Automatic Recognition of Affective Body Movement in a Video Game Scenario

Nikolaos Savva and Nadia Bianchi-Berthouze

UCLIC, University College London, MPEB Gower Street, London, WC1E6BT, UK
{nikolaos.savva.09,n.berthouze}@ucl.ac.uk

Abstract. This study aims at recognizing the affective states of players from non-acted, non-repeated body movements in the context of a video game scenario. A motion capture system was used to collect the movements of the participants while playing a Nintendo Wii tennis game. Then, a combination of body movement features along with a machine learning technique was used in order to automatically recognize emotional states from body movements. Our system was then tested for its ability to generalize to new participants and to new body motion data using a sub-sampling validation technique. To train and evaluate our system, online evaluation surveys were created using the body movements collected from the motion capture system and human observers were recruited to classify them into affective categories. The results showed that observer agreement levels are above chance level and the automatic recognition system achieved recognition rates comparable to the observers' benchmark.

Keywords: Body movement, automatic emotion recognition, exertion game.

1 Introduction

The gaming business is changing with one of the latest highlights being the inclusion of body movement in their games (e.g., Nintendo Wii, Microsoft Kinect). As more and more companies move towards this new type of technology, researchers are exploring new ways to improve and measure the player's experience by considering the role of body movement in the game [1, 21]. An important aspect of the user experience is the affective one. Until recently, the main modality used to measure the affective state of people was their facial expressions [4]. Recent psychology studies, however, have revealed that body expressions are also a very good indicator of affect [e.g., 2, 3, 10]. These studies encouraged us into researching the possibility to create an automatic recognition system that would use the players' body movement to detect their affective state.

Previous work on this subject has been carried by various researchers even if on a smaller scale than automatic recognition of affect from facial expression. An interesting work is presented in [5] and aims at detecting emotions from non-stylised acted body motions. The movement analysed in this study are cyclic knocking arm

A. Camurri, C. Costa, and G. Volpe (Eds.): INTETAIN 2011, LNICST 78, pp. 149–159, 2012.

movements expressing either basic emotions (i.e., angry, happy, sad) or a neutral state. Using SVMs classifiers, the correct recognition rate of affective states reached 50%. However, by taking into account individual idiosyncrasies in the description of the movement, the performances increased to 81%. The recognition performances were comparable to human observers' performances (varying between 59% and 71%) for the same set of stimuli, as discussed in [2]. Another interesting study aimed at recognizing affective states is the one by Gunes and Piccardi [20]. It exploits both facial expressions and upper-body gestures. The expressions considered are anger, anxiety, disgust, happiness and uncertainty. Using BayesNet, the recognition performances reached 90% by using body expressions only.

Using acted affective postures, Kleinsmith et al. [24] explored cultural differences in expressing and recognizing affect from body expressions. The analysis, based on the set of low-level descriptive features proposed in [24], highlighted some differences between the cultures but showed also the possibility to build automatic recognition models that reflect the recognition of human observers from different cultures. Similar results were obtained for the Japanese culture on affective dimensions as discussed in [25].

In all theses studies, like many others [6, 9, 10, 11, 7], the affective states are acted and hence very stereotypical and even exaggerated making the generalization of these studies to real application scenarios more difficult.

Recently, there have been some attempts to model non-acted body expressions. A study that aims at detecting emotional states from non-acted body expression is presented in [19]. This is very relevant to our work as the scenario considered is whole-body computer games. However, the body expressions used in this study are static postures rather than movement. The recognition rates for the automatic systems were 60% on average for four affective states (concentrating, defeated, frustrated and triumphant). Their results were comparable with the human observers' level of agreement (i.e., 67% recognition rate) reached for the same set of stimuli. In the same work, the authors explore the possibility to recognize the level of arousal and valence from the postures of the players. Again, the results are comparable to human observers' agreement levels and well above chance level.

All these studies obtained quite interesting results highlighting the importance and the feasibility of using body expressions for automatic affect recognition. However, each of these studies explores a very particular type of body movement or body expression making the generalization of the results limited. Also, most of these studies focus on acted expressions.

Our focus on this study is to create a system to automatically recognize non-acted, affective expressive movements in the context of computer games. A benchmark is created from an analysis of the agreement between human observers in order to evaluate the system. The benchmark and the system are built using a dataset collected from players playing Nintendo Wii tennis games. The body movement of the players is collected during matches and represents the affective expressions that occur between the start and the end of a match point (winning or losing the point). The next session presents the method used to create the data. Section 3 describes the surveys used to build the benchmark on human observers. Finally, section 4 presents the automatic recognition system and its performance. We conclude with a result discussion and a comparison with human observers' agreement level.

2 Methodology and Data Analysis

The first step of our methodology was to obtain the body movement data. A motion capture system, Animazoo IGS-190, was used to record the movements of the participants during game play. The motion capture system has 17 sensors placed on the head, neck, spine, shoulders, arms, forearms, wrists, upper-legs, knees and feet. Nine players, ranging in age from 20 to 30 years old were recruited for the experiments. Since psychology studies suggest that players feel more emotionally engaged when they are familiar with their opponent [12], we asked the participants to bring a friend to compete with. The participants were asked to play the Wii Grand Slam Tennis game for 15 minutes while being recorded with the motion capture system and by a camera.

After collecting the motion captured data, we segmented them into ‘playing’ and ‘non-playing’ frame windows. We were able to collect 423 significant playing windows containing either winning or losing points. Each window time length varied between 10 and 40 seconds (i.e., between 600 and 2400 frames per window). By examining all the data, it was found that 248 out of 423 windows were very noisy (due to gimbal lock problem [13]) and we decided to exclude them as sufficient data would be available. As a result, our final data set consisted of 175 windows.

In order to identify the set of affective states to focus on, we first asked the participants to freely list the emotions they had felt during the game. Furthermore, we also observed the set of collected videos. At the end, eight emotion labels were selected: *Frustration, Anger, Happiness, Concentration, Surprise, Sadness, Boredom* and *Relief.*

In order to collect the affective ground truth for the collected movement, i.e, assign an affective label to each movement, and build the automatic recognition system, an online evaluation survey was conducted using computer animated avatars (See Fig. 1). These animations were built using the motion captured data corresponding to the selected 175 windows. Computer animated avatars were used instead of the video of the actual human participants to create a faceless non-gender, non-culturally specific ‘humanoids’ in an attempt to eliminate bias. The reason to use external observers rather than the players to build the ground truth is due to the unreliability of post-task reported feelings and to the fact that it is not possible to stop players during the gaming session to ask them their current affective state. Furthermore, because the complete affective state is expressed through a combination of modalities in a non-acted scenario, it is difficult for the players to be aware through which modality affect was expressed [25].

A forced choice survey was created and nine observers were recruited for the labeling task. The survey required the observers to assign one of the eight labels to each animated avatar according to the affective expression its body conveyed. We then used the most frequent label associated by the observers to an avatar as the representative affective state for that avatar. We call this label ground truth following the approach used in [19]. Fig. 2 shows the distribution of the 175 windows (animated avatars) grouped according to the associated ground truth. The results of the survey were also used to set a benchmark for evaluating the performances of the automatic recognition system. This will be discussed in section 5.

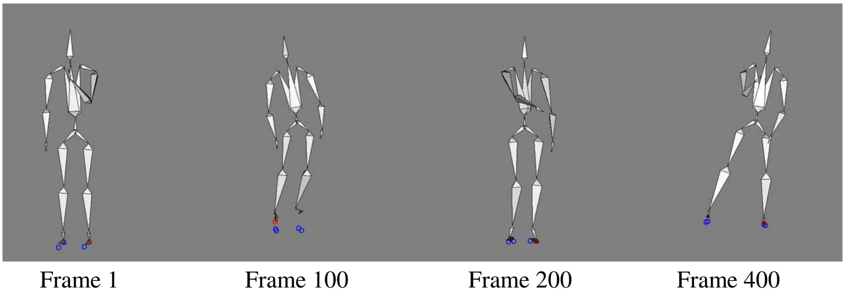

Fig. 1. The figure shows four frames of one of the avatar animations

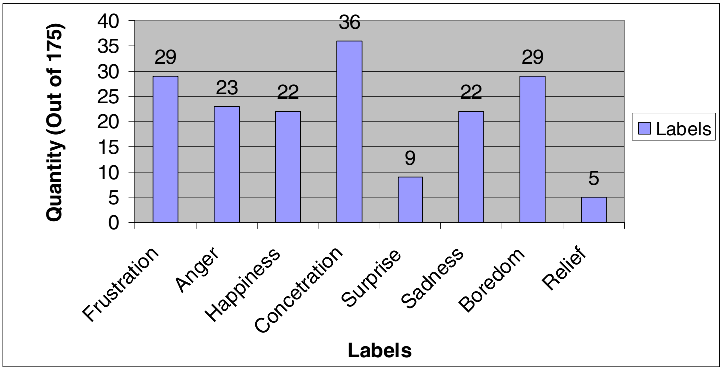

Fig. 2. Distribution of the most frequent labels associated to the 175 avatar animations

3 Low-Level Motion Description

In order to build our recognition system, the following dynamic features were selected on the basis of previous studies [e.g., 14]: *Angular Velocity, Angular Acceleration, Angular Frequency, Orientation, Amount of Movement, Body Directionality* and *Angular rotations*. As the motion capture data provided the 3D rotational information for each segment of the body (17 sensors were used as discussed on Section 2), a visual analysis of these set of features (see Fig. 3 for examples) was conducted for each body segment along the 3 rotational axes (x, y, z). An extensive graph analysis (by calculating all the features for each of the 17 sensors and for all the emotional states) was conducted in order to find the most discriminative features. From this analysis, we noticed that there was excessive variability between the participants for the data gathered from their leg sensors, so these data were discarded as they were contradictory and inconsistent. The final set of the most discriminative features (listed in Table 1) were selected to build the automatic recognition system.

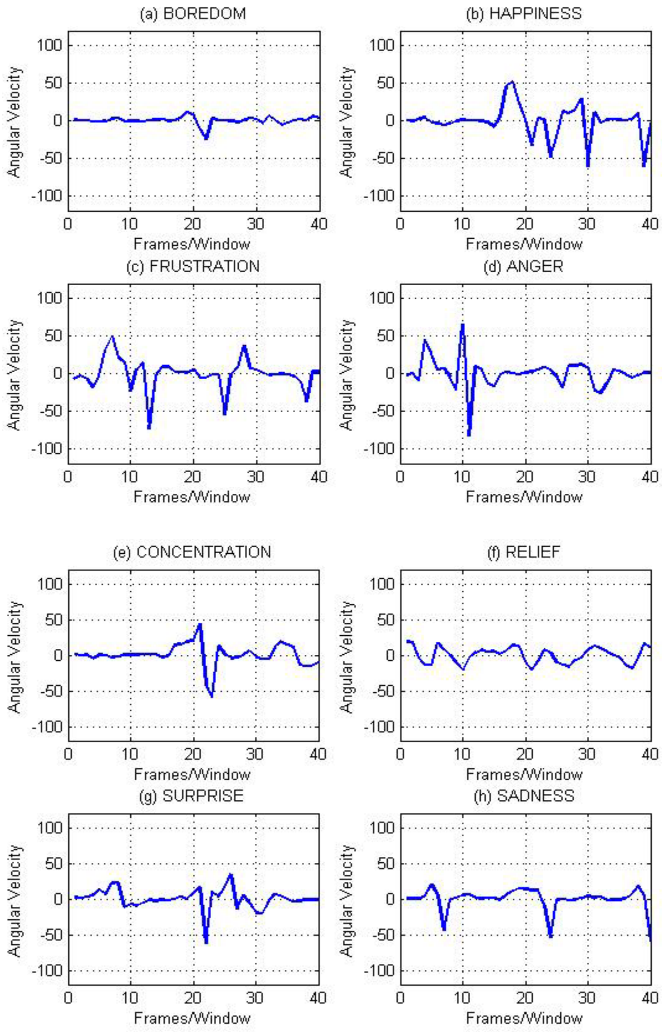

Fig. 3. Angular velocity for the X-rotation of the right forearm (Window=10)

Table 1. Identified set of discriminative features

Motion Features	Frame Interval Features
Angular Velocity$_{XYZ}$: Right Forearm, Arm, Hand Angular Accellaration$_{XYZ}$: Right Forearm, Arm, Hand Angular Frequency$_{XYZ}$: Right Forearm, Arm, Hand Body Directionality$_{X}$:Spine, Head BodySegment Rotation$_{XYZ}$: Right Forearm, Arm, Hand Angular Speed$_{XYZ}$ - Right Forearm, Arm, Hand	Amount of movement with respect to each sensor

4 Automatic Recognition System and Evaluation

Since we are dealing with time related features, a dynamic learning algorithm was better suited for building our system. A Recurrent Neural Network algorithm (RNN) [15, 16, 17] was selected. The parameters of the RNN can be seen in Table 2. The inputs to the network correspond to the set of features listed in Table 1. The number of output nodes corresponds to the number of selected emotion labels.

The testing of the learning algorithm was conducted for both the ability to generalise to new observers and to new data. The 5-fold cross validation method was employed to ensure that. The training was conducted using four subsets of the training set and then the remaining subset was used to test the algorithm's ability to generalise to new data as well as to new observers. Our first experiment showed recognition rate lower than 35%. The analysis of the results showed that most of the errors were due to misclassifications of very similar expressions: frustration with anger, sadness with boredom. Furthermore, the low number of data for surprise and relief (see Fig. 2) was also one of the main causes of misclassifications. It was hence decided to refine the set of labels to be recognized.

Table 2. Initial Network Parameters

Parameter	Value	Parameter	Value
Input nodes:	47	Momentum:	0.3
Hidden layer nodes:	90	Recurrency parameter:	0.5
Output nodes:	8	Network window size:	10
Learning rate:	0.7		

According to the literature, affective states can be divided into larger categories such as negative, positive, and neutral affective states. According to Storm et al. [18], frustration and anger are both negative and high intensity emotions and their main difference is in the intensity levels of the expression. Anger normally has higher intensity than frustration. As a result, we decided to group these two emotions into one category called 'high intensity negative emotion'. Instead, sadness and boredom are negative emotions characterized by low energy/intensity. Thus, we grouped these two emotions into one category called 'low intensity negative emotion'. Furthermore, given the low number of samples for 'surprise' and 'relief', these two labels were removed from the data set. Therefore, we are left with four classes: 'high intensity negative emotion', 'happiness', 'concentration' and 'low intensity negative emotion' and 161 windows as data set. The distribution of affective states with respect to the data set is illustrated in Fig. 4. These 4 classes of emotions cover the four quadrants of valence-arousal space generally used to describe emotional states, with Concentrated being a neutral state and Happiness, in this case, representing the high intensity positive emotions.

Various experiments to identify the best set of features were executed. The best results were obtained by using only *angular velocity, angular speed* and *amount of* *movement* as input features. Angular velocity is a vector quantity which specifies the angular speed (a scalar) of an object along with the axis which the object is rotating around. The 175 windows in our data set were further segmented into smaller frame intervals. Since each window varied between 600 and 2600 frames (10 to 40 seconds), we segmented them into smaller, equal frame intervals in order to import them into the RNN. The best performance was achieved using a network window size equal to ten. For example, a window consisting of 600 frames was segmented into 60 frame intervals containing ten consequent frames each. As a result, for one data point we extracted sixty sub-windows.

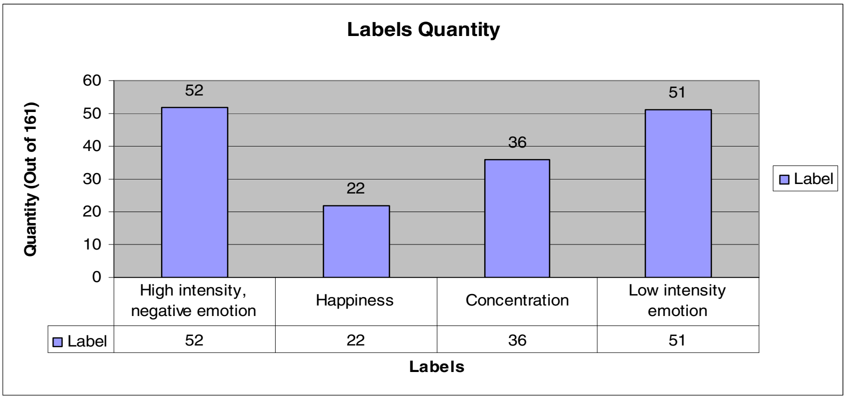

Fig. 4. The graph shows the number of animated avatars for each emotion category

Table 3. Confusion matrix for the testing set

		Predicted				
		High intensity, negative emotion	Happy	Concentr.	Low intensity emotion	Multiple classes
Actual	High intensity, negative emotion	223 (64%)	42	46	6	31
	Happy	13	124 (58%)	24	11	43
	Concentration	51	37	102 (36%)	32	62
	Low intensity emotion	9	33	49	263 (67%)	38

The resulting data set was split in two parts (training and testing set) reserving the 1/3rd, 1239 samples, to be used as a testing set and the remaining 2/3^{rds}, 3720 samples, for the training set, from overall 4959 instances. Table 3 shows the recognition performance over the testing set. Overall the network was able to categorize correctly 712 samples corresponding to 57% of the testing set. In particular, 64% of the 'high intensity negative emotion' samples were correctly classified, 58% accuracy was obtained for 'happiness' and 67% for low intensity negative emotion. Only 36% accuracy was, instead, obtained for 'concentration'. The low accuracy obtained for 'concentration' could be due to the fact that the human observers may have used this label when the avatar's expression did not express any of the other affective states as discussed in [19]. Finally, the column named as 'Multiple classes' in table 3 contains the number of the test samples that our algorithm was not able to categorize into only one class. An analysis of the results highlighted, also, the large variability between expressions belonging to the same category. This was due to the large diversity of the players' playing styles. Thus, for every class we had a variety of different input patterns. An example is provided in Fig. 5.

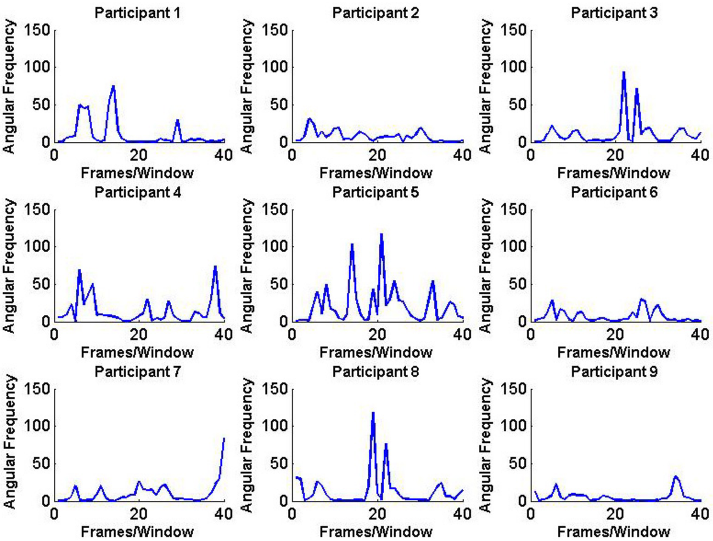

Fig. 5. The differences in angular frequency between the participants that portray anger

From Fig. 5, we can notice that players (participants) can be categorized into two groups, the ones that do not move a lot during the game (P2, P6, P7 and) and the others that move a lot (P1, P3, P4, P5 and P8). This difference between them exists because some of the participants tend to play the game using only their hand/wrist in comparison with the other group that uses their arm and shoulder as well. As discussed in [22], players adopt different body movement strategies according to their level of expertise but also according to their motivations for engaging in the game play.

5 Discussion

To evaluate the performance of the automatic recognition system, we followed a simplified version of the method proposed in [19]. The evaluation method proposed in [19] requires three groups of observers in order to fully separate the computation of the benchmark from the testing of the system. This was not possible in this case as the number of observers available was quite small. Hence, we divided the observers into two groups; the first group of observers was used to create the training set and the second group of observers was used for the testing set. The agreement level between the two groups of observers resulted in 61.49% since the two groups agreed only on 99 out of the 161 instances. Finally, we can observe that our system's accuracy (57.46%) is comparable to, even if slightly lower than, the observers' agreement. The results are hence very encouraging given the complexity of our data set. The results are also comparable to the results obtained for complex expressions in the acted and non-acted studies discussed in the introduction.

Bernhardt et al. [5] is one of the studies we can compare with ours since they used motion data instead of single postures. The researchers used arm movement features to recognize emotions from 'knocking' movements reaching similar performance with our system (59% accuracy). However, when individual idiosyncrasies were considered, their results increased to 81%. As we pointed out in Fig. 5 and various studies [19] show that, players not only have their own idiosyncrasy but they employ different strategies when playing. By taking into account such differences in the modelling process, it could be expected that the performance of our system would improve. We still have however to remember that in [5], the expressions are acted and hence simpler to discriminate, whereas in our study the expressions are non-acted and often very subtle making even the human observer recognition task much harder. Also, differently from our study, their movements were repeated and hence easily to segment into movement phases before describing them. Hence, by adding a segmentation of playing movement in our study, it is possible that our method could reach much better results.

Besides discussing the evaluation of our system, we should consider the limitations of our approach. By analysing the features visually and individually, we have possibly discarded some important ones. It is possible that combination of features that individually appear to have low discrimination may instead result in being very informative. Therefore, it would be important to perform a more thorough statistical analysis of the features and their combinations (e.g. by using PCA). Finally, by adding to the recognition system information about the type of shots being played (back-hand, fore-hand, etc) together with its features may bring better performances in the recognition of each emotion. In fact, biomechanical aspects of the type of shot may have an effect on the kinematic features considered independently of the emotion expressed. These observations will be our guide for our next step.

References

1. Kim, J.H., Gunn, D.V., Schuh, E., Phillips, B., Pagulayan, R.J., Wixon, D.: Tracking real-time user experience (TRUE): a comprehensive instrumentation solution for complex systems. In: Proceedings of the 26th Annual SIGCHI Conference On Human Factors In Computing Systems, pp. 443–452. ACM, New York (2008)

2. Pollick, F., Paterson, H., Bruderlin, A., Sanford, A.: Perceiving affect from arm movement. Cognition 82, 51–61 (2001)
3. Mehrabian, A., Friar, J.: Encoding of attitude by a seated communicator via posture and position cues. Journal of Consulting and Clinical Psychology 33, 330–336 (1969)
4. Mandler, G.: History of Psychology. Emotion, vol. 1, ch. 8. Wiley (2002)
5. Bernhardt, D., Robinson, P.: Detecting Affect from Non-stylised Body Motions. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 59–70. Springer, Heidelberg (2007)
6. Castellano, G., Villalba, S., Camurri, A.: Recognising Human Emotions from Body Movement and Gesture Dynamics. In: Paiva, A., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 71–82. Springer, Heidelberg (2007)
7. Kleinsmith, A., Fushimi, T., Bianchi-Berthouze, N.: An incremental and interactive affective posture recognition system. In: Carberry, S., De Rosis, F. (eds.) International Workshop on Adapting the Interaction Style to Affective Factors, in conjunction with the International Conference on User Modeling (2005)
8. Kleinsmith, A., Bianchi-Berthouze, N.: Recognizing Affective Dimensions from Body Posture. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 48–58. Springer, Heidelberg (2007)
9. Coulson, M.: Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. Journal of Nonverbal Behavior 28, 117–139 (2004)
10. Kleinsmith, A., De Silva, R., Bianchi-Berthouze, N.: Cross-cultural differences in recognizing affect from body posture. Interacting with Computers 18(6), 1371–1389 (2006)
11. Camurri, A., Mazzarino, B., Ricchetti, M., Timmers, R., Volpe, G.: Multimodal Analysis of Expressive Gesture in Music and Dance Performances. In: Camurri, A., Volpe, G. (eds.) GW 2003. LNCS (LNAI), vol. 2915, pp. 20–39. Springer, Heidelberg (2004)
12. Mandler, G.: History of Psychology. Emotion, vol. 1, ch. 8. Wiley (2002)
13. Kitagawa, M., Windsor, B.: MoCap for Artists: Workflow and Techniques for Motion Capture, pp. 190–194. Focal Press (2008)
14. Roether, C., Omlor, L., Christensen, A., Giese, M.A.: Critical features for the perception of emotion from gait. Journal of Vision 8(6), 15, 1–32 (2009)
15. Elman, J.L.: Finding Structure in Time. Cognitive Science 14, 179–211 (1990)
16. Haykin, S.: Neural Networks: A Comprehensive Foundation, 2nd edn., pp. 754–777. Prentice-Hall (1999)
17. Bodén, M.: A guide to recurrent neural networks and backpropagation, in The DALLAS project. Report from the NUTEK-supported project AIS-8: Application of Data Analysis with Learning Systems, 1999-2001. Holst, A. (ed.), SICS Technical Report T2002:03, SICS, Kista, Sweden (2001)
18. Storm, C., Storm, T.: A taxonomic study of the vocabulary of emotions. Journal of Personality and Social Psychology 53(4), 805–816 (1987)
19. Kleinsmith, A., Bianchi-Berthouze, N., Steed, A.: Automatic Recognition of Non-Acted Affective Postures. IEEE Transactions on Systems, Man and Cybernetics, Part B (2011)
20. Gunes, H., Piccardi, M.: Bi-modal emotion recognition from expressive face and body gestures. Journal of Network and Computer Applications 30, 1334–1345 (2007)
21. Muller, F., Bianchi-Berthouze, N.: Evaluating Exertion Games Experiences from Investigating Movement Based. Human-Computer Interaction Series, Part 4, pp. 187–207. Springer, Heidelberg (2010)

22. Pasch, M., Bianchi-Berthouze, N., van Dijk, B., Nijholt, A.: Movement-based Sports Video Games: Investigating Motivation and Gaming Experience. Entertainment Computing 9(2), 169–180 (2009)
23. De Silva, R., Bianchi-Berthouze, N.: Modeling human affective postures: An information theoretic characterization of posture features. Journal of Computational Animation and Virtual Worlds 15(3-4), 269–276 (2004)
24. Kleinsmith, A., de Silva, R., Bianchi-Berthouze, N.: Cross-cultural differences in recognizing affect from body posture. Interacting with Computers 18, 1371–1389 (2006)
25. Kleinsmith, A., Bianchi-Berthouze, N.: Recognizing Affective Dimensions from Body Posture. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 48–58. Springer, Heidelberg (2007)
26. Russell, J.A., Feldman-Barrett, L.: Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. J. Pers. Social Psychol. 76, 805–819 (1999)