Natural Interaction expressivity modeling and analysis*

George Caridakis†
Intelligent Systems, Content and Interaction Lab
National Technical University of Athens
gcari@image.ntua.gr

Konstantinos Moutselos
Dept. of Computer Science and Biomedical Informatics
University of Central Greece
kmouts@ucg.gr

Ilias Maglogiannis
Department of Digital Systems.
University of Piraeus
imaglo@unipi.gr

ABSTRACT

Behavior, including non verbal, expressiveness is an integral part of the communication process since it can provide information on the emotional state and the user's performance when the aim of the interaction is measurable. Long term temporal measurements can also assist in monitoring the user for either emergencies or long term mood instabilities. Current article presents research work on the computational formalization and analysis of full body 3D expressivity in Natural (bodily) Interaction within the framework of Pervasive Assistance. Expressivity dimensions are selected as the most complete approach to body expressivity modeling, since they cover the entire spectrum of expressivity parameters related to emotion and affect. In this study five expressivity parameters are computationally formalized, using different approaches based on silhouette, limbs position and joints rotation, for each expressivity feature. These approaches are then evaluated in terms of their effectiveness in modeling the expressivity aspect in question. The modeling effectiveness of each approach is assessed using Linear Discriminant Analysis (LDA) and its coefficients on the automatically extracted parameters, defined in the computational formalization, against an experimental dataset consisting of extreme expressions (positive and negative) of the investigated expressivity aspects. The The experimental results confirm that the proposed Fading Silhouette Motion Volumes (FMSV) approach, is the most effective in modeling body expressivity.

PETRA "13, May 29 - 31 2013, Island of Rhodes, Greece

http://dx.doi.org/10.1145/2504335.2504378

†also affiliated with the Dept. of Cultural Technology, University of the Aegean

General Terms

Affective Computing, Pervasive Assistance

Keywords

Natural Interaction, Body expressivity, Linear Discriminant Analysis, Behavioral Analysis

Categories and Subject Descriptors

J.3 [**Life And Medical Sciences**]: Medical information systemsPerformance, Design

1. INTRODUCTION

An increasing number of studies from various disciplines have shown that body expressions are as powerful as facial expressions in conveying emotions [24] [23]. According to Mehrabian and Friar [17] and Wallbott [26], changes in a person's affective state are also reflected by changes in body posture. Recently, introduced and established technologies become more and more ubiquitous and the interaction with these technologies becomes natural. A typical example is that of the computer game industry where body movement is not only a means to control the interaction between us and the games, but also a way to capture and affect our own emotional and cognitive performances. In a wider perspective, these emotional and cognitive performances could be monitored in Pervasive Assistive context in order to serve as behavior-driven assistance or rehabilitation and patients during therapy progress evaluation tools.

Alternative to conventional means of interaction, Natural Interaction (NI), and bodily interaction specifically, is increasingly attracting the attention of researchers in related research areas. Within Natural Interaction context body actions, movement and postures, either intentional or not, convey important emotional content, enhanced with qualitative expressive cues. Body motion or posture qualitative aspects (formulated using different approaches) communicate affective and emotional content and are embodied in the direct and natural emotional expression of body movement [6]. Accordingly, the application areas of Natural Interaction are constantly increasing both in numbers as well as in breadth, including pervasive and assistive environments.

Non verbal behavioral cues are by definition connected to alternative means of interaction such as NI. An abundance of research within the fields of psychology and cognitive science related with non verbal behavior and communication stress out the importance of qualitative expressive characteristics of body motion, posture, gestures and in general human action during an interaction session. Expressivity of body movement is a qualitative cue that is, or at least should be, incorporated in the design process of such applications. Alex Pentland [18] wrote in the Scientific American: "The problem, in my opinion, is that our current computers are both deaf and blind: they experience the world only by way of a keyboard and a mouse. ... I believe computers must be able to see and hear what we do before they can prove truly helpful". Moving a step further, we might add, that they should also interpret fully and appropriately what they see and hear.

2. RELATED WORK

Within the wider research area of Affective computing, research has been carried out towards gesture or body interaction analysis and related articles can be found both on the IEEE Transactions on Affective Computing (TAC) as well as in the two books that have been recently published [19] and [20] and deal with the entire spectrum of research related to Affective Computing. Investigating though Natural Interaction in three dimensions and performing comparative studies regarding full body expressivity formalization, remains a scarcely studied domain, although some research work has been performed recently on actor portrayals corpus [10].

The majority of affective recognition systems of body posture and movement have focused on extracting emotion information from dance sequences [3]. Kapur et al. [13] used acted dance movements, both from professional and amateur dancers. The use of dance movements for building affect recognition systems is interesting; however, these movements are exaggerated and purposely geared toward conveying affect. Body movements and postures that occur during day-to-day human interactions and activities are typically more subtle and not overtly emotionally expressive. On the other hand, our research focuses on acted extremes of specific expressivity parameters which provide a solid experimental corpus for evaluating the modeling effectiveness.

Affective analysis, aiming to classify interaction segments into emotions based on gestures or body information, has been proposed [1], [14] and [7]. Additionally, such information has been fused with modalities used widely in Affective computing such as facial expressions and speech prosody [11], [4] and [22]. Recently, an extensive survey [15] has been published discussing the literature review on affective body expression perception and recognition. It investigates whether there are universal aspects in such models and finally it provides and overview of automatic affect recognition systems using body expressions as at least one input modality. On the other hand, Zeng et al. in [27] provide a survey of multimodal, including body gestures, affect recognition methods in spontaneous (non-acted) natural interactions. Kleinsmith and Bianchi-Berthouze in [2] examined automatic recognition of affect from whole body postures using a low-level posture description in an acted situation while more recently [16] they studied non-acted postures and subtle affective states in video games.

Although the automatic affect recognition, either unimodal or in a multimodal setting, of body expressivity has been recently quite active, it is questionable if the basis of such approaches, the modeling aspect of expressivity has been sufficiently studied. This aspect is included in the recognition process but the modeling effectiveness is not investigated per se. The effectiveness of the expressivity modeling is measured usually through the respective recognition accuracy. Other aspects, such as the corpus construction, stimuli, feature extraction and processing, classification and fusion issues, are also included in the recognition process making it almost impossible to derive conclusions on the modeling itself. Research presented here focuses solely on the modeling aspect of body expressivity and its effectiveness.

3. EXPRESSIVITY MODELING

Expressivity modeling, in this work, consists of dimensionsal description of non verbal behavior [5]. Five parameters modeling behavior expressivity have been defined at the analysis level, as a subset of features derived from the field of expressivity synthesis. These are namely: 1. Overall activation 2. Spatial extent 3. Temporal 4. Fluidity 5. Power. Such an approach is an established way of describing expressivity, tackling all emotion expression parameters [26].

3.1 Experimental dataset

The corpus used as experimental dataset, as also discussed in section 4, was designed and recorded in order to maximize the disassociation of the two extremes for every expressivity feature, which will be formalized in section 3.2. Although its size is clearly small, four users, it is helpful in simplifying the problem and assisting in the extraction of safer conclusions regarding the optimal modeling approach.

In more details the users were asked to perform two extremes variants, corresponding to their interpretation of maximum and minimum value, of the same movement for every expressivity feature. The meaning of each expressivity feature was clarified previously and the user performance was recorded using Microsoft's Kinect. The subject performed two variants of a gesture of his choice for every expressivity feature. The two variants consisted of the two extremes for that expressivity parameter and 'high expressivity' corresponds to a high value, as perceived by the subject, for that parameter and respectively 'low expressivity' to a low value. The body motions/gestures were not acted, in the sense that they were not predefined or instructed by the experiment designers but rather were decided by the subjects. Although, not directly correlated to emotion, some expressivity parameters could be directly correlated to affective dimensions, given a dimensional emotion representation approach. For example, Overall Activation is clearly and directly related to Activation/Arousal axis of the Evaluation/Activation or Valence/Arousal emotional space.

Silhouette binary images S, depth image maps D and skeleton joint rotations J were extracted from the recordings. S, D and J will consist the input to the full body expressivity formalization process and are illustrated in Figures 1 and 2.

Figure 1: Depth and silhouette images provided by Kinect

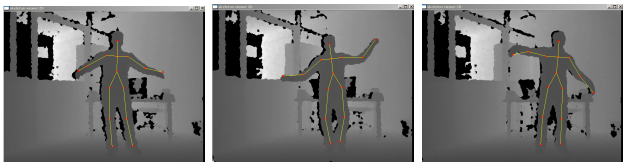

Figure 2: Fused depth and skeleton (using calculated joint rotations) information images

3.2 Expressivity computational formalization

Body pose P is formally defined as a sequence, of T frames $i \in [1, T]$, consisting of:

$$P = [\vec{l}, \vec{r}, S, D, F, J] \tag{1}$$

where $\vec{l}$ *and* $\vec{r}$ are the 3D coordinates of the left and right hand, S the silhouette binary image, D the depth image map, F the face information, describing position, size and depth, and J the skeleton joint rotations for left/right arm J_l/J_r for shoulder, elbow, hip and knee. Given the above definition of pose, expressivity features are formulated using different approaches, based on a) Silhouette b) Limbs and c) Joints.

Although silhouette is usually used in full body expressivity analysis, as already discussed in section 2, limb based expressivity formalization presents interest, since it has been used before in half-body, desktop interaction context. One could argue that limb based analysis is a subcase of silhouette based one, but on the other hand the corresponding computer vision techniques used to extract them are different. More specifically, Silhouette extraction is a trivial task for fixed background [8] and feasible when depth information is available. Limb, actually limb's end effectors, detection and tracking, especially for the case of skin colored hands could be applied to wider range of applications and interaction contexts. Finally, joint expressivity formalization is quite innovative since robustly extracting relative features is an extremely challenging task and researchers opted to simpler and more robust approaches. This feature can be calculated with the following equations based on different modeling approaches:

Overall activation is considered as the quantity of movement during a dialogic discourse.

1. For a given time window of w frames define fading silhouette motion volumes $FSMV$ adding a degrading weight depending on time and volume:

$$FSMV_t = ((\sum_{i=1}^{w} \frac{w-i}{w} S_{t-i}) - S_t)(|D_t - D_{t-w}|) \tag{2}$$

The general equation of silhouette based overall activation would be:

$$OA = \frac{volume\ of\ motion}{volume\ of\ silhouette} \tag{3}$$

or better defined as:

$$\frac{FSMV}{SV} \tag{4}$$

$$FSMV = \sum_{t=1}^{T} FSMV_t \tag{5}$$

$$SV = \sum_{i=1}^{T} S_i D_i) \tag{6}$$

SV being a normalization factor for distance and size invariant results.

2. limb based OA defined as:

$$OA = \sum_{i=1}^{T} \left|(r\vec{/}l)h_i - (r\vec{/}l)h_{i-1}\right| + \left|(r\vec{/}l)f_i - (r\vec{/}l)f_{i-1}\right| \tag{7}$$

3. weighted sum of joints rotations derivative:

$$OA = W_1(J'_{(r/l)s} + J'_{(r/l)h}) + W_2(J'_{(r/l)e} + J'_{(r/l)k}) \tag{8}$$

$$s = shoulder\ e = elbow\ h = hip\ k = knee \tag{9}$$

Spatial extent is expressed with the expansion or the condensation of the used space in front of the user (gesturing space). Let SE_0 be the spatial extent (according to each definition) of the neutral/calibration position. Then SE can be calculated as follows:

1. 2D silhouette based:
   (a) max and median of area of polygon consisting of left hand, head, right hand, right foot, left foot normalised by SE_0
   (b) max and median of sum of diagonals of Quadrilateral consisting of right hand/left foot and left hand/right foot normalised by SE_0
2. limb based is already included in silhouette based
3. joint based does not make sense since rotation is independent of spatial extent

Fluidity differentiates smooth / elegant from the sudden / abrupt gestures. This concept attempts to denote the continuity between movements. It is formally defined as the variance of the OA as described previously:

$$FL = Var(\frac{FSMV}{SV_0}) \tag{10}$$

Please note that the quantity FL corresponds to is reversely proportional to the notion of fluidity. Thus, a motion with high value of FL expressive parameter demonstrates low fluidity and consequently is categorized as a sudden/abrupt movement. Inverting the definition of fluidity is not a trivial process since the upper and lower bound of the measure are not a priori known.

Temporal expressivity parameter denotes the speed of hand movement during a gesture and dissociates fast from slow gestures:

$$TE = \frac{mean(FSMV)}{SV_0} \tag{11}$$

$$SV_0 = S_0 D_0 \tag{12}$$

again SV_0 as SE_0 is a normalizing factor

Power is associated qualitatively with the first derivative of which refers to acceleration:

$$PO = FSMV^{'} \tag{13}$$

Additionally, variations of the ideas behind expressivity features extraction for each expressivity aspect are formalized in order to conclude on which of the variations best describes the feature in question. This validation was performed through the LDA and related coefficients analysis described in Section 4. More specifically, for OA silhouette based formalization $FMSV$ was defined additionally with a more gradual approach using $S_{t-i}(D_{t-i} - D_{t-w}) - S_t(D_t - D_{t-w})$ and a cumulative and more rough one using $(S_t - S_{t-1})(|D_t - D_{t-1}|)$. As also described above, Spatial Extent for 2D silhouette is defined dually with the joints polygon area and sum of Quadrilateral diagonals. Respectively, expressivity features that are calculated as derivatives or variance of another one are also multiply defined.

4. EXPRESSIVITY ANALYSIS

As a first phase in analysing the five expressivity features, we applied Linear Discriminant Analysis (LDA) on the whole dataset, as well as to the parameters of each feature subset, in order to assess the discrimination between two distinct expressivity levels: high and low. LDA uses the class information for calculating the projection that maximises the distance between the two levels. Although LDA requires certain conditions to be met at the dataset in order to find the best possible linear classifier (normal multivariant distribution of the variables and homoscedasticity), it can still be a suitable method for projection directions that maximise the separation between elements of different classes and this purely geometric interpretation is not affected by hypotheses on the distribution of data [12]. LDA allows for the interpretation of the linear coefficients as predictive strengths of the corresponding variables [9]. All the programming of the workflow was implemented in R [21]. At the pre-processing stage the data where inserted in an R dataframe and transformed (centered and scaled) using the function preProcess from the R package caret (classification and regression training). LDA was performed by the LDA function from the MASS R package [25]. The results from the analysis on the full expressivity dataset are depicted at Figure 3. Along

Table 1: SVD values in decreasing order

Expressivity Feature	SVD Value
Spatial Extent	70.46
Overall activation	49.36
Fluidity	44.74
Temporal	34.87
Power	29.91

the Y axis label the Singular Value Decomposition (SVD) value is shown, which gives the ratio of the between- and within-group standard deviations on the linear discriminant variables. Its square is the canonical F-statistic.

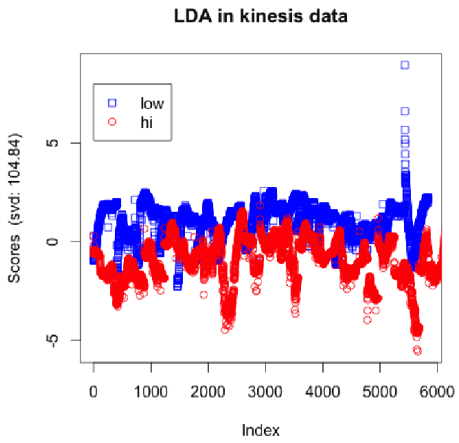

Figure 3: LDA Analysis on the full dataset

Aiming at the identification of the best variables at each of the five expressivity features, we applied again LDA on each expressivity subset. At Figures 4 and 5 are the results for the Spatial Extent (SE) expressivity feature, which offered the best SVD value (70.46) amongst the expressivity features. A list of the SVD values for all the five features is shown at Table 1. Figure 5 showcases the comparative importance of the variables which are relative to the SE feature, according to their absolute coefficient value resulting from the LDA.

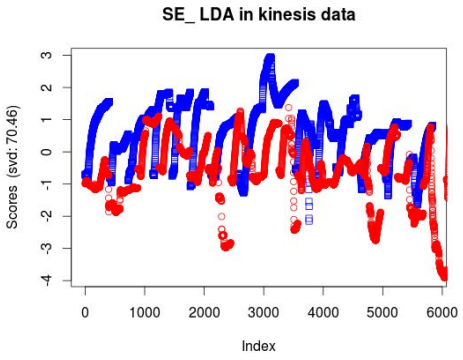

Figure 4: LDA on the Spatial Extent (SE) expressivity feature

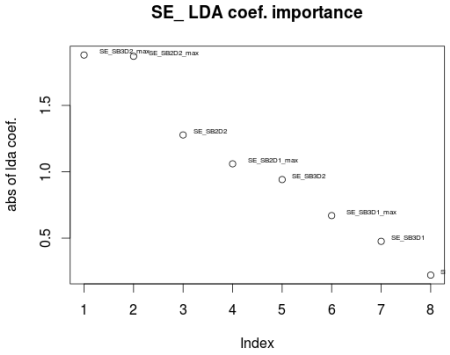

Figure 5: Variables' importance of the SE expressivity feature

Table 2: Overall accuracy on predicting the activation level (High or Low)

Expressivity Feature	Overall Accuracy
Spatial Extent	0.7486
Overall activation	0.5971
Fluidity	0.6607
Temporal	0.6131
Power	0.5692

The LDA function has two working modes. One having the parameter CV=False (the default), allowing then to obtain an object that includes discriminant scores, and the other with CV=True, were predictions of class memberships are derived from leave-one-out (LOO) cross-validation. Running LDA in LOO CV mode, we obtained the overall accuracies in predicting the activation class of the objects by using only the specific expressivity feature. The results are depicted at Table 2.

We also illustrate the LDA and coefficients analysis for the rest of the expressivity features in Figures 6 and 7 respectively.

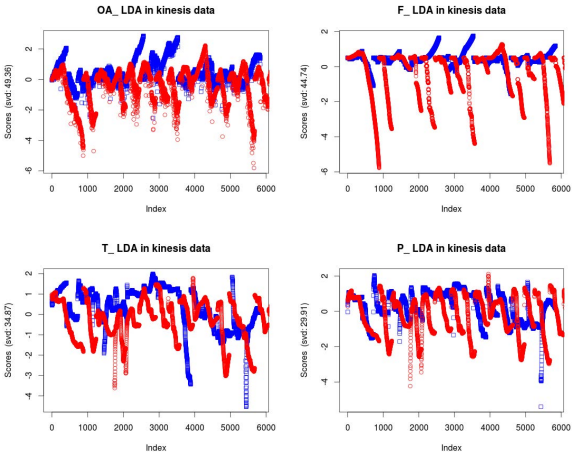

Figure 6: LDA on the expressivity features

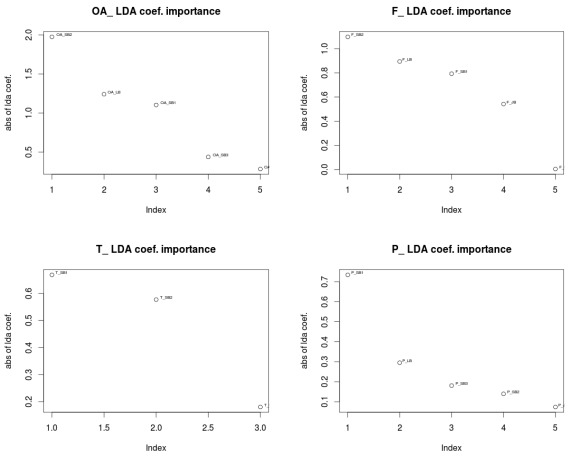

Figure 7: Variables' importance of expressivity features

5. CONCLUSIONS AND FUTURE WORK

Automatic affect recognition, either unimodal or in a multimodal setting, of body expressivity has been recently quite active. On the other hand it is questionable if the basis of such approaches, the modeling aspect of expressivity has been sufficiently studied. This aspect is included in the recognition process but the modeling effectiveness is not investigated per se. The effectiveness of the expressivity modeling is measured usually through the respective recognition accuracy. Other aspects, such as the corpus construction, stimuli, feature extraction and processing, classification and fusion issues, are also included in the recognition process making it almost impossible to derive conclusions on the modeling itself. Research presented here focuses solely on the modeling aspect of body expressivity and its effectiveness.

Behavior, including non verbal, expressiveness is an integral part of the communication process and current article presents research work on the computational formalization and analysis of full body 3D expressivity in Natural (bodily) Interaction. Five expressivity parameters are computationally formalized, using different approaches which are in turn evaluated in terms of their effectiveness in modeling the expressivity aspect in question. Current research work confirms that the silhouette based, especially using the Fading Silhouette Motion Volumes (FMSV), approach is the most effective in modeling body expressivity. It's superiority, related to limb and joint based approaches, is proved using LDA analysis and respective coefficients for all the parameters investigated in the presented study (Overall activation, Spatial extent, Temporal, Fluidity and Power). Regarding recognition which, as discussed is not the main focus of the research work presented here, Spatial Extent was more accurately recognized. Intuitively, this is expected since it is also more straightforward and perceivable during human to human communication.

Regarding ongoing and future work, we are working on, further validating on naturalistic user behavior both during

different, but always natural, interaction context. Finally, appropriate ways, and hopefully an integrated architecture, to incorporate extracted expressivity features will constitute a challenging future research direction.

6. ACKNOWLEDGMENTS

Current research work has been partially funded by the Greek Ministry of Education, Education and Lifelong Learning program, project code MIS 380328 "Cross-disciplinary research in Affective Computing for physiology and activity recognition in assistive enviroments - STHENOS".

7. REFERENCES

[1] D. Bernhardt and P. Robinson. Detecting affect from non-stylised body motions. In *Proc. of Affective Computing and Intelligent Interaction (ACII 2007)*, pages 59–70. Springer, 2007.

[2] N. Bianchi-berthouze and A. Kleinsmith. A categorical approach to affective gesture recognition. *Connection science*, 15(4):259–269, 2003.

[3] A. Camurri, B. Mazzarino, M. Ricchetti, R. Timmers, and G. Volpe. Multimodal analysis of expressive gesture in music and dance performances. *Gesture-based communication in human-computer interaction*, pages 357–358, 2004.

[4] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, and K. Karpouzis. Multimodal emotion recognition from expressive faces, body gestures and speech. *Artificial Intelligence and Innovations 2007: From Theory to Applications*, pages 375–388, 2007.

[5] G. Caridakis, A. Raouzaiou, K. Karpouzis, and S. Kollias. Synthesizing gesture expressivity based on real sequences. In *Workshop on multimodal corpora: from multimodal behaviour theories to usable models, LREC 2006 Conference, Genoa, Italy*, pages 24–26. Citeseer, 2006.

[6] G. Castellano, M. Mancini, C. Peters, and P. McOwan. Expressive copying behavior for social agents: A perceptual analysis. *IEEE Transactions on Systems, Man and Cybernetics, Part A - Systems and Humans*, 2011.

[7] G. Castellano, M. Mortillaro, A. Camurri, G. Volpe, and K. Scherer. Automated analysis of body movement in emotionally expressive piano performances. *Music Perception*, pages 103–119, 2008.

[8] A. Christodoulidis, K. Delibasis, and I. Maglogiannis. Near real-time human silhouette and movement detection in indoor environments using fixed cameras. In *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments*, page 1. ACM, 2012.

[9] D. Garrett, D. A. Peterson, C. W. Anderson, and M. H. Thaut. Comparison of linear, nonlinear, and feature selection methods for eeg signal classification. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 11(2):141–144, 2003.

[10] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer. Towards a minimal representation of affective gestures. *Affective Computing, IEEE Transactions on*, PP(99):1, 2011.

[11] H. Gunes and M. Piccardi. Automatic temporal segment detection and affect recognition from face and body display. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(1):64–84, 2009.

[12] L. J. Hargrove, E. J. Scheme, K. B. Englehart, and B. S. Hudgins. Multiple binary classifications via linear discriminant analysis for improved controllability of a powered prosthesis. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 18(1):49–57, 2010.

[13] A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis, and P. Driessen. Gesture-based affective computing on motion capture data. *Affective Computing and Intelligent Interaction*, pages 1–7, 2005.

[14] A. Kleinsmith and N. Bianchi-Berthouze. Recognizing affective dimensions from body posture. *Affective Computing and Intelligent Interaction*, pages 48–58, 2007.

[15] A. Kleinsmith and N. Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 2012.

[16] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed. Automatic recognition of non-acted affective postures. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 41(4):1027–1038, 2011.

[17] A. Mehrabian and J. T. Friar. Encoding of attitude by a seated communicator via posture and position cues. *Journal of Consulting and Clinical Psychology*, 33(3):330, 1969.

[18] A. Pentland. Smart rooms. *Scientific American*, 274(4):54–62, 1996.

[19] P. Petta, C. Pelachaud, and R. Cowie, editors. *Emotion-Oriented Systems, The Humaine Handbook*. Springer, Series: Cognitive Technologies, February, 2011.

[20] K. R. Scherer, T. Banziger, and E. Roesch, editors. *A Blueprint for Affective Computing, A sourcebook and manual*. Oxford University Press, November, 2010.

[21] R. C. Team et al. R: A language and environment for statistical computing. *R Foundation Statistical Computing*, 2008.

[22] M. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 38–45. ACM, 2007.

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

[24] C. Van Heijnsbergen, H. Meeren, J. Grezes, and B. de Gelder. Rapid detection of fear in body expressions, an erp study. *Brain research*, 1186:233–241, 2007.

[25] W. N. Venables and B. D. Ripley. *Modern applied statistics with S*. Springer, 2002.

[26] H. Wallbott. Bodily expression of emotion. *European journal of social psychology*, 28(6):879–896, 1998.

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