Towards Real-Time Continuous Emotion Recognition from Body Movements

Weiyi Wang[1], Valentin Enescu[1], and Hichem Sahli[1,2]

[1] Dept. Electronics & Informatics (ETRO), Vrije Universiteit Brussel (VUB), Belgium
[2] Interuniversity Microelectronics Center (IMEC), Leuven, Belgium
{wwang,venescu,hsahli}@etro.vub.ac.be

Abstract. Social psychological research indicates that bodily expressions convey important affective information, although this modality is relatively neglected in the literature as compared to facial expressions and speech. In this paper we propose a real-time system that continuously recognizes emotions from body movements data streams. Low-level 3D postural features and high-level kinematic and geometrical features are through summarization (statistical values) or aggregation (feature patches), fed to a random forests classifier. In a first stage, the MoCap UCLIC affective gesture database has been used for training the classifier, which led to an overall recognition rate of 78% using a 10-fold cross-validation (leave-one-out). Subsequently, the trained classifier was tested with different subjects using continuous Kinect data. A performance of 72% was reached in real-time, which proves the efficiency and effectiveness of the proposed system.

Keywords: emotion recognition, real-time, bodily expressions.

1 Introduction

In real life, humans interact with each other and the environment with multi cues including facial expressions, gestures and voice. These cues together convey not only part of the content in the interactions, but also the information of affective states. In the past decades, a great effort has been made to automatically recognize emotions and affective states through different cues or their combinations. According to De Gelder [11], 95% of the research was conducted with face stimuli, the majority of the remaining 5% with audio, followed by bodily expressions. Supported by psychological studies, bodily expressions recently started to attract increasingly more research interest [11][14][30][2].

In this paper we propose an approach for real-time continuous recognition of expressive gestures from body movements. In the proposed metodology, both 3-dimensional low-level postural features and high-level (kinematic and geometrical) features are estimated, from which statistical cues or temporal patches, representing the dynamics of body movements, are calculated. The UCLIC affective gesture database [22] has been employed to train a random forests classifier,

A.A. Salah et al. (Eds.) : HBU 2013, LNCS 8212, pp. 235–245, 2013.

then both sequence-based (one single emotional gesture sequence) and continuous input tests were conducted to validate the recognition capabilities of the proposed approach. The obtained recognition rate, around 78%, outperformed the best result of apex posture-based recognition on the UCLIC database. Finally the trained random forests classifier has been used on continuous Kinect sensor data to provide real-time recognition of emotional gestures.

The rest of this paper is organized as follows: Section 2 introduces the research background, state-of-the-art and the trends that are related to the scope of this paper. Section 3 gives details of our proposed system, including feature extraction, recognition framework, as well as the introduction of the UCLIC database we used to test our system, followed by Section 4 of experimental results and comparison. Finally, in Section 5, we discuss the proposed approach.

2 Background

In [16], a survey of recent representative research work on automatic emotion analysis found the bodily information as an important modality to predict affective states. Berthouze and Kleinsmith [7][22] chose the most representative static postures that were acted by participants to analyze both categorical emotions and affective dimensions. A motion capture (MoCap) device was used to capture the body movements and low-level geometric features were calculated. They continued their work to spontaneous affects analysis in a body-controlled game scenario with a similar representation of selected postures [21], while the body's dynamics was not considered. Metallinou et. al in [25] described an emotional trends tracking system in a dyadic scenario performed by professional actors. The body movements were represented by both individual features and relative ones between two persons in dyadic interactions, in three dimensionalities. Gaussian Mixture Models and Long-Short Term Memory regression were used to continuously track the changes of three affective dimensions: *Activation*, *Valence* and *Dominance*. Convincing tracking results were reported for *Activation* and *Dominance*, while *Valence* was less tracked by the proposed feature sets. Castellano et. al [10][9] conducted categorical emotion recognition based on pre-segmented sequences. High-level features, such as quantity of motion, contraction index, velocity and acceleration, were calculated on a per-frame basis for each sequence. Then sequence-based statistical indicators such as initial/final slope, centroid of energy, etc... were computed for each feature sequence as inputs for the used machine learning algorithms. Due to the nature of proposed body movement descriptors, it is not easy to perform the recognition in real-time. Indeed the approach requires the complete expressive gesture sequence before recognition. Glowinski et. al [13] used similar body motion descriptors to investigate a minimal representation of expressive gestures. In their approach, Principal Component Analysis (PCA) revealed four components (roughly indicating motion activity, temporal and spatial excursion of movement, spatial extent and postural symmetry and motion jerkiness respectively) which could represent the expressions efficiently. Piana et. al [27] proposed a set of features to differentiate emotions, which includes both 3D movement information from RGB-D cameras or MoCap devices, and 2D features from the image segmentation of the shape of the users. Sparse Coding and Dictionary Learning were employed to combine the feature vectors in an unsupervised manner, followed by the normal SVM for classification. Similar feature set was used in [29] as a component of the social active music systems to analyze listener's affective states and then to control the music processing and playing. Some other studies were also conducted either by using bodily information as a complementary cue to facial expressions and speech [3][26], or in a constrained context [5][24]. Generally speaking, emotion recognition from bodily information is still an unsettled research question due to the differences in individual expressions, body movement sensors, as well as affective model definition. Wallbott [30] conducted comprehensive experiments and tried to reveal certain connections between the emotion categories and specific body movement patterns and postures. However, unlike Facial Action Coding System (FACS and its variants EMFACS and FACSAID for emotion analysis) [30] that is widely used in facial expression studies, there still is no agreement on affective bodily expressions within the research community.

3 Methodologies

3.1 Body Representation

The way to acquire and represent human bodily information influences directly the approaches to model the body movements and postures, and hence the affective analysis. Color cameras were widely used either to track the positions of certain parts of body (mostly hands and head) [13][26] or to abstract high-level body motions [10][17]. In contrast, marker-based motion capture devices give the positions or angular values of body joints, which are considered to be more precise and easier to process. However, participants are required to wear the marker suits. With the advent of increasingly mature RGB-D sensors (e.g. Microsoft Kinect sensor), depth information and body skeleton could nowadays easily be extracted from these devices and provide similar accuracy to the motion capture ones.

There exist several databases consider bodily stimuli as one of the modalities for affective analysis, such as the GEMEP (Geneva Multimodal Emotion Portrayals) corpus [4] and UCLIC affective gesture database [22]. In this work, we use the UCLIC database to train and evaluate our approach. Moreover, the trained classifier has been applied to continuous Kinect data. To build the UCLIC database, 13 candidates, from different cultural regions, were asked to act certain emotions (*Anger*, *Fear*, *Happiness*, and *Sadness*) with body languages [22]. A MoCap device with 32 markers attached on different segments of the body was used to obtain the joint positions. Although affective postures were the main research target of the UCLIC database, the acquired continuous frames make it also valuable for dynamic analysis. Fig. 1 depicts few frames of a sample sequence of 'fear' emotion.

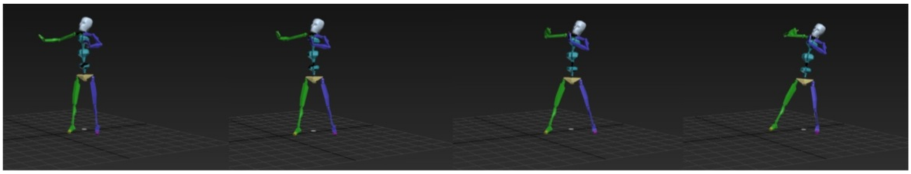

Fig. 1. An example of fear emotion at four time points from the UCLIC database

3.2 Feature Extraction

Inspired by the results of psychological experiments conducted by Wallbott [30], both low-level postural features and high-level kinematic and geometrical features are extracted.

Postural Features: Some specific postures are generally bound to certain emotional expressions according to social psychological research. [30] has shown that upper body, especially arms and head, plays the most important role for expressions. Based on this observation, we calculate the spatial distances among hands, elbows and shoulders in each of the three dimensions, as well as the angles of two elbows. Moreover, we calculate the distance between feet, the orientation of feet and the orientation of shoulders. All these lead to 28 postural features in total. Note that, in contrast to the approaches proposed in [6][20], we assume that expressive postures are evolving both spatially and temporally rather than being static, which had been investigated and supported in previous work indicating that bodily expressions could also be segmented to *onset*, *apex*, *offset* as facial expressions [15]. Therefore, the 28 low-level postural features are calculated on a per-frame basis.

High-Level Features: These features are designed to represent the abstract characteristics of bodily expressions, such as *movement activity*, *movement power*, *body spatial extension* and *body bending*.

Body movement activity and power: Human motions could be thought of as being composed of different physical segments and each segment can move independently and exhibit an independent degree of activity [1]. As illustrated in Fig. 2, these body segments have a hierarchical structure, which allows estimating the body movement activity and power composed of three parameters: *force*, *kinetic energy* and *momentum* calculated hierarchically (from the bottom to the top) [19]:

$$segment\_Force = segment\_Mass \times segment\_Acceleration \tag{1}$$

$$segment\_KineticEnergy = 0.5 \times segment\_Mass \times segment\_Velocity^2 \tag{2}$$

$$segment\_Momentum = segment\_Mass \times segment\_Velocity \tag{3}$$

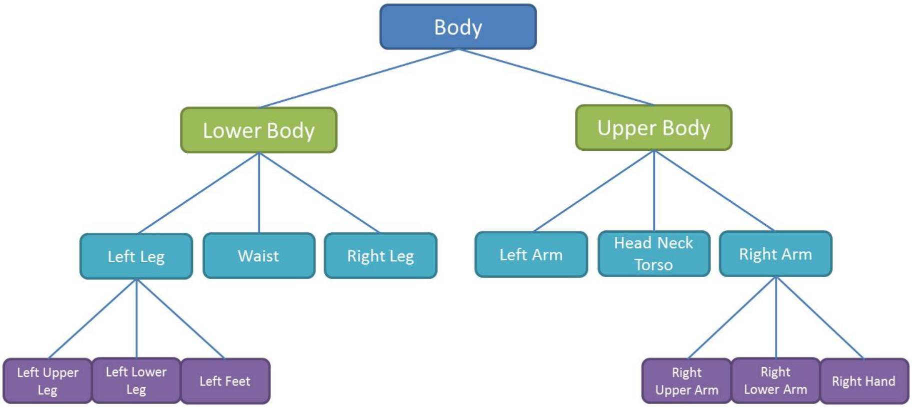

Fig. 2. Hierarchical representation of human body

where $segment\_Mass$ are estimated according to ergonomic definitions, which relate the body weight to the masses of different body segments (refer to [23] for details).

Body Spatial Extension: From the positions of body joints, we extract the body bounding box at each frame. Then three spatial extent indexes are calculated in x, y, z axis separately:

Symmetry: Spatial symmetric indexes are considered for the two hands in three coordinates. A relative coordinate system is used to make sure these indexes are position independent.

Body Bending: This feature is designed to catch full body bending forward or backward (mostly happens in fear and disgust emotions) movements.

Thus, ten high-level kinematic and geometrical features are extracted, together with the postural feature a frame-based38-dimensional feature vector is obtained. Note that, most of the above defined features are body size dependent, which means the height of the body will affect the feature values. This adverse effect could be practically eliminated by normalizing the feature values with the body size.

Statistical Features: In order to capture the gist of the body motion, we further calculate 6 statistical cues, *mean*, *standard deviation*, *min*, *max*, *kurtosis* and *skewness*, within a fixed-size time.

Temporal Patches: We also create temporal feature patches that concatenates, for the considered time window, the 38 per-frame features. Such temporal feature patch could catch not only the current states of the features, but also how they evolve within the time window.

3.3 Recognition

We are dealing with a multi-class, multi-variable time series classification problem. According to a comprehensive survey on different machine learning approaches [8], the *Random Forests* classifier achieves the best recognition performance for high dimensionality problems. Moreover, some unique characteristics of Random Forests, such as its low computational cost, its capability to handle feature redundancy as well as avoiding over-fitting problems, render it suitable for our research. Additionally, Random Forests could also be used for regression, which is of value for our future study.

Basically, a *Random Forests* consists of many single decision trees, each tree is grown individually without pruning. Samples used to train the individual trees are randomly selected from the whole set, as well as the variables used in each node. These introduced randomness efficiently de-correlates the trees and guarantees the reliability of the results.

4 Experiments

Baseline Classifiaction: We firstly estimated the 27 postural features proposed in [20], for which the UCLIC database was originally designed, for the apex frame of each affective gesture motion in the database, and used the WEKA toolbox [18] to run a 10-fold cross validation with leave-one-out strategy using different classifiers. The best results were achieved by the Random Forests classifier with an overall recognition rate of 75.41% (138 correctly classified out of 183). This result has been used as the baseline of our experiments.

Sequence Testing: Pre-segmented (single expression) gesture sequences from the UCLIC database were used to test the proposed features and classifier. We borrowed the idea from *Random Forests* to randomly sample a certain percentage of frames (e.g. 50 percent) within a testing sequence with replacement. Features with statistical cues and/or temporal patches were calculated for these frames. Then, we summed up the classification results of these frames and the mostly occurred label has been assigned to the considered sequence. Two evaluations were conducted:

Feature set comparison: we compared the contribution of low-level and high-level features, respectively. We conducted 10-fold cross validation with leave-one-out strategy for the 183 single expression gesture sequences. 100 trees were grown without pruning in the forest during training.

The recognition performance is presented in Table 1. When using only postural features, we could reach a higher classification accuracy as compared to high-level features, and the high-level ones are not sufficient to differentiate expressive gestures. This conclusion was also supported by a χ^2 *ranking* on all these features. It could be, on the other hand, partly explained by the *acted nature* of the UCLIC database.

Table 1. Recognition accuracy when using postural features and high-level features separately. The last row gives the baseline recognition rate with apex frames.

	Postural Features				High-level Features			
	Fear	Happy	Angry	Sad	Fear	Happy	Angry	Sad
Fear	28	10	8	3	21	15	8	5
Happy	6	36	4	1	11	32	1	3
Angry	5	3	28	5	4	14	16	7
Sad	1	5	6	34	4	3	3	36
Rate	68.85%				57.38%			
Baseline	75.41%							

Statistical cues and temporal patches comparison: we then analyzed the recognition rate when adding either statistical cues or temporal feature patches. The window size introduces an extra parameter to the system. A larger value would comsume too much calculation power, and on the other side if this size is too small, it would not be sufficient to capture the dynamics. Analyzing the UCLIC sequences, we observed that for most gestures, the motion lasts for $0.5s$ to $2s$. By taking into account the affordable latency (i.e., how fast we have to provide a classification result), we set the size of the time window to 30 frames ($0.5s$). Table 2 summarizes the the classification results.

Table 2. Recognition accuracy using the frame-based feature set, plus either statistical cues or temporal feature patches. The last row gives the baseline recognition rate with apex frames.

	Frame-based Features				+Statistical Cues				+Temporal Patches			
	Fear	Happy	Angry	Sad	Fear	Happy	Angry	Sad	Fear	Happy	Angry	Sad
Fear	31	11	4	3	34	5	5	5	35	7	3	4
Happy	5	36	4	2	4	38	2	3	4	39	3	1
Angry	4	5	29	3	4	3	30	4	3	4	30	4
Sad	2	3	4	37	3	2	3	38	2	2	3	39
Rate	72.68%				76.50%				78.14%			
Baseline	75.41%											

We can see that adding statistical cues or concatenating previous frames on the original frame-based feature set (temporal feature patches) could, to a certain extent, improve the recognition performance. Note that, although the feature set dimensionality would largely increase (6 times for statistical cues and 30 times for temporal patches), the *Random Forests* classifier delivers results in real-time due to its efficient tree-searching nature. Another conclusion from this experiment is that statistical cues and temporal patches reach very similar recognition rates. This is because statistical cues could be considered as abstract descriptions of the temporal evolvement of features, which convey similar information with temporal feature patches.

In this experiment, we also proved our proposed features and recognition framework outperform the baseline result. Manual expressive apex frame selection is not necessary in our system, thus it is feasible to be run with continuous, un-segmented data streams.

Continuous Recognition: Owing to the computational efficiency of *Random Forests*, the continuous prediction result could be given for each incoming frame in soft real-time (statistical cues or the temporal patches will introduce latency depending on the window size). We used a Kinect sensor to capture the body skeleton and feed the classifier trained with the UCLIC database sequences. It has to be noted that there are slight differences between the definition of joints in Kinect SDK and the ones defined in the UCLIC database, while during our experiments we observed it is not necessary to match them perfectly since the system has a certain tolerance. We asked six participants to *act* randomly the four emotions contained in the UCLIC database as naturally as possible (some examples of each emotion category in the UCLIC database were presented to provide some hints. However, the participants could freely act as their own understanding). To avoid nervousness and "over-acted" artifacts, the participants rehearsed as many times as they needed before the real experiments. In total 82 expressive segments, in which seven were not usable due to the errors in skeleton extraction, were collected. The recognition system was running in real-time giving the feedback of *neutral state* or recognized emotions. In the UCLIC database, the neutral gesture was defined to be a *T-Pose*, and could be simply detected by measuring both the body energy and body spatial extension. More sophisticated approaches could be introduced, but it is out of the scope of this paper. A simple window-based smoothing filter was applied to remove sharp errors. For the 75 expressive segments, 54 were correctly recognized, giving the recognition rate of 72%. Fig. 3 depicts the recognition results for a long sequence containing a random repetition of the four emotion segments.

From this experiment, one can see that even the subjects did not appear in the training data, the expressive gestures could still be well classified.

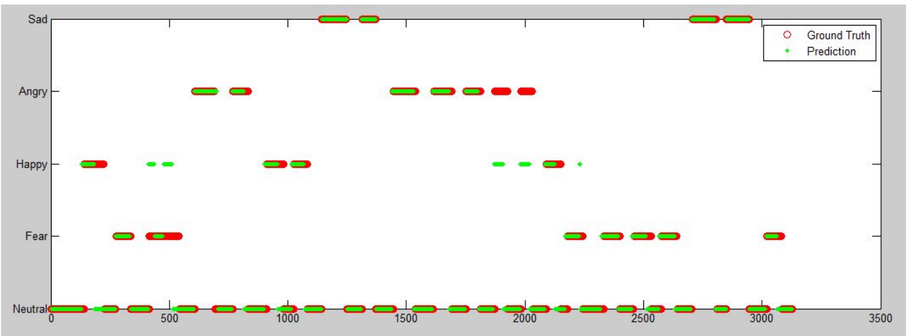

Fig. 3. Continuous recognition result of a sequence recorded by Kinect sensor. Red bars are labels and green bars are predicted results. Overlaps represent correctly recognized frames.

5 Discussion and Future Work

There are three important criteria of creating automatic affect analyzers: *correctness* (to match human observers labeling), *robustness* (not too sensitive to changing conditions) and *efficiency* (real-time processing and responsiveness) [16]. In this paper we proposed a real-time expressive gesture recognition system to meet those criteria. The system continuously calculates postural and high-level bodily features, as well as statistical cues. A random forests classifier is applied to make efficient predictions. The UCLIC affective gesture database was employed to train and test our model. Kinect sensor recordings were also used for testing in both offline and online recognition mode. Overall results are convincing and outperform the affective posture recognition conducted on the same database, which requires manual frame selection and is not suitable for continuous analysis. Obviously, when using Kinect skeleton streams as the inputs, the quality of the skeleton extraction directly influences the recognition performance. Moreover, due to the lack of available public 3D affective gesture corpus, our experiments were constrained to *acted* expressions and limited to four emotion categories. In the future, we will conduct other experiments that aim at eliciting and recording *spontaneous* affective expressions in multi-modalities. To have better body tracking quality, these recordings will be made using multiple RGB-D cameras (e.g. 2 Kinects placed with a certain angle), and the tracking will be implemented following the ideas from Girshick et. al [12] and Taylor et. al [28]. This system could be naturally extended to dimensional affective analysis (e.g. *Valence, Arousal, Dominance etc.*) using random forests regression, this being another part of our future work. Last but not least, we are planning to employ the approaches and system proposed in this paper to study the emotional expressions and perception of children in child-robot interaction scenarios, within the EU FP-7 Aliz-E project (www.aliz-e.org).

Acknowledgements The research work reported in this paper was supported by the EU FP7 project ALIZ-E grant 248116, and by the CSC-VUB scholarship grant. We also would like to thank the referees for their comments and suggestions, which helped improving the paper.

References

1. Aggarwal, J., Cai, Q.: Human motion analysis: a review. In: Proc. of Nonrigid and Articulated Motion Workshop, pp. 90–102 (1997)
2. Alaerts, K., Nackaerts, E., Meyns, P., Swinnen, S.P., Wenderoth, N.: Action and emotion recognition from point light displays: an investigation of gender differences. PLoS One 6(6), e20989 (2011)
3. Baltrusaitis, T., McDuff, D., Banda, N., Mahmoud, M., el Kaliouby, R., Robinson, P., Picard, R.: Real-time inference of mental states from facial expressions and upper body gestures. In: Proceedings of 2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011), pp. 909–914 (2011)

4. Bänziger, T., Mortillaro, M., Scherer, K.R.: Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. Emotion (Washington, D.C.) 12(5), 1161–1179 (2012), http://www.ncbi.nlm.nih.gov/pubmed/22081890
5. Bernhardt, D.: Emotion inference from human body motion. Tech. Rep. 787, Computer Laboratory, University of Cambridge, Cambridge (2010)
6. Bianchi-Berthouze, N., Cairns, P., Cox, A.: On posture as a modality for expressing and recognizing emotions. In: Emotion in HCI: Joint Procs. of the 2005, 2006, and 2007 Intl. Workshops, pp. 74–80. Citeseer (2008)
7. Bianchi-Berthouze, N., Kleinsmith, A.: A categorical approach to affective gesture recognition. Connection Science 15(4), 259–269 (2003)
8. Caruana, R., Karampatziakis, N., Yessenalina, A.: An empirical evaluation of supervised learning in high dimensions. In: Proceedings of the 25th International Conference on Machine Learning, ICML 2008, pp. 96–103 (2008), http://portal.acm.org/citation.cfm?doid=1390156.1390169
9. Castellano, G., Kessous, L., Caridakis, G.: Multimodal emotion recognition from expressive faces, body gestures and speech. Affective computing and intelligent interaction 4738, 71–82 (2007)
10. Castellano, G., Villalba, S., Camurri, A.: Recognising human emotions from body movement and gesture dynamics. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 71–82. Springer, Heidelberg (2007)
11. de Gelder, B.: Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. Philosophical Tran. of the Royal Society B: Biological Sciences 364, 3475–3484 (2009)
12. Girshick, R., Shotton, J., Kohli, P., Criminisi, A., Fitzgibbon, A.: Efficient regression of general-activity human poses from depth images. In: 2011 International Conference on Computer Vision, pp. 415–422 (November 2011), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6126270
13. Glowinski, D., Dael, N., Camurri, A., Volpe, G., Mortillaro, M., Scherer, K.R.: Toward a Minimal Representation of Affective Gestures. IEEE Transactions on Affective Computing 2(2), 106–118 (2011)
14. Gross, M.M., Crane, E.A., Fredrickson, B.L.: Methodology for Assessing Bodily Expression of Emotion. Journal of Nonverbal Behavior 34(4), 223–248 (2010), http://www.springerlink.com/index/10.1007/s10919-010-0094-x
15. Gunes, H., Piccardi, M.: Automatic temporal segment detection and affect recognition from face and body display. IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics: A Publication of the IEEE Systems, Man, and Cybernetics Society 39(1), 64–84 (2009)
16. Gunes, H., Schuller, B.: Categorical and dimensional affect analysis in continuous input: Current trends and future directions. Image and Vision Computing (July 2012), http://linkinghub.elsevier.com/retrieve/pii/S0262885612001084
17. Gunes, H., Shan, C., Chen, S., Tian, Y.: Bodily Expression for Automatic Affect Recognition. In: Advances in Emotion Recognition (2012)
18. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. Newsl. 11(1), 10–18 (2009)
19. Kahol, K., Tripathi, P., Panchanathan, S.: Gesture Segmentation in Complex Motion Sequences. In: Proc. of International Conference on Image Processing (ICIP 2003) (2003)

20. Kleinsmith, A., Bianchi-Berthouze, N.: Recognizing affective dimensions from body posture. In: Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction, pp. 45–58 (2007)
21. Kleinsmith, A., Bianchi-Berthouze, N.: Automatic recognition of non- acted affective postures: A video game scenario. IEEE Trans. on Systems, Man, and Cybernetics Part B 41(4), 1027–1038 (2011)
22. Kleinsmith, A., De Silva, P.R., Bianchi-Berthouze, N.: Cross-Cultural Differences in Recognizing Affect from Body Posture. Iteracting with Computers 18(6), 1371–1389 (2006)
23. Kroemer, K.H.E., Kroemer, H.B., Kroemer-Elbert, K.E.: Ergonomics: how to design for ease and efficiency. Prentice Hall (1994)
24. Leite, I., Castellano, G., Pereira, A., Martinho, C., Paiva, A.: Modelling Empathic Behaviour in a Robotic Game Companion for Children. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI 2012) (2012)
25. Metallinou, A., Katsamanis, A., Narayanan, S.: Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. In: Image and Vision Computing (September 2012), `http://linkinghub.elsevier.com/retrieve/pii/S0262885612001710`
26. Nicolaou, M.A., Member, S., Gunes, H., Pantic, M., Member, S.: Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence Arousal Space. IEEE Trans. on Affective Computing, 1–15 (2011)
27. Piana, S., Staglianò, A., Camurri, A., Odone, F.: A set of Full-Body Movement Features for Emotion Recognition to Help Children affected by Autism Spectrum Condition. In: IDGEI International Workshop (2013)
28. Taylor, J., Shotton, J., Sharp, T., Fitzgibbon, A.: The Vitruvian Manifold: Inferring Dense Correspondences for One-Shot Human Pose Estimation Optimization of Model Parameters, pp. 103–110 (2012)
29. Volpe, G., Camurri, A.: A system for embodied social active listening to sound and music content. Journal on Computing and Cultural Heritage 4(1), 1–23 (2011), `http://dl.acm.org/citation.cfm?doid=2001416.2001418`
30. Wallbott, H.G.: Bodily Expression of Emotion. European Journal of Social Psychology 28(6), 879–896 (1998)