* is the action during which a crossed leg is clamped by the hands. Although it is not specifically linked to disagreement, it signifies stubbornness, as if the conversation participant was saying: "My ideas, like my body, are clamped firmly in position and will not budge an inch" [28]. Similarly, the *Neck Clamp* and the *Clenched Fist* signal anger with what is being said. The positioning of the feet (*Feet-pointing*) might also play a role in disagreement as there have been observations that for example, jurors unconsciously point their feet away from solicitors they disagree with [30].

However, body postures are considered more fundamental as mood signs than are leg and arm postures [30]. *Body posture and position* relative to others is important, as, for example, one can show agreement, liking, or loyalty by aligning the upper body with someone one agrees with, or angle away from people one dislikes or disagrees with [30]. *Sideways Leaning*, for instance leaning on a wall due to relaxation is referred to as an agreement cue by Bull [34] and Argyle [18]. [3] On the contrary, turning the spinal column away from the person seated beside oneself is "a reliable – and wholly unconscious – sign of disagreement, disliking, or shyness", whereas *Gross Body shifts* may also be used to explicitly convey disagreement [30].

Finally, a whole family of possible cues are the ones that could be considered as unconscious *Self-manipulation*, e.g., a finger on the lips, massaging a hand, or a chin rub. These can provide self-comfort when politeness prevents a listener from expressing disbelief and disagreement [28,30].

3.5. Auditory, second-order and other cues

Cohen [1] states that *Laughter* could also increase the reliability of any reasoning about detecting agreement. Laughter, however, could also be directed at enemies or be part of disagreement or disliking [30]. Possible audiovisual cues of similar nature are *Sighing* and *Yawning*, usually considered a sign of drowsiness or perhaps boredom, which could also occur in certain cases as a sign of mild disagreement [30,21]. Another very interesting cue is the *Throat Clearing*. Givens [30] states that disagreement and uncertainty can act like chemicals or food irritants and cause this cue. However, a conscious throat clearing can be also used to announce somebody's presence or arrival.

The human communication system is fairly complex and it is unlikely that receivers will form intricate representations of attitude on the basis of a single cue. In fact, people most probably infer attitudes like agreement by using a combination of such cues, or through the perception of second order dynamic processes that involve these cues. For example *Mimicry* is a mutual imitation of the interlocutor's nonverbal behaviour and is believed to foster affiliation, agreement, and liking [51]. Mimicking the other person's positive behaviour such as nod or smile could therefore be interpreted as agreement; while the presence of the cue on its own might just signal something else, like submissiveness or interest.

(Dis)agreement could also be inferred by second order cues such as interruption, delay in responding, or utterance length. For example, Greatbach et al. [42] argued that disagreement can be stronger if *Interruptions/Overlapping Speech* occur. However, there are also cases where it has been shown that overlapping speech could be a

[3] However, it is specifically discredited by Bull himself [50] as a weaker sign of agreement.

cue for collaboration and agreement rather than confrontation [52]. *Delays* in responding could be characteristics of a dispreferred activity, such as a disagreement act [23,22]. In these two examples, it is not the act of speaking or not speaking per se that conveys disagreement but the act of violating implicit rules of turn-taking in a conversation. Note, however, that there are certain cases where disagreement becomes the preferred activity, as is the case with responses to compliments [53]. Finally, *Utterance Length* has been shown to be particularly longer in disagreement than in agreement acts [1,41].

3.6. Discussion

The most important conclusion to draw from this section is that no single cue can be unequivocally matched to a social attitude. It will be the temporal interaction of all these cues that will be able to allow us to detect agreement and disagreement. Modelling this interaction is not an easy task, as discussed in Section 7. Also, this section highlights the importance of context and culture in the detection of agreement and disagreement and how different combination of cues can mean different things in different situations.

4. Relevant databases

In order to develop and evaluate automated systems capable of detecting and analysing cues relevant to (dis)agreement, as described above, and to further infer the presence of agreement and disagreement based on this analysis, large collections of training and testing data are needed. This data has to be recorded in naturalistic settings, and be rich in both episodes of agreement and disagreement.

Televised political debates provide an interesting platform for analysing agreement and disagreement-related cues. Since the first televised political debates of the 1960's, debates have become more common, and the audience actually expects the participation of political figures in them [21]. At the same time, the presentation of such debates has evolved from a single-screen approach to multiple split screens, where every reaction each participant shows is available for examination, regardless of who the speaker is. [19] Even if only a single screen is used, the director of the debate will often use close-ups of the speaker or the listeners to give access to the nonverbal aspect of their behaviour [54]. Research has suggested that those watching the debates perceive as less likable the participants who attempt to belittle a debate opponent via cues of nonverbal listener's disagreement. Interestingly enough, political figures are still prepped to display certain cues for that purpose, and hence creating an interesting case of acted agreement and disagreement in a natural context.

Canal9[3] [47] is an example of a database of political debates. The database contains a total of 43 h and 10 min of 70 real televised debates on Canal 9, a Swiss television network. There is always a moderator and two sides that argue around a central issue, with one or more participants on each side. Although this is a "political" debate database, the participants are not always politicians, and the public opinion does not matter to them as much as it would to career politicians. Hence, instances of masked or acted (dis)agreement mentioned above, are rare. Although the recording quality and resolution is suitable for behaviour analysis, including facial expression analysis which requires relatively high-resolution recordings, the debates are pre-edited in one feed and multiple camera angles are used, as one can clearly see in Fig. 1. This means that not all participants are visible at all times and there are times where the camera angle makes automatic visual analysis of, e.g., facial actions very difficult. The database includes (*a*) manual and automatic audio speaker segmentation, i.e., all speaker turns are identified with a label unique to each individual; (*b*) the role of every speaker – moderator or participant – and the stance each participant holds with respect to the central question of interest; and (*c*) manual and automatic shot segmentation and annotation, i.e., all camera angle changes are clearly marked as boundaries of shots which are labelled as *personal* or *other*. Additional annotation of agreement and disagreement episodes was done for [55]. The latter annotation, which is still under construction, currently consists of 53 episodes of agreement, 94 episodes of disagreement, and 120 neutral episodes of neither agreement nor disagreement. These episodes feature 28 participants and they occur over a total of 11 debates. They were selected on the basis of verbal content, and thus, only episodes of direct and indirect agreement and disagreement were included (see Section 2). As the debates were filmed with multiple cameras, and edited live to one feed, the episodes

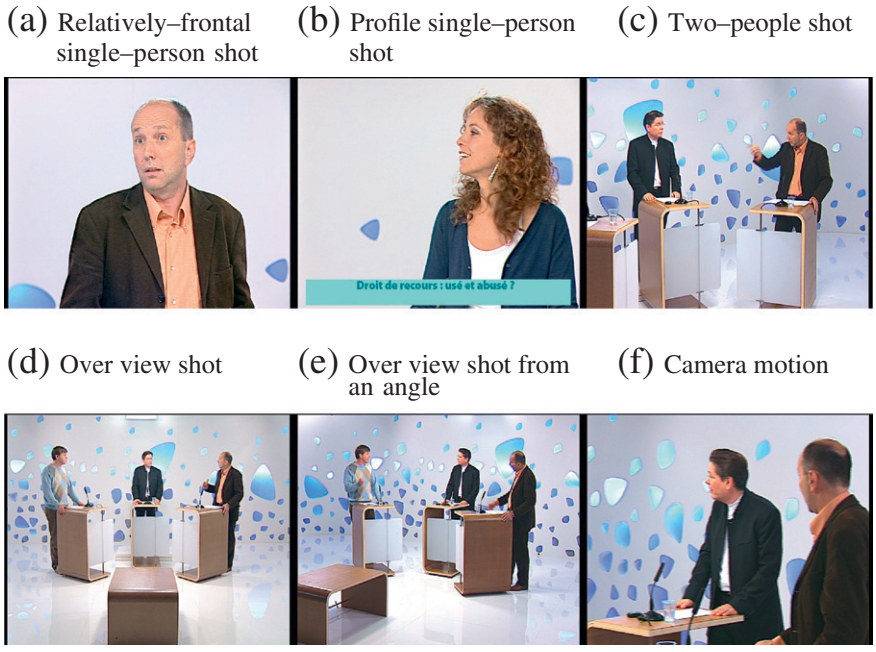


Fig. 1. An example of the difficulty posed by the different kinds of shots and unconstrained environment, partially due to the fact that the participants are not seated, in the Canal9 Database of Political Debates.

selected for this dataset were only the ones that were contained within one personal, close-up shot of the speaker. These episodes were also manually annotated for the head and hand gestures in Tables 1 and 2. These annotations provide ground truth for these gestures and can serve as means to readily evaluate the results of automated detectors of such cues.

Group meeting recordings like the *AMI and AMIDA Meeting Corpus*[3] [56] could also be useful for training and testing automated tools for (dis)agreement detection as they capture instances of human–human interaction, in which occurrences of (dis)agreement are frequent. AMI consists of 100 h of meeting recordings by individual and room-view video cameras, as displayed in Fig. 2. The data, mostly centred around the idea of role-playing in a meeting with the purpose of designing a new remote control, also includes output from a slide projector and a smartboard. A rich set of annotations include transcriptions of the meetings, dialogue act segmentation and labelling, topic segmentation and labelling, summarization of each meeting, head and hand gestures including nods and shakes, higher-level activities (e.g., note-taking), subject visibility (e.g., occlusion), manual head, face, mouth and hand localization, and focus of attention. Most importantly for the task at hand, the AMI corpus is annotated for agreement and disagreement for the 20 out of 170 sessions–meetings for a total of 636 episodes of agreement and 70 episodes of disagreement manifested by 80 participants. What is particularly interesting in the AMI annotation for (dis) agreement is the fact that both the source and target of the (dis)agreement can be analysed. For these 20 meetings, adjacency-pairs of dialogue acts are identified and labelled as 'Support/Positive Assessment', 'Objection/Negative Assessment', and 'Partial Agreement/Support', based on their verbal content. An adjacency-pair is a pair of dialogue acts *A* and *B*, the later of which, *B*, is an assessment of the earlier one, *A*. *B* is then considered the 'source' and *A* the 'target' of this assessment. If the target is an opinion, dialogue act *B* is an episode of agreement or disagreement, per the definitions in Section 2. One issue with AMI is the low quality of the video recordings and their potential for reliable automatic facial expression analysis. Additionally, the participants are often out-of-view, too close, or too far from the camera. However, this corpus can still be used by using prosodic and gestural features for (dis)agreement detection. AMI has been successfully used already for examining (dis)agreement [57], and also for analysing dominance [58], cohesion and leadership [59,60], among others.

AMIDA consists of 10 h of meeting recordings, designed under the same principles as AMI. Although no similar annotations of (dis) agreement expressions exist for AMIDA, it is equally rich in episodes of agreement and disagreement and could be harvested by future researchers for such episodes. The *ICSI* [61] *corpus*, another database of meeting recordings, consists of 75 h of meetings, which however were only recorded for audio.

The *Green Persuasive Dataset*[3], the *IDIAP Wolf Corpus* and the *Mission Survival Corpus* are publicly available databases that could be used to develop and test tools to automatically detect (dis)agreement. Although explicit annotations of agreement and disagreement do not exist for these databases, an interested researcher is expected to find plenty of examples of (dis)agreement. The *Green Persuasive Dataset* was specifically recorded to induce and capture persuasion. The database consists of 8 recorded instances of attempts by one strong pro-green individual to convince others to adopt a 'greener' lifestyle and, naturally, it includes many instances of agreement and disagreement. Each discussion is a dyadic interaction and lasts from 25 to 48 min. The 8 participants are recorded by different cameras which capture the face at a 45-degree angle, making it hard for, e.g., facial action unit detection. Annotations include dimensional labelling of persuasiveness by both the persuadees and third observers. The *IDIAP Wolf Corpus* [62] consists of audio-visual recordings of groups of people playing a total of 15 competitive role-playing games for a total of approximately 7 h. The database features 36 different participants in four groups. It contains plenty of examples of conflict, agreement and disagreement. The *Mission Survival Corpus* [63] is an audiovisual database of multi-party meetings. The participants were asked to reach a consensus on what items are essential for survival. Given the intensive engagement required to reach a consensus, a large number of (dis)agreement examples can be observed. The database consists of 12 meetings featuring 4 participants each and for a total length of approximately 6 h. Finally, 3D tracking of body activity – head, hands and body fidgeting – is also publicly available.

Other naturalistic databases that might not explicitly be rich in examples of (dis)agreement can nevertheless be useful in training automatic tools for detecting cues that could be relevant to (dis) agreement. Human-virtual character interaction recordings like the *SAL* [3,65] and *SEMAINE* [3,66] Datasets could be used in this respect. In SAL and SEMAINE, each user is recorded while interacting with an emotionally-coloured virtual agent, impersonated by an operator, holding a discussion of approximately 5 min each time. SAL consists of approximately 10 h worth of recordings, whereas SEMAINE, which includes recordings from higher-quality videocameras and microphones, also consists of approximately 10 h of material. Annotations

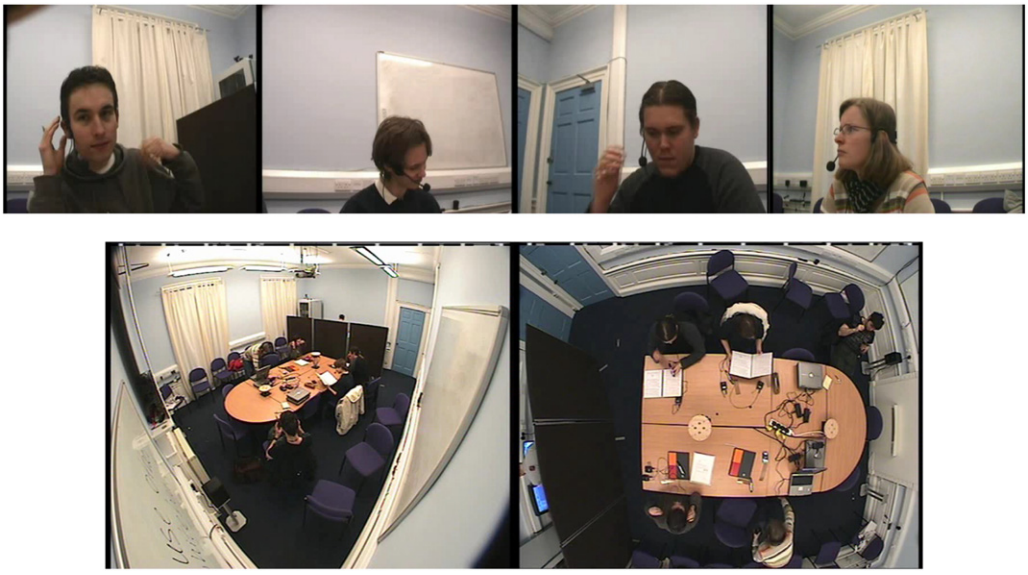


Fig. 2. A snapshot of all cameras in an AMI meeting. Top row: personal shots of each participant. Bottom row: snapshots from the two overview cameras.

include categorical and dimensional labelling of emotions. Finally, these databases could be annotated for episodes of agreement and disagreement; given that the agents have different "personalities", some are bound to cause more agreement or disagreement than others. It is important to note, however, that although these are indeed recordings of spontaneous behaviour, the agent impersonators are very much acting out certain pre-defined characteristics and could not be used for the goal of analysing spontaneous expressions of (dis)agreement.

For an exhaustive overview of such databases that can further be used for training detection tools for (dis)agreement-relevant cues, like facial expressions, see [6,7].

As evident in this section, and particularly Table 3, there is a lack of data for (dis)agreement detection. As we discuss in Section 7, the field of detection social attitudes in general will not be able to move significantly forward without specialised datasets and a significant annotation effort.

5. Face and gesture detection tools

Although in some cases detecting the cues in Tables 1 and 2 is relatively straightforward, as is the case with cues that correspond to Action Units, there are cues that are known to be hard to detect. Two such examples are *Arm Folding* and *Head and Chin Support on a Hand* [5]. However, there are known techniques that would be able to detect most of the cues listed.

5.1. Facial action unit detection tools

When it comes to automatically detecting facial actions, significant advances have been made over the past ten years [78,74,86,77,70,79,82,83,87] (for exhaustive overview see [6,7]). Table 4 lists a few examples of the state-of-the-art systems, omitting older ones that cannot detect combinations of Action Units (AUs) and have not been tested on naturalistic data.

The AU detection systems that exist are usually divided up based on the kind of features they use: geometric or appearance-based. Systems that rely on geometric-based features usually employ the coordinates of a number of fiducial facial points, such as in [86,77], whereas the ones that use appearance-based features such as [78,79] use features like Gabor wavelets or Haar-like features.

The most comprehensive works in automatic AU detection in terms of the number of AUs detected are those of Koelstra et al. [82,83], which detects a total of 30 AUs, and Vural et al. [79], which detects a total of 31 AUs. Hence, these works detect most of the AUs defined in FACS [36], including most of those that could be cues of (dis)agreement, as evident in Table 4. The former work also enables the analysis of the temporal dynamics (onset, apex, offset) of AUs, which could prove very important when distinguishing, for example, a smile (slow symmetric action) from a smirk (fast asymmetric action) [87,88].

Koelstra et al. [82,83] propose an appearance-based approach to automatic coding of AUs and their temporal segments. It presents a dynamic-texture-based approach based on non-rigid registration using free-form deformations, in which the extracted facial motion representation is used to derive motion orientation histogram descriptors in both the spatial and temporal domain, which, in turn, form further input to a set of AU classifiers based on Gentleboost and Hidden Markov Models. This work represents the first appearance-based approach to explicit segmentation of detected AUs into temporal segments, showing that modelling the temporal dynamics of facial expressions significantly improves the performance of such automated systems. Vural et al. [79] also used appearance-based features, a bank of 72 Gabor filters, which are input to separate Support Vector Machines (SVMs) trained for each target AU. However, neither of these methods will work particularly well if rigid head movements are not properly dealt with, which is usually a problem with naturalistic, spontaneous data.

The work of Valstar and Pantic [77] can also detect many of the AUs listed in Tables 1 and 2, including their temporal dynamics (onset, apex, offset), while handling problems with head movement registration rather well. In this work, after a number of facial characteristic points are detected and tracked, a set of spatiotemporal features is extracted from the trajectories of the tracked points, and a combination of Gentleboost, Support Vector Machines (SVMs) and Hidden Markov Models (HMMs) is used to detect AUs and their temporal segments. Fig. 3 shows an example of using the system of Valstar and Pantic [77] on a sequence extracted from the Canal9 database.

The only work reported so far on modelling the temporal correlation among different AUs is that by Tong et al. [84,70]. It applies an appearance-based approach to AU recognition, similar to that by Littlewort et al. [78], using Gabor features and a set of Gentleboost classifiers, one for each target AU. Furthermore it uses a Dynamic Bayesian Network to model the relationships among different AUs.

In their latest work (2010), Tong et al. [70] also use a set of geometric features and take into account rigid head motion by not only modelling the head pan angle (left–right movement), but also the dynamics between head pose (which, however, are discretized to states 'left', 'frontal', 'right') and the AUs. They personalise a 3D shape model – 28 facial feature point coordinates – to a given subject on the frontal-view face, which is subsequently projected on a 2D plane and tracked using active shape models and Gabor wavelets. This 2D geometric shape model is used to improve AU recognition performance. This latter system has been tested on spontaneous data with natural head movement and

Table 3
Summary of the freely-available databases that could be used for automatic (dis)agreement analysis. 'Used by' refers to works using a database specifically for (dis)agreement recognition. In 'Naturalness' S stands for 'Spontaneous', textbfE stands for 'Elicited', and A for 'Acted'.

	Databases							
	Canal 9	AMI/AMIDA	ICSI	SAL	SEMAINE	Green persuasive	Wolf	Mission survival
	2009	2007	2003	2007	2010	2007	2010	2007
Data availability								
Agreement cases	53	656	Not public	–	–	–	–	–
Disagreement cases	94	70	Not public	–	–	–	–	–
Neutral cases	120	None	Not public	–	–	–	–	–
Sessions annotated	11	16	Not public	–	–	–	–	–
Subjects represented	28	64	Not public	–	–	–	–	–
Rich in agreement and disagreement	✓	✓	✓			✓	✓	✓
Naturalness	S	S	S	E	E	E	A	E
Restricted motion			N/A	✓	✓	✓	✓	✓
Video quality	High	Low	No video	Low	High	High	High	High
Used by	[55]	[57]	[40,41,64]	–	–	–	–	–
Reference	[47]	[56]	[61]	[65]	[66]	–	[62]	[63]

Table 4
A few of the most recent cutting-edge AU detection systems published. The AUs listed are only the ones that could be relevant to (dis)agreement detection. P means that an AU was only tested for posed data for a spontaneous AU detection system. The 'Rigid' row mentions how each work handles the case of rigid head motion, if at all. The 'Spont.' row signifies if the work has been tested for spontaneous expressions. 'Temporal' refers to the onset–apex–offset phases of AUs. The references show the evolution of the system in time, but the rest of the fields refer to the latest work in each case.

	Systems						
	Whitehill et al.	Lucey et al.	Valstar et al.	CERT	Yang et al.	Koelstra et al.	Tong et al.
	2006	2007	2007	2008	2009	2010	2010
AU1	✓	✓	✓	✓	✓	✓	✓
AU2	✓	✓	✓	✓	✓	✓	✓
AU4	✓	✓	✓	✓	✓	✓	✓
AU5	✓	✓	✓	✓	✓	✓	✓
AU7	✓	P	✓	✓	–	✓	P
AU9	–	P	✓	✓	–	✓	P
AU10	–	–	✓	✓	✓	✓	–
AU11	–	–	–	✓	–	✓	–
AU12	–	P	✓	✓	✓	✓	✓
AU13	–	–	✓	✓	–	✓	–
AU14	–	–	–	✓	✓	✓	–
AU15	✓	P	✓	✓	–	✓	✓
AU17	✓	P	–	✓	–	✓	P
AU18	–	–	✓	✓	–	✓	–
AU19	–	–	–	✓	–	–	–
AU23	–	P	–	✓	–	✓	✓
AU24	–	P	✓	✓	–	✓	✓
AU25	✓	P	✓	✓	–	✓	✓
AU26	–	–	✓	✓	–	✓	–
AU32	–	–	–	✓	–	–	–
AU38	–	–	–	✓	–	–	–
AU43	–	–	✓	–	–	✓	–
Rigid	–	–	Affine Registration	Simple Eye Alignment	–	Affine Registration	Explicit Pose-AU modelling
Temporal	–	–	✓	–	–	✓	–
Spont.	–	✓	–	✓	–	✓	✓
Features	Haar wavelets	AAM	20 points	Gabor	Haar-like features	FFD	28 points & Gabor
Method	AdaBoost	SVM	GentleBoost, SVM–HMM	SVM	AdaBoost	GentleBoost, HMM	AdaBoost, DBN
Data	DFAT-504 (P) [67]	DFAT-504 (P) [67]; RU-FACS (S) [68]	MMI (P/S)	DFAT-504 (P) [67]; RU-FACS (S) [68]	DFAT-504 (P) [67]	MMI (P/S) [69]; DFAT-504 (P) [67]; SAL (S) [65]	DFAT-504 (P) [67]; ISL [70] (P); Belfast [71] (S); MAD [72] (S); YouTube [70] (S)
Reference	[73]	[74]	[75–77]	[78,68,79]	[80,81]	[82,83]	[84,85,70]

reconfirms the results reported in Valstar and Pantic [77] — the integration of AU relationships and AU dynamics with AU measurements yields a significant improvement of AU recognition.

The system presented in this latter work is the only, to our knowledge, attempt towards head pose-invariant AU recognition. However, when the target data is spontaneous and fairly unrestricted, keeping subjects still and their faces nearly-frontal is not an option. Although pose-invariant AU recognition seems to only now start becoming a research focal point, there have been a fair number of attempts towards pose-independent facial expression recognition in recent years. Most of these have been based on 3D face models, e.g., [89–93]. Although such methods have the advantage of decoupling head pose and facial expressions analysis, they are usually resource-intensive, require time-consuming initialization and the resulting models are often person-dependent and need to be retrained for each expression and head pose intended to be recognized. There exist two 2D and shape-free methods towards solving this issue. Hu et al. [94] recognize facial expressions at five distinct head pan angles. Rudovic et al. [95] map 2D fiducial facial points from head poses within a large range of pan and tilt rotations to the frontal one, enabling the usage of traditional facial expression methods, which require a nearly-frontal face to perform well. Although the works discussed above focus on facial expression recognition of basic emotions, their methods could be adapted for the goal of head pose-invariant AU recognition.

We presented, in this subsection, some of the state-of-the-art AU detection systems. The interested reader is encouraged to consult the exhaustive surveys on the topic by Pantic and colleagues [87,96,6].

5.2. Head gesture detectors

Another set of cues for which there have been explicit detection attempts is head gestures, and particularly *Head Nods* and *Head Shakes*, probably the most important cues for our objective (Fig. 4).

Kawato and Ohya [97] developed a method for head nod and shake detection by using the coordinates of the midpoint between the eyes as a feature for a rule-based system. However, the data used was only 450 frames of three subjects who were following instructions in a lab setting. The number of nod and shake instances was not revealed, but an 86.2% of accuracy was reported for a 13.7% of false positive rate.

Kapoor and Picard [98] used the eye pupils' coordinates as features for their system. A total number of 62 nods and 48 shakes from 10 subjects were recorded using an IR camera, which was post-processed to obtain the pupil coordinates.

The gestures were invoked by questions, to which the participants were instructed to nod for 'yes' and shake for 'no'. Two discrete three-state HMMs were trained for nods and shakes and a ten-frame sliding window was used for temporal localization. 40% of the data was used for training and 60% for testing. The results reported were 81.08% accuracy for nods and 75% for shakes, but there was no discussion about continuous recognition.

Tan and Rong [99] also used the coordinates of the midpoint between the eyes, although it was detected in a different way than in [97]. Similar to Kapoor and Picard [98] they induced the head gesture by asking participants a number of factual 'yes/no' questions, and trained two discrete three-state HMMs for nods – with states 'Up',

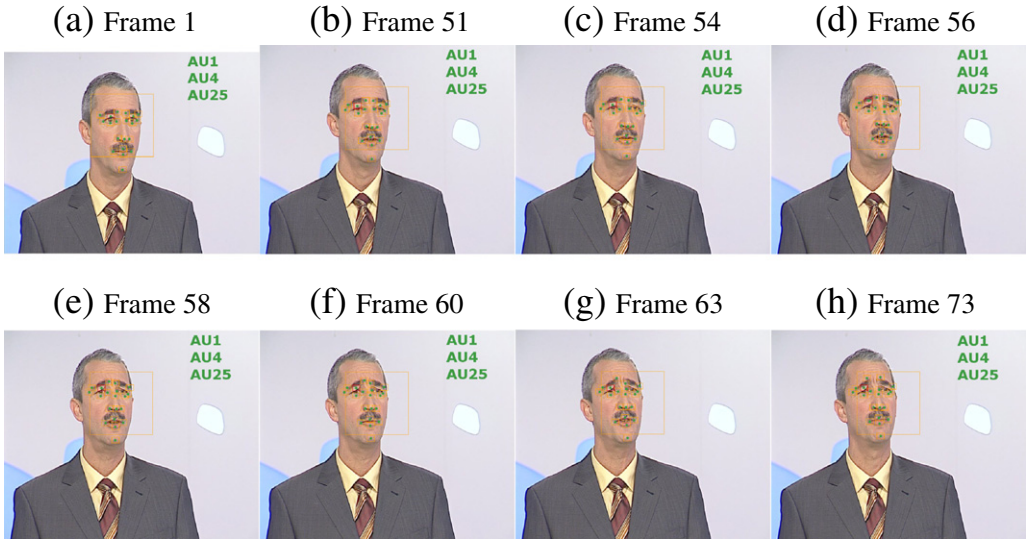


Fig. 3. The output of the AU detection system of Valstar et al. [77] on selected frames from a 75-frame Canal9 sequence of particularly low rigid head motion. The images show the automatically detected facebox, and the 20 automatically tracked fiducial facial points (green dots) used as features in the detection system to detect which AUs are present in each frame — also displayed on each frame above. The system correctly detected the intensely displayed AU1 and AU4, but also the speech-related AU25 in this sequence.

'Down', and 'None' – and shakes – with states 'Left', 'Right', and 'None'. They only performed recognition of presegmented sequences with an 82% accuracy for nods and 89% for shakes.

Fujie et al. [100] also implemented a nod/shake detector in the context of Robot–Human Interaction. They used the mean optical flow over the quartiles of the head region as their features to four continuous HMMs for nod, shake, tilt, stillness and other movements. Their data consisted of 114 min of posed gestures and 90 min of robot–human interaction. They reported a 79.8% accuracy for nod and 61.4% for shakes with an 85.5% and 93.1% precision respectively.

Morency et al. [101] used WATSON [102], a head pose tracker that outputs the three angular head velocities. Each component of the velocity vector is then independently converted into a frequency-based feature which is used in a two-class SVM. They trained the system with 10 natural head gesture sequences taken from interactions with an embodied agent and 11 posed gesture sequences. The test data consisted of 30 video recordings of 9 subjects interacting with an interactive robot, for a total of 20,672 frames, out of which 18,246 were non-gesture frames. True detection rates reported were 75% for nods and 84% for shakes for a fixed false positive rate of 0.05. Wang et al. [103] used similar data and the same system for feature extraction but focused only on recognition, assuming segmentation was already complete. They compared 4-state HMMs – one for each class: nods, shakes and other – (64.3% accuracy), CRFs with one state per class (68.24% accuracy) and a new model they introduced for gesture recognition, Hierarchical CRFs (hCRFs) with 12 states for all classes (85.25% accuracy). Morency et al. [104] extended this work for continuous gesture recognition, by introducing yet another model the Latent-Dynamic CRFs (LDCRFs). They tested the model only for head nods using 79 min of data containing 269 instances of nods. The accuracy for the new model ranged from 65 to 75% for a false positive rate of 20–30% outperforming CRFs, SVMs, HMMs and hCRFs.

Adapted versions of systems like the ones described here could be used to also detect other head gestures, like the *Head Roll* or the *Cut Off*. In fact, depending on the target data, most of the current computer-vision-based head pose estimation systems (for an exhaustive survey refer to Murphy-Chutorian and Trivedi [105]) can be adjusted for detection of an array of head gestures. However, no system reported so far attempts the actual detection of these gestures; only nods and shakes are typically detected.

5.3. Hand and body action detection

Although most of the hand and body-related cues in Tables 1 and 2 have not been explicitly modelled or addressed in any published work of our knowledge, they can either be detected with adapted versions of vision-based human activity recognition methods, or with hand gesture and pose estimation systems. However, the latter have severe limitations, especially when we do not have 3D hand information. This is because vision-based techniques for hand gestures have to deal with problems like self-occlusion, the high dimensionality of the problem (more than 20° of freedom), the environment (background/lighting conditions, clutter), and the speed of hand motions, which can be particularly fast and make the problem of hand tracking and hand gesture recognition rather difficult, considering the low sampling rates of the widely available monocular cameras [106]. Most of the techniques available can only deal with a very limited number of hand gestures and the current limitations – e.g., the hand having to be viewed from a certain angle or the palm having to face the camera [106] – are forbidding when our target data is fairly unconstrained naturalistic human behaviour. For exhaustive surveys of Computer Vision-based approaches to hand tracking and hand gesture recognition, the interested reader is referred to [106].

For our purposes, locating the hand, tracking it, and perhaps being able to tell if a finger is erect or a palm open would be sufficient. These are similar to the requirements of automatic sign language translators, which need hand shapes and locations to interpret signs [107]. Such a system which could be adapted into detecting our (dis)agreement-relevant hand action cues is the one suggested by [107], which segments the head region and extracts hand locations with reference to it, by using a skin colour model and then extracting the hand motion trajectories. Another system that could be adapted to detect the hand actions of interest is the work of Ding and Martinez [108] which uses a particle filter tracker to track hand fiducial points (knuckles, fingertip and wrist), and can handle cases where the hand occludes the face, which occur frequently during natural human interactions. The

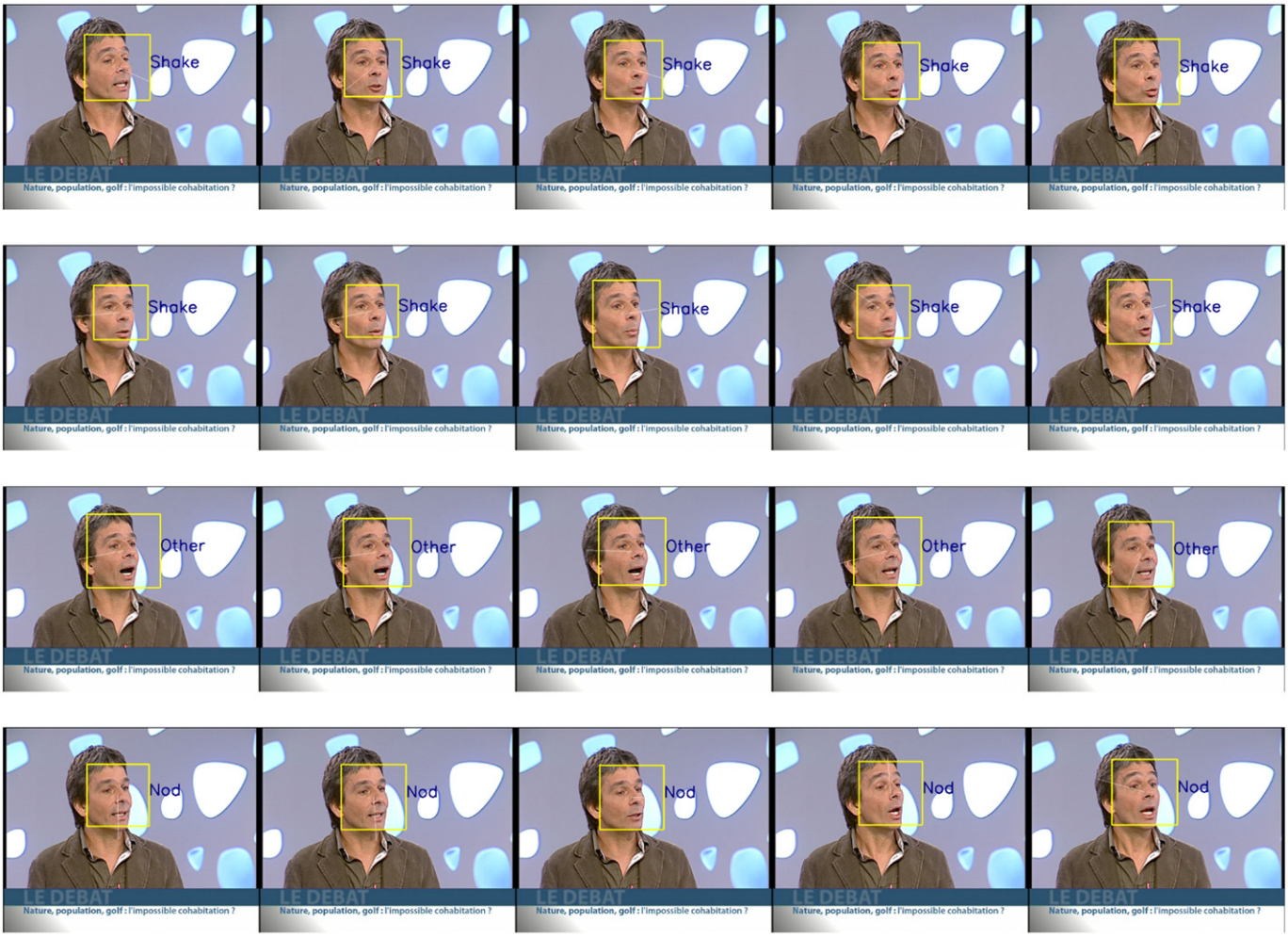


Fig. 4. Our optical flow and HMM-based nod and shake detector in action on difficult Canal9 data. Top two rows: a detected subtle head shake. Bottom two rows: the transition from neither gesture (third row) to a subtle nod (fourth row). The detected face and head angles are shown in each of the displayed frames.

detection of face occlusions by the hand can also be handled by the recent work of Mahmoud et al. [109]. Note, however, that putting such a system together has not been reported yet.

Existing work in the highly active field of human activity recognition and detection can also be used as is or adapted to detect cues, such as the *Forefinger Raise*, *Hand Wag*, *Hand Cross* and *Hands Scissor*. Actions like *Leg* or *Neck Clamp* and *Arm Folding* could also be detected with adaptions of these methods, but with more difficulty, and both dynamic and static features would have to be used for better results. Poppe [139] divides existing work on human activity recognition based on the image representations and the classification method used. "Global representations," in which regions of interest are encoded as a whole to form the image descriptor, have limited applicability as they assume reliable localization of these areas of interest, which might not be realistic in spontaneous data. "Local representation" methods usually follow a bottom-up approach where salient points are detected with respect to both space and time, local patches are calculated using these points, and finally the patches are combined to form a global representation, e.g., [140]. These local patches are represented by different descriptors throughout the literature such as extensions of well-known methods to accommodate for the dimension of time, such as SURF by Willems et al. [141], histogram of oriented flow or gradients (HOG/HOF) by Wang et al. [111], and SIFT by Scovanner et al. [110]. Local representation methods are more robust in the presence of noise and partial occlusion and might prove helpful in detecting hand and body actions relevant to (dis)agreement. Such works include, but are by no means limited to the work of Oikonomopoulos et al. [140,142,138], Marszałek et al. [126], Mikolajczyk et al. [116], Laptev et al. [113], Niebles et al. [123], Shechtman et al. [131], Dollár et al. [124], Wang et al. [111], Willems et al. [141], and Rapantzikos et al. [143,120].

However, most of the existing works on human activity analysis assume segmentation is a pre-processing step. Recent works [127,128,131, 129,130,132–134,136–138] take this into account and have suggested methods for spatial and/or temporal localization and recognition, which mainly rely on correlation or voting. Fig. 5 shows an example of using the work of Oikonomopoulos et al. [138] to detect episodes of 'Forefinger Raise' by employing temporal voting in a 350-frame unsegmented sequence from the highly spontaneous Canal9 database (see Section 4). Table 5 summarizes works on human action analysis along with the features used, whether they can deal with occlusions and dynamic background, and whether they require segmentation as a pre-processing step.

Finally, recent advances in automatic body posture analysis could prove beneficial in detecting (dis)agreement, as discussed in Section 3. Significant recent works have managed to reliably estimate body orientation, such as those by Ando et al. [144], Zhao et al. [145], Van der Bergh et al. [146], and Enzweiler and Gavrila [147]. The problems of finding the alignment of one's body in a given interaction, as well as, the detection of sudden body movements could be solved by using recent human motion analysis methods. Such methods are loosely separated in those that

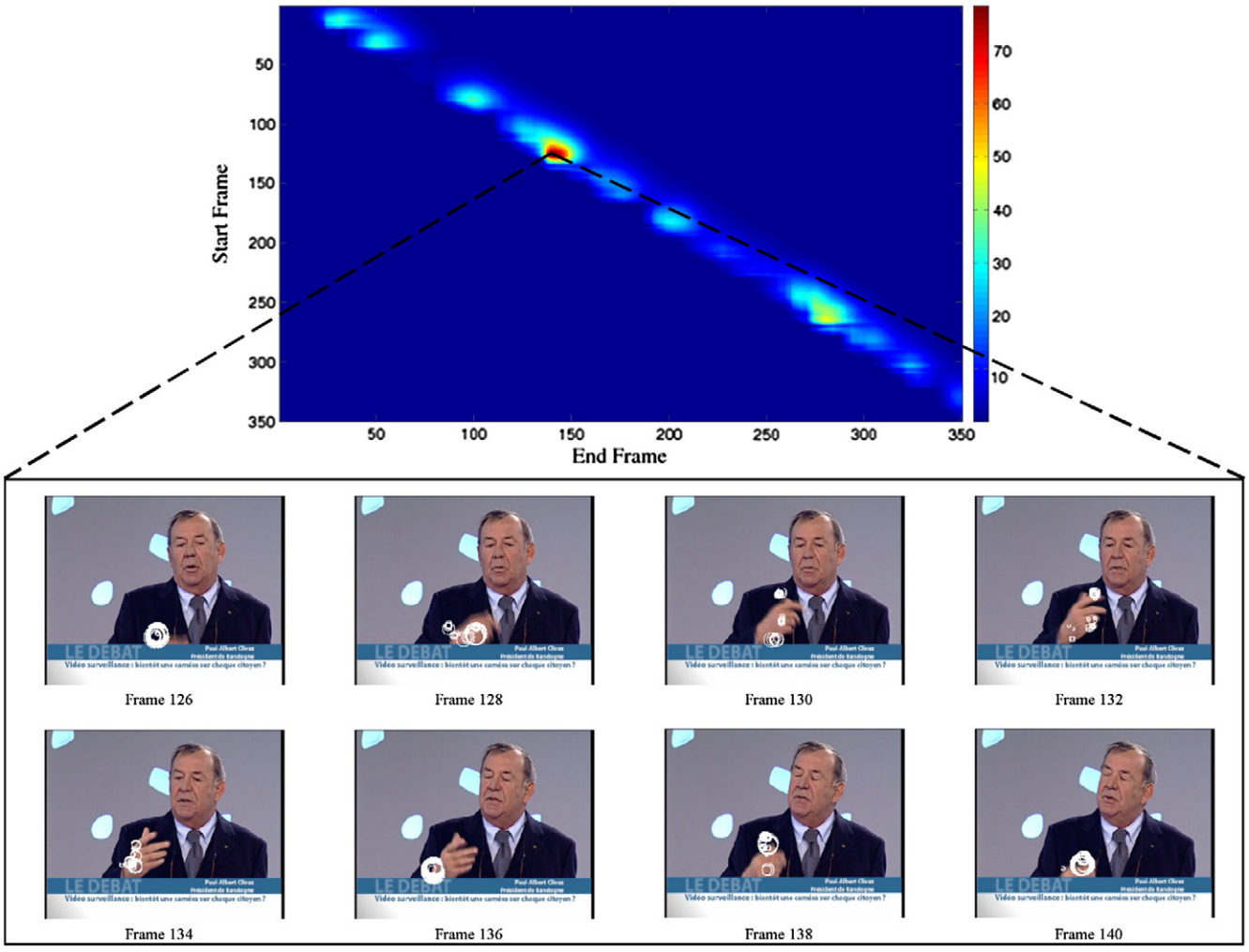


Fig. 5. An example of a forefinger raise detection on Canal9 data using an adaptation of [138]. The top row shows the temporal voting space, where activated ensembles of features in each frame vote for the starting frame of the action given their location. The frames shown below represent the highest-voted start–end frame combination candidate (Start: 126, End: 140) in a 350-frame sequence. The centres of the circles denote the salient point location, whereas their radius denotes their saliency scale. The second highest 'peak' is around the point (281, 264), which is also a true positive for the gesture of interest.

use human body models, e.g., [148–151], and those that are model-free, e.g., [152–156]. For further information, the interested reader should read the exhaustive surveys of Poppe [139], Wang et al. [157], Wang and Singh [158], Gavrila [159], and Aggarwal and Cai [160].

5.4. Detection methods for other cues

In recent years, a lot of research work has also been invested into gaze tracking. *Gaze Aversion* and *Eye Roll* could be a direct application of any monocular gaze-tracking system, as those are surveyed by Hansen and Ji [161]. However, most of those use methods based on infrared (IR) light sources and cameras, which make their use impossible on existing databases that contain episodes of (dis)agreement, such as the ones presented in Section 4. A few methods presented in [161], however, do work with natural light [162–167], but they are still at an early stage and have severe limitations such as a constant headscale assumption. The only work, to our knowledge, that explicitly models and detects *Gaze Aversion* in natural situations is that of Morency et al. [168], who distinguished between gaze aversion and other eye gestures in the context of human–virtual agent interaction using a monocular visual system. *Eye Roll* could also be detected as a direct extension of this work. Finally, methods aimed at detecting the focus of one's attention in natural interactions such as the recent work of Voit et al. [169,170], Ba and Odobez [171] and Asteriadis et al. [13][4] could also be adapted to detect such cues.

Yet another cue, the automatic detection of which has received significant attention over the past decade is *Laughter*, which could be relevant to (dis)agreement detection as discussed in Section 4. Recent work has analysed laughter and can distinguish it from speech, using auditory features [172–178] or a fusion of auditory and visual features [179–182]. The auditory features that have been used for laughter classification include some of the most frequently used features for automatic speech and emotion recognition, namely Mel-frequency cepstral coefficients (MFCC), Perceptual Linear Predictive (PLP) coefficients, pitch, and energy among others. Visual features include geometric features like fiducial facial points [181] or appearance-based features like mean intensities over the cheeks as in [179]. Similarly, a variety of methods for classification have been used including Neural Networks, Hidden Markov Models, Support Vector Machines, and Gaussian Mixture Models among others. Although most of the works mentioned above deal with classification assuming segmentation as a pre-processing step, there are a few researchers, e.g., Kennedy and Ellis [172], Laskowski and Schultz [176], and Knox et al. [177], who have extended their work to

[4] This tool is also available online at the SSPNet web portal (http://www.sspnet.eu).

Table 5
Summary of human action analysis methods that could be used for human action analysis in our context.

Method	Features	Occlusion/dynamic background	Action detection
Scovanner et al. [110]	3D-SIFT	No	–
Wang et al. [111]	3D gradients	No	–
Laptev et al. [112]	HOG–HOF	Yes	–
Laptev et al. [113]	Space–time interest points	No	–
Gilbert et al. [114]	Space–time interest points	No/yes	–
Han et al. [115]	HOG/HOF	Yes	–
Mikolajczyk et al. [116]	Combination	Yes	–
Bregonzio et al. [117]	Space–time interest points	Yes/no	–
Jhuang et al. [118]	C-features	No	–
Reddy et al. [119]	Space–time cuboids	Yes	–
Rapantzikos et al. [120]	Salient points	No	–
Schindler et al. [121]	C-features	No	–
Nowozin et al. [122]	Space–time cuboids	No	–
Niebles et al. [123]	Gabor filters	Yes	–
Dollár et al. [124]	Space–time cuboids	No	–
Schüldt et al. [125]	Space–time interest points	No	–
Marszałek et al. [126]	SIFT–HOG–HOF	Yes	–
Zelnik-Manor and Irani [127]	HOG	No	✓
Ning et al. [128]	Gabor filters	Yes	✓
Shechtman and Irani [129]	Space–time cuboids	Yes	✓
Matikainen et al. [130]	Space–time harris matrices	No	✓
Shechtman and Irani [131]	Space–time cuboids	Yes	✓
Seo et al. [132]	Space–time local steering kernels	Yes	✓
Hu et al. [133]	MHI/HOG	Yes	✓
Yuan et al. [134]	Space–time invariant points	Yes	✓
Junejo et al. [135]	HOG	Yes	✓
Boiman and Irani [136]	Space–time cuboids	Yes	✓
Rodriguez et al. [137]	3D-MACH	Yes	✓
Oikonomopoulos et al. [138]	Space–time interest points	Yes	✓

automatic detection of laughter in unsegmented sequences. Finally, an interesting work that detects smiles vs. laughter in multi-party contexts is that of Kumano et al. [183].

Other auditory cues like *Sighing* and throat clearing can also be detected by adopting methods used for laughter detection, as they are very similar in nature. This adaptation is made easier as there exist freely-available packages, e.g., PRAAT [184], for extracting auditory features like the ones mentioned above. The work of Schuller et al. [185] can detect sighs, whereas the work of Matos et al. in [186] can specifically detect *Throat Clearing* as a sub-goal to cough detection.

Recent work on multimodal silence detection during spontaneous speech [187] could help identify *Silent Pauses* and calculate *Utterance Length*. The latter can also be extrapolated by using the work of Liu et al. teLiu2006. Liu et al. compare Maximum Entropy Models, Conditional Random Fields and Hidden Markov Models in detecting *Interruptions* on telephone conversations and on news data. They found that the former, discriminative models outperform the generative Hidden Markov Models. Lee et al. [189] can also classify interruptions and achieve a large improvement by using multimodal cues. *Filled Pauses* can also be detected using the work of Goto et al. [190], Gabrea and O'Shaughnessy [191], Wu and Yan [192], Schuller et al. [185] and Stouten et al. [193]. Audhkasi et al. [194] attempted the detection of this cue during spontaneous speech based on the premise that prosodic characteristics are stable during a filled pause and outperformed the standard methods of [190] and [193].

When it comes to *Mimicry* there are only a few attempts of automatic detection, one of which is by Keller et al. [35] who mention the possibility of using Motion Energy Analysis [195] to analyse the synchrony between the movements of the participants in a dyadic conversation. Pentland [196] measures mimicry (or "mirroring", as it is called in [196]) in conversational audio patterns, by using auditory backchannels and short words. Kim et al. [197] also measured body movement mimicry, however by using an accelerometer attached on the body, which makes it impossible for using in our context of detecting cues related to agreement and disagreement solely by audiovisual means. Finally, Kim et al. [198] manually analysed the synchrony of genuine smiles via visual means in spontaneous human–human dyadic conversations, i.e., they used human coders to annotate the smiles, which however could presumably be replaced by automatic annotation tools, e.g., AU detection systems like the ones in Table 4, for a fully automatic smile mimicry pattern recogniser. Other important works on measuring mimicry include the work by Madan et al. [199], Veenstra and Hung [200], and Kalimeri et al. [201].

5.5. Discussion

This section can serve as a starting point for a researcher who wants to build a fully-automated system to detect (dis)agreement. Such a researcher would without doubt have to tweak most of the methods described in this section for the dataset in mind and measure those tools that are accurate in their detection. The real difficulty would arise when creating detectors for hand and body-related cues that have not been explicitly modelled in publicly available tools. This would require the collection of a number of episodes of said gestures for the training and testing of each detector. The databases described in Section 4 are candidates for collecting such data.

6. Existing automatic methods for agreement and disagreement

Tables 4 and 6 summarize some of the above discussed, recently proposed methods that could be used as is or adapted to detect the cues relevant to agreement and disagreement, as those listed in Tables 1 and 2. Yet, in spite of this obvious progress in automatic analysis of various behavioural cues, not much effort has been reported so far towards automatic analysis of social attitudes in naturalistic data, let alone the analysis of (dis)agreement. In this section, we discuss all, to our knowledge, works (7) that have dealt with the automatic analysis of (dis)agreement thus far in literature. This discussion includes important works on (dis)agreement classification as a dialogue act.

Hillard et al. [40] attempted speaker (dis)agreement classification on pre-segmented 'spurts', speech segments by one speaker with pauses not greater than 500 ms. The spurt segmentation is reported as an automatic process with human adjustment, without any further explanation of the process. The authors used a combination of word-based and prosodic cues to classify each spurt as 'positive-agreement', 'negative-disagreement', 'backchannel', or 'other'. Most of the results reported included word-based cues, however an overall classification accuracy of 62% was reported for a 17% confusion rate between the agreement and disagreement classes. Similar works by Galley et al. [41] and Hahn et al. [64] also deal with classifying spurts as disagreement and agreement, with [41] also dealing with finding the addressee of the action. Germesin and Wilson [57] also deal with these issues. Wang et al. [220] attempted (dis)agreement detection

Table 6
Methods that could be used as is or adapted for detecting cues for agreement and disagreement. The bolded references are surveys that would aid in the creation of such tools.

Cue	References
Head nod/shake/roll, cut off	[97–100,202,101,103,104], **[105]**
Facial action units	*see Table* 4, **[203–205,5,87,96]**
Smiles vs smirks	[206]
Hand actions	*see Table* 5, [107,108], **[207,106,208]**
Body actions	*see Table* 5, **[157,209–211,139]**
Gaze aversion, eye roll, cut off	[162,163,202,164–168,212,171,13], **[161]**
Laughter, sighing	[213,172–178,214,179–182,172,176,177]
Throat clearing	[186]
Utterance length	[40,215]
Filled pause	[194,192,191,190]
Pause	[216,215]
Interruption	[188,189]
Mimicry	[35,196,198]
Body posture	**[139,157–160]**

not on the spurt level, but on the utterance level. However, the features used by these works included lexical, structural and durational cues and are not comparable with other systems based on nonverbal cues. It is significant to mention, though, that Wang et al. found that incorporating prosodic features improves the performance over using lexical-only features.

The first system that was based solely on nonverbal cues is that by el Kaliouby and Robinson [202], which attempted (dis)agreement classification of acted behavioural displays based on head and facial movements. They used 6 classes: 'agreeing', 'disagreeing', 'concentrating', 'interested', 'thinking', and 'unsure'. They tracked 25 fiducial facial points, out of which they extrapolated rigid head motion (yaw, pitch, and roll), and facial action units (eyebrow raise, lip pull, lip pucker), but also utilized appearance-based features to summarise mouth actions (mouth stretch, jaw drop, and lips parting). They used Hidden Markov Models (HMMs) to detect each head and facial action, by sliding a window of 30 frames – 1 second for their data – at a sliding step of 5 frames. A Dynamic Bayesian Network (DBN) per class was trained to perform the higher-level inference of each of the 'mental states' mentioned above, allowing for the co-occurrence of states. The recognition accuracies were a true positive rate of 76.5% for a false positive rate of 5.4% for agreement and a true positive rate of 81% for a false positive rate of 0.7% for disagreement.

Sheerman-Chase et al. [218] are, to our knowledge, the first research group who has attempted recognition of agreement based on nonverbal cues in spontaneous data. They distinguished between 'thinking', 'understanding', 'agreeing' and 'questioning'. However, they did not include disagreement as a class, because of the lack of data. Their spontaneous data was obtained by capturing the four 12–minute dyadic conversations of 6 males and 2 females. The participants were seated limiting their body and head movements significantly. 21 annotators rated the clips with each clip getting on average around 4 ratings that were combined to obtain the ground truth label. For the automatic recognition, tracking of 46 fiducial facial points was used, which required manual initialization and re-initialization when failures occurred. The output of the tracker was then processed to obtain a number of static and dynamic features to be used for classification. Principal Component Analysis (PCA) was performed on the tracked points in each video frame, and the PCA eigen values were used as features. Similarly to el Kaliouby and Robinson [202], the head yaw, pitch and roll, the eyebrow raise, lip pucker and lip parting were calculated as functions of these tracked facial points. Gaze was also estimated in a similar fashion — the eye pupils were among the points tracked. Rigid head motion, which can often dominate subtle non-rigid facial motion, is accounted for by performing both affine and Levenberg–Marquardt (LM) head pose estimation. Based on the above, it is implied that a neutral, frontal-view frame of each participant is required for some of the static features mentioned above, i.e., the Action Units and the affine head pose estimation. In addition to the features above, they used a temporal frame window of four different scales (80 ms, 160 ms, 320 ms and 640 ms) and fitted a quadratic polynomial to the evolution of their representative static features. The polynomial coefficients were then used as the dynamic features. Feature selection and recognition were accomplished with AdaBoost and the authors claim the results are comparable to human performance, with the area under the ROC curve for agreement being 0.70.

Bousmalis et al. [55] is, to the best of our knowledge, the only work that has attempted recognition of agreement and disagreement based on multimodal (audiovisual) spontaneous data. The (dis)agreement episodes used were part of the Canal 9 Political Debates Database (see Section 4 for more details on this dataset). The dataset was manually annotated for the hand and head gestures listed in Tables 1 and 2, with the exception of a number of them that never appeared in the dataset, and the addition of the 'Shoulder Shrug' and 'Forefinger Raise-Like' gestures. The latter is a 'Forefinger Raise' without an erect index finger. Fundamental frequency (F0) and energy were automatically extracted. This work attempted feature-level fusion of the multiple modalities and presents results obtained by applying Support Vector Machines (SVMs), Hidden Markov Models (HMMs), and Hidden Conditional Random Fields (HCRFs) to the problem of automatic (dis)agreement recognition. It was shown that the latter technique is more suitable for learning the dynamics of the different modalities. Moreover, an automatic model analysis technique for HCRFs is presented, which allows the ranking, according to importance, of the information used by the model. The findings support the fact that the Head Nod and Head Shake, which are considered the most prevalent cues in agreement and disagreement respectively (see Section 3), are also found to be the most discriminative cues by this analysis. Bousmalis et al. also experimented with a nonparametric version of the Hidden Conditional Random Fields (iHCRF) in [219] on the same data, and found that the iHCRF is able to learn the latent structure of the model without specifying a priori the appropriate number of hidden clusters of cues.

There are many possible avenues for future research in this area. One of the most interesting ones involves research with analysing (dis)agreement as a dimension. This presents particular challenges, as discussed in Section 7 and any advancement in that front will advance relevant research in continuous and dimensional analysis of behavioural data in general.

7. Challenges

The automatic analysis of spontaneous agreement and disagreement is still very much a daunting task. There are many challenges in detecting spontaneous agreement and disagreement as is evident in the works that have attempted it.

Agreement and disagreement like all social attitudes are intrinsically ambiguous, high-level semantic events, which typically include interactions with the environment and causal relationships. Nonverbal behavioural cues cannot always unequivocally be associated to a specific emotion or social attitude. As mentioned in Section 2, it is the temporal interaction of a variety of cues that will allow us to identify a specific social attitude. It will be interesting to see if it is possible to differentiate these social attitudes from others e.g. disagreement from disinterest or dislike. *Culture* is another challenging factor that has to be taken into serious consideration when analysing such behaviour. We discussed in Section 2 some of the cues, such as the head nod, the head shake, and the forefinger raise, that may have different interpretation in different cultures. In fact, no complete system of detecting (dis)agreement or any other social attitude can be successful if culture-specific cues and intricacies are not taken into account. Similarly, it's important to consider *context* as a factor that may affect behaviour. Context plays a crucial role for the interpretation of social attitudes. For example, environmental aspects such as the level of visibility and noise influence the behaviour that is

manifested, e.g. lack of proximity due to physical constraints. On the other hand, societal aspects such as the formality of the situation and previously established roles and relations of the persons involved, and individual aspects such as the personality and affective state influence not only the choice of cues to be shown but the interpretation of the observed collection of cues as well.

However, the challenges of achieving this goal start from *naturalistic data collection and annotation*. The domain is still in its early stages and no major efforts have been done yet for the collection of data specifically aimed at the analysis of social attitudes. Most of the works in the literature use data originally aimed at different purposes (e.g., broadcast material, like the Canal9 Database, which has severe disadvantages, primarily the pre-edited single feed by multiple cameras, as discussed in Section 4) and annotated ad-hoc for analysing some specific social phenomena (e.g., the subset of the AMI Meeting Corpus annotated in terms of dominance while originally aimed at speech recognition and computer vision goals). Nevertheless, the need for data is of paramount importance, as one can notice by examining Table 3 of available data for (dis) agreement analysis. Obtaining the ground truth can be very challenging and requires a strict data annotation protocol regarding the definition of (dis)agreement for the annotators, but also regarding the starting and ending points of such an episode. However, social interactions involve a large variety of aspects and no standard annotation or data collection protocol seems to be easy to implement. There is currently a significant need for such data in order for the field to be able to move forward. Without a significant effort for data collection and annotation it will be impossible to tackle solutions to problems like distinguishing disagreement from disinterest and dislike; distinguishing between real and fake agreement; recognizing prefaced disagreement.

Approaching *agreement and disagreement in a dimensional way* presents its own challenges. The main reason for it is the absence of data annotated in such a way and the considerable difficulty to obtain it. Annotation of spontaneous social attitudes is particularly hard even in a categorical approach. Dimension-based annotation of a spontaneous database, e.g. Canal9 [47], in terms of agreement and disagreement would require an annotation tool such as the FeelTrace, which is commonly used for continuous annotation of dimensional data [222]. Such a tool could allow observers to watch a recording and move their cursor within an emotional space to rate their impression about the emotional state of the subject. This could be the 2D emotional space of valence-arousal or a 1D space of an agreement–disagreement dimension, where the highest value could signify strong agreement, whereas the lowest value strong disagreement. This would involve a large number of annotators, as annotator agreement in dimensional annotations tends to be one of the most challenging issues in dimension-based behaviour analysis [7]. However, the biggest challenge yet, when it comes to approaching the automatic detection of (dis)agreement in a dimensional way, is the lack of machine learning techniques that are able to sufficiently tackle the problem. A model that could be used for this task is the Conditional State–Space Model [223], however it is unable to capture latent structure which is vital for such complex behaviour. Specifically, there is a need for effective regression models that are able to capture latent structure, but also model temporal interactions.

The problem of the appropriate *computational model* for the task is still present even when our labels are discrete. The question of what is the most appropriate model for the interaction of all the potential cues is still unanswered. Choosing, for example, the number of potential hidden states for a model with latent variables is not intuitive for our task. An interesting approach is the use of nonparametric models, which allows the convergence to an appropriate model driven by the data. Preliminary results for using nonparametric models to agreement and disagreement analysis [219] show promising results for this kind of approach to modelling data for social attitudes. Another modelling challenge to consider is the *fusion of the different modalities*. We know that the integration of multiple modalities produces superior results in human behaviour analysis when compared to single-modal approaches. The analysis of agreement and disagreement is no different as one can see in [55]. Many of the multimodal systems in the field perform decision-level data fusion in which the input from each modality is modelled independently and the individual recognition results from each classifier are fused at the end. However, this results in the loss of information of mutual correlation between the modalities. A number of model-level fusion methods have been proposed that make use of the correlation between auditory and visual streams [6], however further work is needed in order to address issues such as modelling the temporal correlations within and between modalities.

Furthermore, the analysis of high-level behavioural events such as (dis)agreement requires very accurate recognition of certain very specific cues. There has been significant progress in computer vision when it comes to *detecting low-level behavioural cues*, including some of the ones needed to detect agreement and disagreement, as one can see in Section 5. However, the quality of these tools is not yet at the point where one can use these tools without significantly sacrificing detection rates of the social attitudes. It is important to keep in mind that these tools, as well as any methodology towards the detection of spontaneous agreement and disagreement, will always be bound by the quality and amount of data available to researchers.

8. Conclusion

This paper has provided an overview of the cues (Tables 1 and 2) that, based on Social Psychology literature, could be useful when attempting to automatically detect episodes of (dis)agreement as they naturally occur in discourse. From this overview, it becomes apparent that it is the temporal interaction of these potential cues that will make the difference in detecting agreement and disagreement. We have also presented a number of databases of spontaneous human behaviour, many of which are rich in (dis)agreement episodes. However, it is evident that there is still a great need for data collection and annotation in order for the field to move forward. Moreover, we surveyed the state-of-the-art methods that could be used as is or extended (Tables 4, 5 and 6) to detect the (dis)agreement-relevant behavioural cues in such databases. Specialized tools for many (dis)agreement-related cues still need to be developed. It is important to keep in mind that any computational model for (dis)agreement will heavily rely on the detection accuracy of these tools. Finally, we discussed the very few attempts (Table 7) reported towards the multi-cue analysis of (dis)agreement episodes, which are however limited to classification. Only one of the proposed systems are multi-modal, and only two of them are based solely on nonverbal features. From these two, only one was tested on spontaneous data. There is still no work on the detection (dis)agreement, nor is there any work on (dis)agreement as a dimension. The latter is a particularly challenging task, yet an approach to (dis)agreement that may better reflect all the different kinds of (dis)agreement that are possible. As discussed in Section 7, we can conclude that automatic detection of (dis)agreement is yet to be achieved and that deep investigations of how best to reach this goal are yet to be conducted. However, we believe that this work can serve as an introductory reading to researchers interested in the problem of automatic detection of spontaneous agreement and disagreement based on nonverbal cues and their temporal dynamics.

Acknowledgements

This work has been supported by the European Community's 7th Framework Programme [FP7/20072013] under grant agreement no. 231287 (SSPNet). The work of Konstantinos Bousmalis is currently funded by Google European Doctoral Fellowship in Social Computing. The work by Maja Pantic is also funded in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). Part of this research was supported by the National Center of Competence in Research Affective Sciences and by the Swiss National Science Foundation.

Table 7
Summary of the existing systems that have attempted (dis)agreement classification.

Method			Features	Classifier	Data	Spontaneous
Hillard et al.	[40]	(2003)	Verbal, pause, fundamental frequency (F0), duration	Decision tree	ICSI [61]	✓
Galley et al.	[41]	(2004)	Verbal	Bayesian network	ICSI [61]	✓
el Kaliouby et al.	[202]	(2004)	Head nod, head shake, head turn, head tilt, AU1, AU2, AU12, AU16, AU19, AU20, AU25, AU26, AU27	HMM, DBN	Mind reading [217]	–
Hahn et al.	[64]	(2006)	Verbal	Contrast classifier, SVM	ICSI [61]	✓
Sheerman-Chase et al.	[218]	(2009)	Head yaw, head pitch, head roll, AU1, AU2, AU12, AU18, AU20, AU25, Gaze, head pose	AdaBoost	own	✓
Germesin and Wilson	[57]	(2009)	Verbal, pitch, energy, duration, pauses, speech rate	Decision tree, CRF	AMI [56]	✓
Bousmalis et al.	[55,219]	(2011)	Pitch, energy, head nod, head shake, forefinger raise, 'forefinger raise'-like, forefinger wag, hand wag, hands scissor, shoulder shrug	SVM, HMM, HCRF, iHCRF	Canal9 [47]	✓
Wang et al.	[220]	(2011)	Verbal, pause, duration, speech rate, pitch, energy, vowel duration	CRF	DARPA GALE [221]	✓

References

[1] S. Cohen, A computerized scale for monitoring levels of agreement during a conversation, Univ. Pa. Working Pap. Linguist. 8 (1) (2003) 57–70.
[2] A. Vinciarelli, M. Pantic, H. Bourlard, Social signal processing: survey of an emerging domain, Image Vision Comput. 27 (12) (2009) 1743–1759.
[3] M. Pantic, R. Cowie, F. D'ericco, D. Heylen, M. Mehu, C. Pelachaud, I. Poggi, M. Schroder, A. Vinciarelli, in: Social Signal Processing: The Research Agenda, 2011, pp. 511–538.
[4] M. Pantic, A. Nijholt, A. Pentland, T.S. Huang, Human-centred intelligent human–computer interaction (HCI^2): how far are we from attaining it? J. Auton. Adapt. Commun. Syst. 1 (2) (2008) 168–187.
[5] M. Pantic, A. Pentland, A. Nijholt, T.S. Huang, Human computing and machine understanding of human behavior: a survey, Lect. Notes Comput. Sci. 4451 (2007) 47–71.
[6] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 39–58 http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.52.
[7] H. Gunes, M. Pantic, Automatic, dimensional and continuous emotion recognition, Int. J. Synth. Emotion 1 (1) (2010) 68–99.
[8] D. Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: a review, Image Vision Comput. 27 (12) (2009) 1775–1787.
[9] A. Vinciarelli, Capturing order in social interactions, IEEE Signal Process. Mag. 26 (2009) 133–137.
[10] A. Kapoor, R.W. Picard, Multimodal affect recognition in learning environments, in: Proc. ACM Int'l Conf. on Multimedia, 2005, pp. 677–682.
[11] D. Gatica-Perez, I. McCowan, D. Zhang, S. Bengio, Detecting group interest-level in meetings, in: Proc. IEEE Int'l Conf. on Acoustics, Speech and Signal Processing, 2005, pp. 489–492.
[12] D. Gatica-Perez, Modeling Interest in Face-to-face Conversation from Multimodal Nonverbal Behavior, Academic Press, 2009. Ch. 15.
[13] S. Asteriadis, P. Tzouveli, K. Karpouzis, S. Kollias, Estimation of behavioral user state based on eye gaze and head pose-application in an e-learning environment, Multimed. Tools Appl. 41 (3) (2009) 469–493.
[14] K. Bousmalis, M. Mehu, M. Pantic, Spotting agreement and disagreement: a survey of nonverbal audiovisual cues and tools, in: Proc. IEEE Int'l Conf. Affective Computing and Intelligent Interfaces, 2009, pp. 1–9.
[15] I. Poggi, F. D'Errico, L. Vincze, Agreement and its multimodal communication in debates: a qualitative analysis, Cogn. Comput. (2010) 1–14.
[16] H.M. Rosenfeld, M. Hancks, The relationship of verbal and nonverbal communication, in: The Nonverbal Context of Verbal Listener Responses, Walter de Gruyter, 1980, pp. 193–206.
[17] P. Ekman, Human ethology, Ch. 3.1: About Brows: Emotional and Conversational Signals, Cambridge Univ. Press, 1979.
[18] M. Argyle, Bodily Communication, 2nd edition Methuen & Co, 1988. Ch. 7.
[19] J. Seiter, Does communicating nonverbal disagreement during an opponent's speech affect the credibility of the debater in the background? Psychol. Rep. 84 (1999) 855–861.
[20] J.S. Seiter, H. Weger, Audience perceptions of candidates' appropriateness as a function of nonverbal behaviors displayed during televised political debates, J. Soc. Psychol. 145 (2) (2005) 225–236.
[21] J.S. Seiter, H.J. Kinzer, H. Weger, Background behavior in live debates: the effects of the implicit ad hominem fallacy, Commun. Rep. 19 (1) (2006) 57–69.
[22] A.M. Pomerantz, Second Assessments: A Study of Some Features of Agreements/disagreements, General sociology, University of California, Irvine, 1975.
[23] A.M. Pomerantz, Structures of social action: studies in conversation analysis, studies in emotion and social interaction, Agreeing and Disagreeing with Assessments: Some Features of Preferred/dispreferred Turn Shapes, Cambridge University Press, 1984.
[24] D.W. Cunningham, M. Kleiner, H.H. Bülthoff, C. Wallraven, The components of conversational facial expressions, in: Proc. Symp. on Applied Perception in Graphics and Visualization, 2004, pp. 143–150.
[25] M. Nicolaou, H. Gunes, M. Pantic, Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space, IEEE Transactions on Affective Computing, 2011.
[26] C. Darwin, The Expression of Emotions in Man and Animals, Oxford University Press, USA, 2002.
[27] U. Hadar, T. Steiner, F.C. Rose, Head movement during listening turns in conversation, J. Nonverbal Behav. 9 (4) (1985) 214–228.
[28] D. Morris, Bodytalk: A World Guide to Gestures, Jonathan Cape, 1994.
[29] V. Manusov, A.R. Trees, "Are you kidding me?": the role of nonverbal cues in the verbal accounting process, J. Commun. 52 (3) (2002) 640–656.
[30] D.B. Givens, The Nonverbal Dictionary of Gestures, Signs and Body Language Cue, Center for Nonverbal Studies Press, Sokane, WA, 2002.
[31] I. Poggi, F. D'Errico, L. Vincze, Types of nods. the polysemy of a social signal, in: Proc. Int'l Conf. Language Resources and Evaluation, 2010, pp. 17–23.
[32] L.J. Brunner, Smiles can be back channels, J. Pers. Soc. Psychol. 37 (5) (1979) 728–734.
[33] A. Dittmann, Developmental factors in conversational behavior, J. Commun. 22 (4) (1972) 404–423.
[34] P. Bull, Posture and gesture, in: The Encoding of Disagreement and Agreement, Pergamon Press, 1987, pp. 62–69, Ch. 5.
[35] E. Keller, W. Tschacher, Prosodic and gestural expression of interactional agreement, Lect. Notes Comput. Sci. 4775 (2007) 85–98.
[36] P. Ekman, W.V. Friesen, J.C. Hager, Facial Action Coding System, Research Nexus, Salt Lake City, 2002.
[37] W. Rinn, The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions, Psychol. Bull. 95 (1) (1984) 52–77.
[38] N. Chovil, Discource-oriented facial displays in conversation, Res. Lang. Soc. Interact. 25 (1-4) (1991) 163–194.
[39] R. Jakobson, Motor signs for 'yes' and 'no', Lang. Soc. (1972) 91–96.
[40] D. Hillard, M. Ostendorf, E. Shriberg, Detection of agreement vs. disagreement in meetings: training with unlabeled data, in: Proc. Conf. North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003, pp. 34–36, http://dx.doi.org/10.3115/1073483.1073495.
[41] M. Galley, K. McKeown, J. Hirschberg, E. Shriberg, Identifying agreement and disagreement in conversational speech: use of Bayesian networks to model pragmatic dependencies, in: Proc. Meeting Association for Computational Linguistics, 2004, pp. 669–676.
[42] D. Greatbatch, Talk at work, On the Management of Disagreement Between News Interviewees, Cambridge University Press, 1992. Ch. 9.
[43] P. Ekman, Telling Lies, Berkley Books, 1985.
[44] J. Gottman, H. Markman, C. Notarius, The topography of marital conflict: a sequential analysis of verbal and nonverbal behavior, J. Marriage Fam. 39 (3) (1977) 461–477.
[45] L. Cerrato, Linguistic functions of head nods, in: Proc. Conf. Multi-modal Communication, 2005, pp. 137–152.
[46] S.J. Duncan, Some signals and rules for speaking turns in conversation, J. Pers. Soc. Psychol. 23 (1972) 283–292.
[47] A. Vinciarelli, A. Dielmann, S. Favre, H. Salamin, in: Canal9: A database of political debates for analysis of social interactions, Proc. IEEE Int'l Conf. Affective Computing and Intelligent Interfaces, 2, 2009, pp. 96–99.
[48] H. Rosenfeld, Conversational control functions of non-verbal behaviour, Non-verbal Behav. Commun. (1978) 291–338.
[49] A. Kendon, Some uses of the head shake, Gesture 2 (2) (2002) 147–182.
[50] P. Bull, Posture and gesture, in: Ch. 6: The Decoding of Interest/Boredom and Disagreement/Agreement, Pergamon Press, 1987, pp. 70–84.
[51] T. Chartrand, J. Bargh, The chameleon effect: the perception–behavior link and social interaction, J. Pers. Soc. Psychol. 76 (6) (1999) 893–910.
[52] D. Tannen, Interpreting Interruption Conversation (1993) 53–83 (Ch. 2).
[53] R. Ogden, Phonetics and social action in agreements and disagreements, J. Pragmat. 38 (10) (2006) 1752–1775.
[54] F. Haumer, W. Donsbach, The rivalry of nonverbal cues on the perception of politicians by television viewers, J. Broadcast. Electron. Media 53 (2) (2009) 262–279.
[55] K. Bousmalis, L.-P. Morency, M. Pantic, Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition, in: Proc. IEEE Conf. on Automatic Face and Gesture Recognition, 2011, pp. 746–752.
[56] J. Carletta, Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus, Lang. Resour. Eval. J. 41 (2) (2007) 181–190.
[57] S. Germesin, T. Wilson, Agreement detection in multiparty conversation, in: Proc. ACM Int'l Conf. Multimodal Interfaces, 2009, pp. 7–14.

[58] D. Jayagopi, H. Hung, C. Yeo, D. Gatica-Perez, Modeling dominance in group conversations from nonverbal activity cues, IEEE Trans. Audio Speech Lang. Process. 17 (3) (2009) 501–513.
[59] D. Jayagopi, D. Gatica-Perez, Discovering group nonverbal conversational patterns with topics, in: Proc. ACM Int'l Conf. Multimodal Interfaces, 2009.
[60] H. Hung, D. Gatica-Perez, Estimating cohesion in small groups using audio–visual nonverbal behavior, IEEE Trans. Multimed. 12 (6) (2010) 563–575.
[61] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, in: ICSI meeting corpus, Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing, 1, 2003, pp. 364–367.
[62] H. Hung, G. Chittaranjan, The idiap wolf corpus: exploring group behaviour in a competitive role-playing game, in: Proc. ACM Int'l Conf. on Multimedia, 2010, pp. 879–882.
[63] F. Pianesi, M. Zancanaro, B. Lepri, A. Cappelletti, A multimodal annotated corpus of consensus decision making meetings, Lang. Resour. Eval. 41 (2007) 409–429.
[64] S. Hahn, R. Ladner, M. Ostendorf, Agreement/disagreement classification: exploiting unlabeled data using contrast classifiers, in: Proc. Human Language Technology Conf. of the NAACL, 2006, pp. 53–56.
[65] 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, Lect. Notes Comput. Sci. 4738 (2007) 483–500.
[66] G. McKeown, M.F. Valstar, R. Cowie, M. Pantic, The SEMAINE corpus of emotionally coloured character interactions, in: Proc. Int'l Conf. Multimedia & Expo, 2010, pp. 1079–1084.
[67] T. Kanade, J. Cohn, Y. Tian, Comprehensive database for facial expression analysis, in: Int'l Conf. Automatic Face and Gesture Recognition, 2000, pp. 46–53.
[68] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Fully automatic facial action recognition in spontaneous behavior, in: Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition, 2006, pp. 223–230.
[69] M. Pantic, M. Valstar, R. Rademaker, L. Maat, Web-based database for facial expression analysis, in: IEEE Conf. Multimedia and Expo, 2005, pp. 317–321.
[70] Y. Tong, J. Chen, Q. Ji, A unified probabilistic framework for spontaneous facial action modeling and understanding, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2) (2010) 258–273.
[71] E. Douglas-Cowie, R. Cowie, M. Schroeder, The description of naturally occurring emotional speech, in: Proc. IEEE Int'l Congress of Phonetic Sciences, 2003, pp. 2877–2880.
[72] Multiple aspects of discourse research lab, http://madresearchlab.org/2009URL.
[73] J. Whitehill, C. Omlin, Haar features for facs au recognition, in: Proc. IEEE Int'l Conf. Face and Gesture Recognition, 2006, pp. 97–101.
[74] S. Lucey, A. Ashraf, J. Cohn, Investigating spontaneous facial action recognition through an representations of the face, Face Recognition, 2007.
[75] M. Pantic, I. Patras, Detecting facial actions and their temporal segments in nearly frontal-view face image sequences, IEEE Trans. Syst. Man Cybern. Part B 4 (2005) 3358–3363.
[76] M.F. Valstar, M. Pantic, Fully automatic facial action unit detection and temporal analysis, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2006, pp. 149–156.
[77] M.F. Valstar, M. Pantic, Combined support vector machines and hidden markov models for modeling facial action temporal dynamics, Lect. Notes Comput. Sci. 4796 (2007) 118–127.
[78] G. Littlewort, M.S. Bartlett, I. Fasel, J. Susskind, J. Movellan, Dynamics of facial expression extracted automatically from video, Image Vision Comput. 24 (6) (2006) 615–625.
[79] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, J. Movellan, Automated drowsiness detection for improved driving safety, in: Proc. Int'l Conf. Automotive Technologies, 2008.
[80] P. Yang, Q. Liu, D. Metaxas, Boosting coded dynamic features for facial action units and facial expression recognition, in: Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2007, pp. 1–6, http://dx.doi.org/10.1109/CVPR.2007.383059.
[81] P. Yang, Q. Liu, D.N. Metaxas, Boosting encoded dynamic features for facial expression recognition, Pattern Recognit. Lett. 30 (2) (2009) 132–139.
[82] S. Koelstra, M. Pantic, Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics, in: Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition, 2008.
[83] S. Koelstra, M. Pantic, I. Patras, Dynamic texture based approach to recognition of facial actions and their temporal models, IEEE Trans. Pattern Anal. Mach. Intell. 99 (2010) 1940–1954.
[84] Y. Tong, W. Liao, Q. Ji, Facial action unit recognition by exploiting their dynamic and semantic relationships, Pattern Analysis and Machine Intelligence, IEEE Trans. 29 (10) (2007) 1683–1699, http://dx.doi.org/10.1109/TPAMI.2007.1094.
[85] Y. Tong, W. Liao, Q. Ji, Affective information processing, in: Ch. 10: Automatic Facial Action Unit Recognition by Modeling Their Semantic and Dynamic Relationships, Springer, London, 2009, pp. 159–180.
[86] M. Pantic, I. Patras, Dynamics of facial expressions — recognition of facial actions and their temporal segments from face profile image sequences, IEEE Trans. Syst. Man Cybern. Part B 36 (2) (2006) 433–449.
[87] M. Pantic, M. Bartlett, Machine analysis of facial expressions, in: K. Delac, M. Grgic (Eds.), Face Recognition, I-Tech Education and Publishing, 2007, pp. 377–416.
[88] P. Ekman, Darwin, deception, and facial expression, Ann. N. Y. Acad. Sci. 1000 (2003) 205–221.
[89] Y. Chang, M. Vieira, M. Turk, L. Velho, Automatic 3D facial expression analysis in videos, in: Proc. Int'l Workshop Analysis and Modelling of Faces and Gestures, 2005, pp. 293–307.
[90] Y. Sun, L. Yin, Facial expression recognition based on 3D dynamic range model sequences, in: Proc. European Conf. on Computer Vision, 2008, pp. 58–71.
[91] J. Sung, D. Kim, Real-time facial expression recognition using STAAM and layered GDA classifier, Image Vision Comput. 27 (2009) 1313–1325.
[92] Y. Cheon, D. Kim, Natural facial expression recognition using differential-AAM and manifold learning, Pattern Recognit. 42 (2009) 1340–1350.
[93] T. Wang, J. Lien, Facial expression recognition system based on rigid and non-rigid motion separation and 3D pose estimation, Pattern Recognit. 42 (2009) 962–977.
[94] Y. Hu, Z. Zeng, L. Yin, X. Wei, J. Tu, T. Huang, A study of non-frontal-view facial expressions recognition, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008, pp. 1–4.
[95] O. Rudovic, I. Patras, M. Pantic, in: Coupled gaussian process regression for pose-invariant facial expression recognition, Proc. European Conf. on Computer Vision, 2, 2010, pp. 350–363.
[96] M. Pantic, Machine analysis of facial behaviour: naturalistic and dynamic behaviour, Philos. Trans. Royal Soc. B 364 (1535) (2009) 3505–3513.
[97] S. Kawato, J. Ohya, Real-time detection of nodding and head-shaking by directly detecting and tracking the "between-eyes", in: Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition, 2000, pp. 40–45.
[98] A. Kapoor, R.W. Picard, in: A real-time head nod and shake detector, Proc. ACM Int'l Conf. Perceptive User Interfaces, 15, 2001, pp. 1–5.
[99] W. Tan, G. Rong, A real-time head nod and shake detector using HMMs, Expert Syst. Appl. 25 (3) (2003) 461–466.
[100] S. Fujie, Y. Ejiri, K. Nakajima, Y. Matsusaka, T. Kobayashi, A conversation robot using head gesture recognition as para-linguistic information, in: Proc. IEEE Int'l Workshop Robot and Human Interactive Communication, 2004, pp. 159–164.
[101] L.-P. Morency, C. Sidner, C. Lee, T. Darrell, Contextual recognition of head gestures, in: Proc. ACM Int'l Conf. Multimodal Interfaces, 2005, pp. 18–24.
[102] L.-P. Morency, J. Whitehill, J. Movellan, Generalized adaptive view-based appearance model: integrated framework for monocular head pose estimation, in: Proc. Int'l Conf. Automatic Face and Gesture Recognition, 2008.
[103] S. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, T. Darrell, in: Hidden conditional random fields for gesture recognition, Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2, 2006, pp. 1521–1527.
[104] L.-P. Morency, A. Quattoni, T. Darrell, Latent-dynamic discriminative models for continuous gesture recognition, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[105] E. Murphy-Chutorian, M. Trivedi, Head pose estimation in computer vision: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 31 (4) (2009) 607–626.
[106] A. Erol, G. Bebis, M. Nicolescu, R.D. Boyle, X. Twombly, Vision-based hand pose estimation: a review, Comput. Vis. Image Underst. 108 (1-2) (2007) 52–73.
[107] M.-H. Yang, N. Ahuja, M. Tabb, Extraction of 2d motion trajectories and its application to hand gesture recognition, IEEE Trans. Pattern Anal. Mach. Intell. 24 (8) (2002) 1061–1074.
[108] L. Ding, A. Martinez, Modelling and recognition of the linguistic components in american sign language, Image and Vision Computing 27 (12) (2009).
[109] M. Mahmoud, R. El-Kaliouby, A. Goneid, Towards communicative face occlusions: machine detection of hand-over-face gestures, Lect. Notes Comput. Sci. 5627 (2009) 481–490.
[110] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: Proc. Int'l Conf. Multimedia, 2007, pp. 357–360.
[111] H. Wang, M.M. Ullah, A. Kläser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: Proc. British Machine Vision Conf, 2009.
[112] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[113] I. Laptev, P. Perez, Retrieving actions in movies, in: Proc. Int'l Conf. Computer Vision, 2007, pp. 1–8.
[114] A. Gilbert, J. Illingworth, R. Bowden, Fast realistic multi-action recognition using mined dense spatio-temporal features, in: Proc. Int'l Conf. on Computer Vision, 2009, pp. 925–931.
[115] D. Han, L. Bo, C. Sminchisescu, Selection and context for action recognition, in: Proc. Int'l Conf. on Computer Vision, 2009, pp. 1933–1940.
[116] K. Mikolajczyk, H. Uemura, Action recognition with motion-appearance vocabulary forest, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008, pp. 1–8, http://dx.doi.org/10.1109/CVPR.2008.4587628.
[117] M. Bregonzio, S. Gong, T. Xiang, Recognising action as clouds of space-time interest points, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009, pp. 1–8.
[118] H. Jhuang, T. Serre, L. Wolf, T. Poggio, A biologically inspired system for action recognition, in: Proc. Int'l Conf. on Computer Vision, 2007, pp. 1–8.
[119] K. Reddy, J. Liu, M. Shah, Incremental action recognition using feature-tree, in: Proc. Int'l Conf. on Computer Vision, 2009.
[120] K. Rapantzikos, Y. Avrithis, S. Kollias, Dense saliency-based spatiotemporal feature points for action recognition, in: Proc. Conf. Computer Vision and Pattern Recognition, 2009, pp. 1454–1461.
[121] K. Schindler, L.V. Gool, Action snippets: how many frames does human action require? in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[122] S. Nowozin, G. Bakir, K. Tsuda, in: Discriminative sub-sequence mining for action classification, Proc. Int'l Conf. on Computer Vision, 1, 2007, pp. 1–8.
[123] J. Niebles, L. Fei-Fei, A hierarchical model of shape and appearance for human action classification, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[124] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: Proc. Int'l Workshop Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65–72.
[125] C. Schüldt, I. Laptev, B. Caputo, in: Recognizing human actions: a local svm approach, Proc. IEEE Conf. Computer Vision and Pattern Recognition, 3, 2004, pp. 32–36.
[126] M. Marszałek, I. Laptev, C. Schmid, Actions in context, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009, pp. 2929–2936.
[127] L. Zelnik−Manor, M. Irani, Statistical analysis of dynamic actions, IEEE Trans. Pattern Anal. Mach. Intell. 28 (9) (2006) 1530–1535.
[128] H. Ning, Y. Hu, T.S. Huang, in: Searching human behaviors using spatial–temporal words, Proc. Int'l Conf. Image Processing, 6, 2007, pp. 337–340.
[129] E. Shechtman, M. Irani, Space–time behavior-based correlation-or-how to tell if two underlying motion fields are similar without computing them? IEEE Trans. Pattern Anal. Mach. Intell. 29 (11) (2007) 2045–2056.
[130] P. Matikainen, M. Hebert, R. Sukthankar, Y. Ke, Fast motion consistency through matrix quantization, in: Proc. British Machine Vision Conf, 2008, pp. 1055–1064.
[131] E. Shechtman, M. Irani, Matching local self-similarities across images and videos, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[132] H.J. Seo, P. Milanfar, Detection of human actions from a single example, in: Proc. Int'l Conf. Computer Vision, 2009, pp. 1–8.
[133] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, T.S. Huang, Action detection in complex scenes with spatial and temporal ambiguities, in: Proc. Int'l Conf. on Computer Vision, 2009, pp. 1–8.
[134] J. Yuan, Z. Liu, Y. Wu, Discriminative subvolume search for efficient action detection, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009, pp. 1–8.
[135] I. Junejo, E. Dexter, I. Laptev, P. Pérez, in: Cross-view action recognition from temporal self-similarities, Proc. European Conf. on Computer Vision, 2, 2008, pp. 293–306.
[136] O. Boiman, M. Irani, Detecting irregularities in images and in video, Int. J. Comput. Vis. 74 (1) (2007) 17–31.
[137] M. Rodriguez, J. Ahmed, M. Shah, in: Action mach: a spatio-temporal maximum average correlation height filter for action recognition, Proc. IEEE Conf. Computer Vision and Pattern Recognition, 8, 2006, pp. 1–8.
[138] A. Oikonomopoulos, Spatiotemporal visual analysis of human actions, Ph.D. thesis, Imperial College London (2010).
[139] R. Poppe, A survey on vision-based human action recognition, Image Vision Comput. 28 (2010) 976–990.
[140] A. Oikonomopoulos, M. Pantic, I. Patras, Sparse B-spline polynomial descriptors for human activity recognition, Image Vision Comput. 27 (12) (2009) 1814–1825.
[141] G. Willems, T. Tuytelaars, L.V. Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, Lect. Notes Comput. Sci. 5303 (2008) 650–663.
[142] A. Oikonomopoulos, I. Patras, M. Pantic, in: An implicit spatiotemporal shape model for human activity localisation and recognition, Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 3, 2009, pp. 27–33.
[143] K. Rapantzikos, Y. Avrithis, S. Kollias, Spatiotemporal saliency for event detection and representation in the 3D wavelet domain: potential in human action recognition, in: Proc. Int'l Conf. Image and Video Retrieval, 2007, pp. 294–301.
[144] S. Ando, X. Wu, A. Suzuki, K. Wakabayahsi, H. Koike, Human pose estimation for image monitoring, NTT Tech. Rev. 5 (11) (2007) 1–8.
[145] X. Zhao, H. Ning, Y. Liu, T. Huang, Discriminative estimation of 3D human pose using gaussian processes, in: Proc. Int'l Conf. on Pattern Recognition, 2008, pp. 1–4.
[146] M.V. den Bergh, E. Koller-Meier, L.V. Gool, Real-time body pose recognition using 2D and 3D haarlets, Image Vision Comput. 83 (2009) 72–84.
[147] M. Enzweiler, D. Gavrila, Integrated pedestrian classification and orientation estimation, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2010, pp. 982–989.
[148] Y. Huang, T. Huang, in: Model-based human body tracking, Proc. Int'l Conf Pattern Recognition, 1, 2002, pp. 552–555.
[149] I. Mikić, M. Trivedi, E. Hunter, P. Cosman, Human body model acquisition and tracking using voxel data, Int. J. Comput. Vis. 53 (3) (2003) 199–223.
[150] G. Mori, J. Malik, Recovering 3D human body configurations using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 28 (7) (2006) 1052–1062.
[151] R. Navaratnam, A. Thayananthan, P. Torr, R. Cipolla, Hierarchical part-based human body pose estimation, in: Proc. British Machine Vision Conf, 2005.
[152] L. Taycher, G. Shakhanrovich, D. Demirdjian, T. Darrell, in: Conditional random people: tracking humans with CRFs and grid filters, Proc. IEEE Conf. Computer Vision and Pattern Recognition, 1, 2006, pp. 222–229.
[153] G. Shakhnarovich, P. Viola, T. Darrell, in: Fast pose estimation with parameter-sensitive hashing, Proc. Int'l Conf. Computer Vision, 2, 2003, pp. 750–759.
[154] E.-J. Ong, A. Micilotta, R. Bowden, A. Hilton, Viewpoint invariant exemplar-based 3D human tracking, Comput. Vis. Image Underst. 104 (2-3) (2006) 178–189.
[155] A. Elgammal, C.-S. Lee, in: Inferring 3D body pose from silhouettes using activity manifold learning, Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2, 2004, pp. 681–688.
[156] A. Agarwal, B. Triggs, A local basis representation for estimating human pose from cluttered images, Lect. Notes Comput. Sci. 3851 (2006) 50–59.
[157] L. Wang, W. Hu, T. Tan, Recent developments in human motion analysis, Pattern Recognit. 36 (3) (2003) 585–601.
[158] J. Wang, S. Singh, Video analysis of human dynamics: a survey, Real-Time Imaging 9 (5) (2003) 321–346.
[159] D. Gavrila, The visual analysis of human movement: a survey, Comput. Vis. Image Underst. 73 (1) (1999) 82–92.
[160] J. Aggarwal, Q. Cai, Human motion analysis: a review, Comput. Vis. Image Underst. 73 (3) (1999) 428–440.
[161] D.W. Hansen, Q. Ji, In the eye of the beholder: a survey of models for eyes and gaze, IEEE Trans. Pattern Anal. Mach. Intell. 32 (3) (2010) 478–500.
[162] D.W. Hansen, J. Hansen, M. Nielsen, A. Johansen, M. Stegmann, Eye typing using markov and active appearance models, in: Proc. IEEE Workshop Applications on Computer Vision, 2003, pp. 132–136.
[163] T. Ishikawa, S. Baker, I. Matthews, T. Kanade, Passive driver gaze tracking with active appearance models, in: Proc. World Congress Intelligent Transportation Systems, 2004, pp. 1–12.
[164] D. Hansen, A. Pece, Eye tracking in the wild, Comput. Vis. Image Underst. 98 (1) (2005) 182–210.
[165] J. Wang, E. Sung, R. Venkateswarlu, Estimating the eye gaze from one eye, Comput. Vis. Image Underst. 98 (1) (2005) 83–103.
[166] O. Williams, A. Blake, R. Cipolla, Sparse and semi-supervised visual mapping with the s3p, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2006, pp. 230–237.
[167] H. Yamazoe, A. Utsumi, T. Yonezawa, S. Abe, Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions, in: Proc. Symp. Eye Tracking Research and Applications, 2008, pp. 140–145.
[168] L. Morency, C. Christoudias, T. Darrell, Recognizing gaze aversion gestures in embodied conversational discourse, in: Proc. of the Int'l Conf. on Multi-modal Interfaces, 2006, pp. 287–294.
[169] M. Voit, R. Stiefelhagen, Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios, in: Proc. ACM Int'l Conf. Multimodal Interfaces, 2008, pp. 173–180.
[170] M. Voit, R. Stiefelhagen, Visual focus of attention in dynamic meeting scenarios, in: Machine Learning for Multimodal Interaction, Vol. 5237 of Lecture Notes in Computer Science, 2008, pp. 1–13.
[171] S. Ba, J. Odobez, Recognizing visual focus of attention from head pose in natural meetings, IEEE Trans. Syst. Man and Cybern. Part B 39 (1) (2009) 16–33.
[172] L.S. Kennedy, D.P.W. Ellis, Laughter detection in meetings, in: Proc. NIST Meeting Recognition Workshop, 2004.
[173] N. Campbell, H. Kashioka, R. Ohara, No laughing matter, in: Proc. Europ. Conf. on Speech Commmunication and Technology, 2005, pp. 465–468.
[174] K.P. Truong, D.A. van Leeuwen, Automatic discrimination between laughter and speech, Speech Commun. 49 (2) (2007) 144–158.
[175] B. Schuller, F. Eyben, G. Rigoll, Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech, Lect. Notes Comput. Sci. 5078 (2008) 99–110.
[176] K. Laskowski, T. Schultz, Detection of laughter-in-interaction in multichannel close-talk microphone recordings of meetings, Lect. Notes Comput. Sci. 5237 (2008) 149–160.
[177] M. Knox, N. Morgan, N. Mirghafori, Getting the last laugh: automatic laughter segmentation in meetings, in: Proc. INTERSPEECH, 2008, pp. 797–800.
[178] R. Cai, L. Lu, H.-J. Zhang, L.-H. Cai, in: Highlight sound effects detection in audio stream, Proc. Int'l Conf. on Multimedia and Expo, 3, 2003, pp. 37–40.
[179] A. Ito, W. Xinyue, M. Suzuki, S. Makino, Smile and laughter recognition using speech processing and face recognition from conversation video, in: Proc. Int'l Conf. on Cyberworlds, 2005, pp. 8–15.
[180] B. Reuderink, M. Poel, K. Truong, R. Poppe, M. Pantic, Decision-level fusion for audio–visual laughter detection, Lect. Notes Comput. Sci. 5237 (2008) 137–148.
[181] S. Petridis, M. Pantic, Audiovisual discrimination between speech and laughter: why and when visual information might help, IEEE Trans. Multimed. 13 (2) (2011) 216–234.
[182] S. Petridis, M. Pantic, J. Cohn, Prediction-based classification for audiovisual discrimination between laughter and speech, in: Proc. IEEE Conf. on Automatic Face and Gesture Recognition, Santa Barbara, CA, USA, 2011, pp. 619–626.
[183] S. Kumano, K. Otsuka, D. Mikami, J. Yamato, Recognizing communicative facial expressions for discovering interpersonal emotions in group meetings, in: Proc. ACM Int'l Conf. Multimodal Interfaces, 2009, pp. 99–106.
[184] P. Boersma, Praat, a system for doing phonetics by computer, Glot Int. 5 (9/10) (2002) 341–345.
[185] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, R. Wöllmer, G. Rigoll, A. Höthker, H. Konosu, Being bored? recognising natural interest by extensive audiovisual integration for real-life application, Image and Vision Computing 27 (12) (2009) 1760–1774.
[186] S. Matos, S. Birring, I. Pavord, H. Evans, Detection of cough signals in continuous audio recordings using hidden markov models, IEEE Trans. Biomed. Eng. 53 (6) (2006) 1078–1083.
[187] I. Gonzalez, I. Ravyse, H. Brouckxon, W. Verhelst, D. Jiang, H. Sahli, A visual silence detector constraining speech source separation, in: Proc. Int'l Conf. on Image and Graphics, 2009, pp. 463–470.
[188] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, M. Harper, Enriching speech recognition with automatic detection of sentence boundaries and disfluencies, IEEE Trans, Audio Speech Lang. Process. 14 (5) (2006) 1526–1540.
[189] C.-C. Lee, S. Lee, S. Narayanan, An analysis of multimodal cues of interruption in dyadic spoken interactions, in: Proc. European Conf. Speech Communication and Technology, 2008, pp. 1678–1681.
[190] M. Goto, K. Itou, S. Hayamizu, A real-time filled pause detection system for spontaneous speech recognition, in: Proc. European Conf. Speech Communication and Technology, 1999, pp. 227–230.
[191] M. Gabrea, D. O'Shaughnessy, in: Detection of filled pauses in spontaneous conversational speech, Proc. Int'l Conf. Spoken Language Processing, 3, 2000, pp. 678–681.
[192] C.-H. Wu, G.-L. Yan, Acoustic feature analysis and discriminative modeling of filled pauses for spontaneous speech recognition, J. VLSI Signal Process. 36 (2-3) (2004) 91–104.
[193] F. Stouten, J. Duchateau, J.-P. Martens, P. Wambacq, Coping with disfluencies in spontaneous speech recognition: acoustic detection and linguistic context manipulation, Speech Commun. 48 (11) (2006) 1590–1606.

[194] K. Audhkhasi, K. Kandhway, O. Deshmukh, A. Verma, Formant-based technique for automatic filled-pause detection in spoken english, in: Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing, 2009, pp. 4857–4860.
[195] A. Bobick, J. Davis, The recognition of human movement using temporal templates, IEEE Trans. Pattern Anal. Mach. Intell. 23 (3) (2001) 257–267.
[196] A. Pentland, Socially aware, computation and communication, Computer 38 (3) (2005) 33–40, http://dx.doi.org/10.1109/MC.2005.104.
[197] T. Kim, A. Chang, A. Pentland, Meeting mediator: enhancing group collaboration with sociometric feedback, in: Proc. Conf. Computer Supported Collaborative Work, 2008, pp. 457–466.
[198] K. Kim, M. Eckhardt, N. Bugg, R.W. Picard, The benefits of synchronized genuine smiles in face-to-face service encounters, in: Proc. Int'l Conf. on Computational Science and Engineering, 2009, pp. 801–808.
[199] A. Madan, R. Caneel, A. Pentland, Voices of attraction, in: Proc. Int'l Conf. on Augmented Cognition, 2005.
[200] A. Veenstra, H. Hung, Do they like me? Using video cues to predict desires during speed-dates, in: Int'l Conf. Computer Vision, 2011, pp. 838–845.
[201] K. Kalimeri, B. Lepri, T. Kim, F. Pianesi, A. Pentland, Automatic modeling of dominance effects using granger causality, in: Human Behavior Understanding, Vol. 7065 of Lecture Notes in Computer Science, 2011, pp. 124–133.
[202] R. el Kaliouby, P. Robinson, in: Real-time inference of complex mental states from facial expressions and head gestures, Proc. IEEE Conf. Computer Vision and Pattern Recognition, 3, 2004, p. 154.
[203] M. Pantic, J. Rothkrantz, Automatic analysis of facial expressions: the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000) 1424–1445.
[204] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recognit. 36 (1) (2003) 259–275.
[205] Y. Tian, T. Kanade, J. Cohn, Facial expression analysis, in: S. Li, A. Jain (Eds.), Handbook of Face Recognition, Springer, 2004, pp. 247–275.
[206] M.F. Valstar, H. Gunes, M. Pantic, How to distinguish posed from spontaneous smiles using geometric features, in: Proc. ACM Int'l Conf. Multimodal Interfaces, 2007, pp. 38–45.
[207] S. Mitra, T. Acharya, Gesture recognition: a survey, IEEE Trans. Syst Man Cybern. Part C 37 (3) (2007) 311–324.
[208] G.R.S. Murthy, R.S. Jadon, A review of vision based hand gestures recognition, Int. J. Info. Technol. Knowl. Manage. 2 (2) (2009) 405–410.
[209] T.B. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, Comput. Vis. Image Underst. 104 (2-3) (2006) 90–126.
[210] V. Krüger, D. Kragic, A. Ude, C. Geib, The meaning of action: a review on action recognition and mapping, Adv. Robot. 21 (13) (2007) 1473–1501.
[211] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, Machine recognition of human activities: a survey, IEEE Trans. Circ. Syst. Video Technol. 18 (11) (2008) 1473–1488.
[212] L.-P. Morency, T. Darrell, Conditional sequence model for context-based recognition of gaze aversion, in: Proc. ACM Int'l Conf. Multimodal Interfaces, 2008, pp. 11–23.
[213] A. Lockerd, F.M. Mueller, LAFCam: leveraging affective feedback camcorder, in: Proc. CHI, Human Factors in Computing Systems, 2002, pp. 574–575.
[214] G. Guo, S. Li, Content-based audio classification and retrieval by support vector machines, IEEE Trans. Neural Netw. 14 (1) (2003) 209–215.
[215] D. Baron, E. Shriberg, A. Stolcke, Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues, in: Proc. Int'l Conf. Spoken Language Processing, 2002, pp. 949–952.
[216] M. Marzinzik, B. Kollmeier, Speech pause detection for noise spectrum estimation by tracking power envelope dynamics, IEEE Trans. Speech Audio Process. 10 (2) (2002) 109–118.
[217] S. Baron-Cohen, O. Golan, S. Wheelwright, J.J. Hill, Mind Reading: the Interactive Guide to Emotions, Jessica Kingsley, London, 2004.
[218] T. Sheerman-Chase, E.-J. Ong, R. Bowden, Feature selection of facial displays for detection of non verbal communication in natural conversation, in: IEEE Int'l Conf. on Computer Vision Workshops, 2009, pp. 1985–1992.
[219] K. Bousmalis, L.-P. Morency, S. Zafeiriou, M. Pantic, A discriminative nonparametric bayesian model: infinite hidden conditional random fields, in: Neural Information Processing Systems (NIPS) Workshop on Bayesian Nonparametrics, 2011.
[220] W. Wang, S. Yaman, K. Precoda, C. Richey, G. Raymond, Detection of agreement and disagreement in broadcast conversations, in: Proc. Association for Computational Linguistics, 2011, pp. 374–378.
[221] S. Strassel, C. Cieri, A. Cole, D. DiPersio, M. Liberman, X. Ma, M. Maamouri, K. Maeda, Integrated linguistic resources for language exploitation technologies, in: Proc. Int'l Conf. Language Resources and Evaluation, 2006, pp. 3052–3056.
[222] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, M. Schroder, FEELTRACE: an instrument for recording perceived emotion in real time, in: Proc. ISCA Workshop on Speech and Emotion, 2000, pp. 19–24.
[223] M. Kim, V. Pavlovic, Discriminative learning for dynamic state prediction, IEEE Trans, Pattern Anal. Mach. Intell. (2009) 1847–1861.