Meeting Behavior Detection in Smart Environments: Nonverbal Cues that Help to Obtain Natural Interaction *

Mannes Poel, Ronald Poppe and Anton Nijholt
University of Twente, Dept. of Computer Science, Human Media Interaction Group
P.O. Box 217, 7500 AE Enschede, the Netherlands
{mpoel,poppe,anijholt}@ewi.utwente.nl

Abstract

Unobtrusive and multiple-sensor interfaces enable observation of natural human behavior in smart environments. Being able to detect, analyze and interpret this activity allows for the implementation of various applications, including real-time surveillance and real-time support in smart home and office environments. We focus on smart meeting environments, where nonverbal behavioral cues sometimes tell more about issues such as discussion participation, involvement and contribution, than information obtained from verbal contributions. An important aspect of this behavior is the interaction between meeting participants. We regard the special case where some of the participants are at physically different locations, which hinders natural interaction. We discuss how we can exploit the ability of a sensor-equipped environment to detect nonverbal interaction cues and use these to allow and improve natural interaction between collaborating participants at distributed locations. We focus on research efforts to detect nonverbal interaction cues by looking at various modalities. Also, we discuss how information obtained from the fusion of these modalities allows us to generate and display behavioral cues that allow remote participants to take part in a distributed meeting in a natural way.

1. Introduction

The automatic detection and analysis of human behavior has become increasingly important in a number of application domains. Anomaly detection in surveillance systems, monitoring of elderly people to facilitate independent living and automatic adjustment of lighting in smart homes are well-known examples of such applications. In this paper, we focus on the detection and analysis of human behavior in smart environments. Specifically, we focus on meetings. In meetings, nonverbal signals are important cues for a number of important reasons. Cues such as gaze, body posture shifts, coughs and laughs have been found indicative for issues such as involvement and attentions, and communicative acts such as addressing, turn-taking and grounding.

While automatic analysis of meetings has become common for off-line browsing (e.g. playback, summarization, search), real-time analysis of human behavior opens up a number of potentially interesting applications. On-line analysis and visualization of interpreted human behavioral cues can help to improve meeting efficiency. Examples are the automatic detection of (dis)agreement among participants, or facilitation of turn-taking. In these cases, the automatic detection provides meta-information about the meeting. While there are certainly scenarios where such information is valuable, most meeting participants are able to obtain this kind of information from their own observations. This is especially true when modalities are considered that can also be processed by humans (i.e. audio and video).

More interestingly, real-time analysis of behavior could help in mitigating communication deficiencies when the meeting participants are geographically distributed. Major drawbacks of video-conferencing systems include the lack of proper nonverbal expression. Examples are the relative difficulty of floor-grabbing by means of posture shifts, and the lack of eye contact, which results in degraded turn-taking abilities. When only a subset of modalities can be used, for example when a participant only has access to a mobile phone, these effects are even more problematic.

Being able to detect and analyze nonverbal cues in a physical meeting room allows for their use in virtual meeting environments. That is, distributed meeting environments where not necessarily all the participants are in the same room, but where we have to accommodate a situation where, for example, a remote meeting participant requires natural interaction with other individuals or groups of people that are in different physical locations. We investigate

*This work was supported by the European IST Programme Project FP6-033812 (Augmented Multi-party Interaction with Distant Access, publication AMIDA-141), and is part of the ICIS program. ICIS is sponsored by the Dutch government under contract BSIK03024.

978-1-4244-2154-1/08/$25.00 ©2008 IEEE

how we can exploit the ability of a sensor-equipped environment to detect nonverbal interaction cues and use these to allow and improve natural interaction between collaborating participants in distributed locations. We discuss the research efforts we and our research partners undertake to detect these cues by looking at various modalities and the fusion thereof. Also, we discus how information obtained from this fusion allows us to generate and display (e.g. using embodied agents) behavioral cues that allow remote participants to take part in a distributed meeting in a more natural way. This allows them to grab the floor, to establish eye contact, to know about (dis)agreement that is building up, and to know about nonverbally announced topic shifts in a discussion. Modelling the interpretation of nonverbal interaction cues certainly allows us to develop tools and environments that help to provide (nonverbal) communication information to distributed meeting participants, allowing them to interact and cooperate close to natural multi-party face-to-face interaction.

Throughout this paper, we will use a running example of a distributed meeting setting. We will regard a meeting room with a small group of participants, and one meeting participant who is in the train. This person has a phone and is further monitored by a small laptop camera. We further assume that limited data transmission between laptop and meeting room is possible. Figure 1 shows an impressions of such a scenario.

The paper is structured as follows. Section 2 introduces the concept of a smart meeting room, and discusses typical human behavior in meetings. In Section 3, we elaborate on a number of real-time detection techniques for these cues. The synthesis of interpreted behavior is important when part of the modalities cannot be communicated. We briefly discuss this topic in Section 4. Finally, we summarize our work and present promising avenues for future work.

2. Behavior in meetings

We consider smart meeting rooms where multiple unobtrusive sensors are available, including cameras and microphone arrays (a number of synchronized microphones in known arrangement). Such environments allow for real-time, synchronized recording of audio and video. Automatic analysis of meetings has many interesting applications, and the domain is interesting from a research perspective as well. The indoor situation allows for good control, which aids in more robust analysis of both verbal and nonverbal human behavior. But more importantly, meetings exhibit a large range of nonverbal behaviors that are performed both consciously and unconsciously. Also, the interaction with other participants makes it an interesting domain for the analysis of realistic human behavior. When distributed meetings are regarded, the interaction with remote participants is likely to change. In a smart meeting room, it can be observed which behavior cues are missing. Moreover, using virtual meeting rooms, we can generate those cues that allow for, and improve interaction between remotely present participants.

(a) (b)

Figure 1. (a) Overview of the IDIAP smart meeting room that has been used to collect multi-party meeting data for the AMIDA project. Note the microphone array with cameras attached in the center of the table. (b) remote participant, working in the train.

2.1. AMIDA project

Within the EU-funded AMIDA project [1], research is conducted in the area of smart meeting rooms. Topics under investigation are automatic speech analysis, detection of nonverbal cues and meeting summarization. Within the project, a large audio-visual corpus of one hundred hours of meetings has been recorded, with both manual and automatic annotations for a large number of observations and events. In the near future, meetings with remote participants will be recorded, either in a video-conferencing scenario, or a scenario with distant access, such as our example with the train. For an overview of the project, the reader is referred to [4].

3. Behavior detection

Meetings are multi-party activities where participants collaborate on a shared task. Apart from verbal communication, nonverbal cues contribute to a large extent to the effectiveness of the meeting. When not all participants are at the same physical location, communication of nonverbal cues is hindered. Automatic detection and analysis of these cues, and the subsequent visualization for the remote party can allow the communication of certain nonverbal cues, and thus improve meeting effectiveness.

In this section, we discuss the automatic detection of behavior cues in the video and audio modalities, and the fusion thereof. To allow for use in distributed meetings, the real-time aspect of the detection and analysis is of key importance.

[1] http://www.amidaproject.org

3.1. Cues from the head and face

The head and face are important sources for relevant behavioral cues. For instance, gaze is the primary source for determining the visual focus of attention (VFOA). Eye gaze is hard to detect and it has been observed that head orientation gives a reliable approximation of gaze direction. Also, facial expressions are often used to communicate social signals.

3.1.1 Head tracking and orientation

A good overview of gaze-based VFOA detection in meeting is presented in [2]. Most of the work is concerned with simultaneously tracking the head and determining the head orientation using particle filters. The observations are composed of texture features determined by Gabor or Gaussian filters, skin color features and silhouette features. A dynamical model is introduced, which can take into account a prior on likely head positions and orientations in meetings. Such an approach allows for fast estimation of VFOA.

A sequence of head orientations is used as input to a Hidden Markov Model (HMM) where the hidden states are a discrete set of possible VFOAs. Such models can be used for a single participant, but can also model the VFOA from multiple participants simultaneously. This approach has some conceptual advantages since movement of participants is often correlated (e.g. looking at the speaker). In a joint model, observations can be conditioned on the context. In recent experiments, an significant increase over the individual model is achieved [3]. Inclusion of audio features or addressee information could further improve the accuracy, since participants are more likely to focus on the speaker or participant that is being addressed [9].

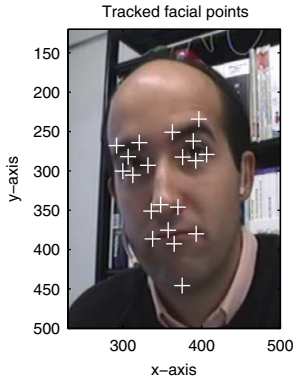


Figure 2. Example of 20 facial points, tracked using [13].

3.1.2 Social signals

The head can be a source of certain social signals (e.g. nodding and shaking). These signals can be determined by classification algorithms based on the output of the head tracking and orientation framework described in the previous subsection. The face is an even more important source of social signals. Facial expressions can be used to estimate the participant's affective state. Facial expressions can be described by the Facial Action Coding System (FACS) introduced by Ekman and Friesen [5]. Detection of facial expressions is challenging, due to the subtleness of the facial actions, inter-personal differences and the presence of head rotations. Also, often only a single camera is used. Within the context of meetings, Patras and Pantic use a coupled particle filter to track 20 facial points and map these to action units [13], see also Figure 2. These action units are then classified into affective states.

Time series of facial points can be analyzed using Principal Component Analysis (PCA) [17]. In realistic scenarios, the first principal components encode the head pose, including translation, rotation and scale. The other components encode interpersonal differences, facial expressions, corrections for the linear approximations and noise factors of the tracked results. Figure 3 shows the first 12 principal components. As such, this technique can be used to determine the head location and orientation by using the first components, and use the remaining components for facial expression analysis. We discuss the use of these features for (multi-modal) laughter detection in Section 3.4.

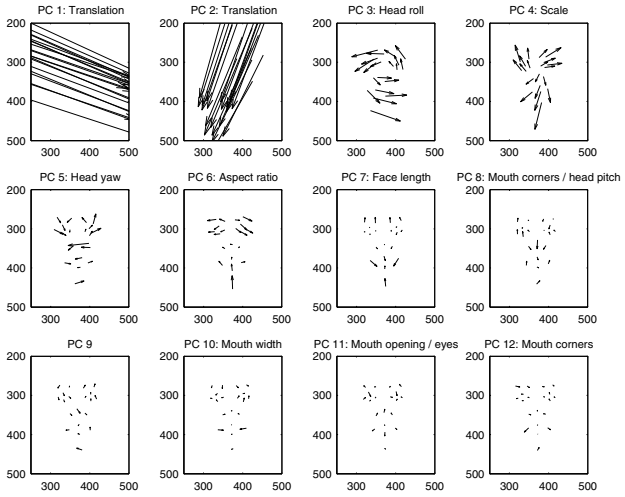


Figure 3. A visualization of the first 12 principal components, applied to located 2D feature points. The arrows point from -3 σ to 3 σ, where σ is the standard deviation.

3.2. Cues from the body

Apart from the head, the body is also informative for a number of communicative cues. We discuss the detection of human poses and actions.

3.2.1 Human pose

Human pose is an important cue for a number of behavioral states, such as attention, attitude, and as an intermediary step to recognize actions and gestures (see also the next section). The recovery of human poses from video is challenging due to variations in human appearance, the large space of physically possible poses and the fact that usually only a single camera can be used. Moreover, real-time performance is required when dealing with a remote participant scenario, as we discuss in this paper.

Although the recovery of human poses from video is an active research domain (see e.g. [15]), relatively few works have focussed on pose recovery in meetings. Smart rooms allow for control of lighting conditions, which makes the task more manageable. However, interactions with the environment (e.g. handling of documents, pen or laptop) and other persons introduce occlusions and make the use of priors on the pose space less powerful.

The work of Lee and Nevatia [11] focusses on meetings. They track persons, use a number of templates to find the face, the shoulders and limbs. In a subsequent step, the 2D locations are lifted to 3D using a data-driven Markov-chain approach. The authors do not report real-time performance, in contrast to [16], who recover poses of presenters in meetings, viewed from the front.

Poses are described in terms of angles of key joints (elbows, shoulders, neck) or the position of the limbs. From such representations, certain behaviors can be derived. When the arms are crossed, the participant conveys a more closed attitude. Similarly, an upright pose is indicative of a higher level of attention. In the train example, such cues can also be used to adapt the transmission directly, e.g. by adjusting the volume when the person leans forward.

3.2.2 Actions and gestures

Actions and gestures can be regarded as sequences of human poses. Usually, these have a clear semantic meaning, such as raising a hand to ask for the floor, or pointing at presentation slides. Unlike activities (e.g. presenting, giving a monologue), gestures and actions have a spatio-temporal character, which has motivated the use of graphical models (e.g. HMMs) for recognition. Within the context of meetings, gesture recognition is performed with classes standing up, sitting down, pointing, writing, shaking head and nodding [1]. While the latter two are best detected using facial cues, the former are typically performed with the entire (upper) body. Pose features are obtained from an adaptation of [16] and yielded reasonable recognition accuracy on pre-segmented sequences. Accuracy on automatically segmented sequences resulted in severely degraded performance. This can be mainly explained by the fact that gestures in meeting contexts are often subtle. Speech supporting gestures show many similarities with semantic gestures, which makes distinction difficult. It would therefore be interesting to see how a multi-modal approach to gesture recognition would perform. Such models could be conditioned on the role of the participant (e.g. speaker, listener, presenter).

Instead of pose features, features can be used that are directly derived from video. While such an approach circumvents the demanding pose recovery task, invariancy to viewpoint, person and specific setting needs to be established in the action recognition task. Recent research on the recognition of actions from spatio-temporal interest points is an interesting venue for future research.

3.3. Cues from audio

We briefly discuss the audio modality in this section. While speech is an informative cue in human behavior detection, nonverbal cues can contribute to a better understanding of the behavior when this is not explicitly communicated. Moreover, the audio modality is often used as a complementary source of cues in multi-modal detection of behavior.

Nonverbal cues from audio include laughs, coughs, sighs and yawns but also acoustic and prosodic features from speech such as pitch, energy, timing and duration can be used. Traditionally, these features have been used for the detection of affect (see [10] for an overview), but are also applied in recognition of other cues that are frequently observed in meetings, such as person tracking and detection of a participant's level of dominance. In this paper, we discuss audio cues in combination with visual cues in Section 3.4.

3.4. Multi-modal signal processing

Fusing cues from different modalities and from different sources, such as head and body, lies at the heart of (social) intelligent human-human interaction [8]. Examples of behavior detection from a single modality have been described in the previous section. In this section, we focus on fusion, also known as multi-modal signal processing, and present three examples of multi-modal algorithms.

Fusion of cues from the different modalities can be done at the feature level, which is termed early fusion. Features of all sources are concatenated into a single feature vector which is used as input for the classifier. Due to this concatenation, the classifier has a large state space and has high learning complexity. The opposite approach is to fuse at the decision level. For each source, a classifier is constructed. The combinatorial complexity in the training phase is avoided, thus significantly reducing learning time. The decisions of each classifier are fused into a global decision. The disadvantage of this high level fusion is the lack of interaction between the information of different sources in the classification process.

3.4.1 Participant tracking

As discussed in Section 3.1, video can be used to find and track faces. When multiple audio sensors are available, the coarse location of the speaker can be obtained by looking at differences in volume and timing between the sensors. In [6], a multi-modal system is presented that combines input from both video and audio into a dynamic graphical model. A particle filter approach is used for inference of both location and speaker activity of multiple participants. The visual observations are derived from the shape and spatial structure of human heads [18]. Audio features are obtained from a microphone array. In addition to increased robustness over a single modality, the multi-modal approach proves to be useful in the presence of occlusion in the video.

3.4.2 Dominance

Given a meeting, some participants are likely to be more dominant, which could lead to irritation and less efficient discussions. In a setting where the interactions between participants are transformed, recognition of dominant participants could be used to achieve a more balanced meeting. To this end, [7] investigate different features from either audio or video, for the recognition of perceived dominance. The fusion of both modalities remains an open issue.

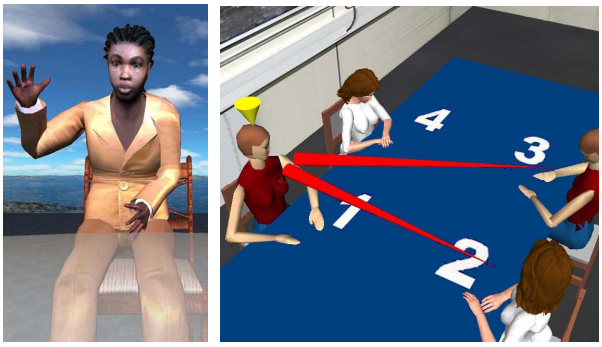
3.4.3 Social signals

Multi-modal recognition of affective behaviors, especially from audio and visual cues, has received much interest [20]. One particular example is that of laughter detection, evaluated on meeting recordings in the AMIDA database. In [17] fusion is applied at the decision level. Audio features are classified based on the approach described in [19]. The visual features are determined by the approach described in Section 3.1. A Support Vector Machine (SVM) is used to fuse the two modalities. In [14] the role of head orientation is investigated into more detail, and was shown to improve recognition accuracy slightly.

4. Behavior synthesis

Consider our running example of a remote participant in a train. The nonverbal behavioral cues of this participant cannot be shown directly to the users in the smart meeting room. For example, the remote participant is unable to make eye-contact with any of the other participants. This inhibits natural interaction. Therefore, it appears advantageous to try to maintain as many nonverbal behavioral cues in the communication. This might require the transformation of the cues to other modalities [12]. In this process, the *cues* remain unchanged, while the *representation* will in general be affected. For example, due to the lack of direct eye contact, a remote participant might cough softly in an attempt to grab the floor. An on-screen representation of the remote participant in the smart room might display an embodied agent that leans forward and makes eye contact, in addition to the cough (see Figure 4(a)). This is likely to be more effective, but it requires models to synthesize the required behaviors in the appropriate modalities.

The remote participant can also enable autonomous listening mode, in which the embodied agent shows natural listening behavior so that the remote participant can check relevant information on her laptop. In the meantime, her behavior is not displayed in the meeting room. This could help to avoid distractions for other participants in the meeting. She could be notified if a relevant topic is discussed in the meeting or when she is being addressed.



(a) (b)

Figure 4. (a) Example of visualization of remote participant performing a floor grab. (b) Virtual meeting room with current speaker and addressees indicated.

Also, the smart meeting room can be represented on the laptop of the remote person, for instance a low-resolution visualization. Figure 4(b) shows a visualization of a virtual meeting room (VMR). Nonverbal cues detected by the sensors in the smart meeting room can be synthesized in this visualization. For instance, the focus of attention can be highlighted and represented in more detail.

Much of the work that we described is ongoing within the AMIDA project. Techniques for detection and recognition are being developed. At the same time, we are performing experiments with synthesis of behavior, and the use of VMRs as a means to achieve transformed social interaction.

5. Summary and future work

We presented research for meeting support in smart rooms. We regarded the special case where one or more participants are not present in the same room, i.e. the meeting is distributed. In such a case, communicative cues between remote participants cannot be observed directly. This inhibits natural interaction, since mechanisms such as turn-

taking, grounding and addressing are not functional. These cues often serve a communicative function, and are important for the progress and the effectiveness of the meeting.

Smart meeting rooms are equipped with microphones and cameras, which allow for the observation of the meeting participants. Research is presented towards the recognition of nonverbal cues and how these can be displayed to the remote participant. The social cues of the remote participant can be synthesized using the available modalities. In this process, we can adapt the rendering of the cues to the current situation. This implies that certain cues can be can be made more explicit by transforming them to other modalities, while others will be suppressed.

Future work will be aimed at improving the detection and recognition of nonverbal behavioral cues. Within the AMIDA project, a demonstrator for remote meeting support will be developed. This demonstrator can be used to study in more detail the consequences of the transformed interaction. Specifically, we aim at investigating the cues that are needed for effective, natural interaction between the remote participants and the participants in the smart meeting room.

References

[1] M. Al-Hames, T. Hain, J. Cernocký, S. Schreiber, M. Poel, R. Müller, S. Marcel, D. van Leeuwen, J.-M. Odobez, S. Ba, H. Bourlard, F. Cardinaux, D. Gatica-Perez, A. Janin, P. Motlícek, S. Reiter, S. Renals, J. van Rest, R. Rienks, G. Rigoll, K. Smith, A. Thean, and P. Zemcik. Audio-visual processing in meetings: Seven questions and some AMI answers. In *Proceedings of the workshop on Machine Learning for Multimodal Interaction (MLMI'06)*, volume 4299 of *Lecture Notes in Computer Science*, pages 24–35, Bethesda, MD, 2006. Springer Verlag.

[2] S. O. Ba and J.-M. Odobez. A study on visual focus of attention recognition from head pose in a meeting room. In *Proceedings of the International Workshop on Machine Learning for Multimodal Interaction (MLMI'06)*, volume 4299 of *Lecture Notes in Computer Science*, pages 75–87, Bethasda, MD, 2006.

[3] S. O. Ba and J.-M. Odobez. Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, pages 2221–2224, Las Vegas, NV, 2008.

[4] H. Bourlard and S. Renals. Recognition and understanding of meetings overview of the european AMI and AMIDA projects. IDIAP-RR 27, IDIAP, Martigny, Switzerland, 2008.

[5] P. Ekman and W. V. Friesen. *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, 1978.

[6] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan. Audiovisual probabilistic tracking of multiple speakers in meetings. *IEEE Transactions On Audio, Speech, And Language Processing*, 15(2):601–616, 2007.

[7] H. Hung, D. Jayagopi, C. Yeo, G. Friedland, S. O. Ba, J.-M. Odobez, K. Ramchandran, N. Mirghafori, and D. Gatica-Perez. Using audio and video features to classify the most dominant person in a group meeting. In *Proceedings of the International Conference on Multimedia (MM'07)*, pages 835–838, Augsburg, Germany, 2007.

[8] A. Jaimes and N. Sebe. Multimodal Human -Computer Interaction: A survey. *Computer Vision and Image Understanding (CVIU)*, 108(1-2):116–134, 2007.

[9] N. Jovanović, R. op den Akker, and A. Nijholt. A corpus for studying addressing behaviour in multi-party dialogues. *Language Resources and Evaluation*, 40(1):5–23, 2006.

[10] P. N. Juslin and K. R. Scherer. *The New Handbook of Methods in Nonverbal Behavior Research*, chapter Vocal expression of affect. Oxford University Press, Oxford, United Kingdom, 2005.

[11] M. W. Lee and R. Nevatia. Human pose tracking in monocular sequence using multi-level structured models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, to appear.

[12] A. Nijholt, J. Zwiers, and J. Peciva. Mixed reality participants in smart meeting rooms and smart home enviroments. *Journal of Personal and Ubiquitous Computing*, to appear.

[13] I. Patras and M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG'04)*, pages 97–102, Seoul, Korea, 2004.

[14] S. Petridis and M. Pantic. Fusion of audio and visual cues for laughter detection. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR'08)*, Niagara Falls, Canada, to appear.

[15] R. Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding (CVIU)*, 108(1-2):4–18, 2007.

[16] R. Poppe, D. Heylen, A. Nijholt, and M. Poel. Towards realtime body pose estimation for presenters in meeting environments. In *Proceedings of the International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision'2005 (WSCG'2005)*, pages 41–44, Plzen, Czech Republic, 2005.

[17] B. Reuderink, M. Poel, K. P. Truong, R. Poppe, and M. Pantic. Decision-level fusion for audio-visual laughter detection. In *Proceedings of the workshop on Machine Learning for Multimodal Interaction (MLMI'08)*, Utrecht, The Netherlands, to appear.

[18] K. Smith, S. Ba, D. Gatica-Perez, and J.-M. Odobez. Tracking the visual focus of attention for a varying number of wandering people. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(7):1212–1229, 2008.

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

[20] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, to appear.