Multimodal Affective State Recognition in Serious Games Applications

Athanasios Psaltis, Kyriaki Kaza, Kiriakos Stefanidis, Spyridon Thermos, Konstantinos C. Apostolakis, Kosmas Dimitropoulos, *Member, IEEE,* and Petros Daras, *Senior Member, IEEE*
Information Technologies Institute,
Centre for Research and Technology Hellas,
{at.psaltis, kikikaza, kystefan, spthermo, kapostol, dimitrop, daras}@iti.gr

***Abstract*— A challenging research issue, which has recently attracted a lot of attention, is the incorporation of emotion recognition technology in serious games applications, in order to improve the quality of interaction and enhance the gaming experience. To this end, in this paper, we present an emotion recognition methodology that utilizes information extracted from multimodal fusion analysis to identify the affective state of players during gameplay scenarios. More specifically, two mono-modal classifiers have been designed for extracting affective state information based on facial expression and body motion analysis. For the combination of different modalities a deep model is proposed that is able to make a decision about player's affective state, while also being robust in the absence of one information cue. In order to evaluate the performance of our methodology, a bimodal database was created using Microsoft's Kinect sensor, containing feature vectors extracted from users' facial expressions and body gestures. The proposed method achieved higher recognition rate in comparison with mono-modal, as well as early-fusion algorithms. Our methodology outperforms all other classifiers, achieving an overall recognition rate of 98.3%.**

Keywords— emotion recognition; multimodal fusion; serious games;

I. INTRODUCTION

Emotion recognition is a challenging issue in Affective Computing that has recently found fertile ground for application in the active research area of serious games. Understanding player's emotional response is a key issue for increasing the challenge of the game, improving the quality of interaction and enhancing the gaming experience. In that context, elicitation, detection and modeling of emotions are of great importance and constitute the initial steps of the so called affective feedback loop [1], which aims to adapt the behavior of the game towards maximizing the user's entertainment or learning. This quantification of emotional experience can either be intrusive or non-intrusive, usually taking the form of questionnaires or processing information cues from external sensors, respectively. Unobtrusive monitoring of user's affective state is important in order to retain his/her engagement level and not interrupting the flow during the game [2]. While conventional methods in human-computer interaction (HCI) research use physiological sensors to monitor emotions [3], current emotion detection technologies incorporated by games commonly include visual and audio sensors, resulting in the extraction of input modalities such as body gestures, facial expressions and verbal response [4]. Methods adopted for processing such information consider both mono-modal and multimodal approaches. For modeling emotions, theories from cognitive psychology are usually adopted. In particular, Plutchik introduces an evolutionary view of emotions, conceived as infinite processes of feedback loops, and proposes a discretized taxonomy called the Wheel of Emotions [5], illustrating their hierarchical interrelations. On the same basis, Ekman's discrete categorization of emotions [6] provides the gold-standard methodology for classifying emotions based on facial expressions. In contrast, theories preserving the continuous nature of emotions, tend to represent them in the multidimensional Euclidean space. The most prominent example is the Valence-Arousal-Dominance space (VAD) [7], which uses a 3-dimensional real valued space to define emotions.

Recent studies in emotion recognition verify the significance of combining modalities from different sources of information in order to enhance the accuracy of the recognition task [8], [9]. Towards that direction, combinations of facial, body, speech, text, and physiological modalities have been surveyed and proven their effectiveness against standard mono-modal approaches [10], [11]. The process of fusing the different input modalities can either be done in the feature level or the decision-making level, leading to early or late fusion schemes, respectively. In both cases, fusion algorithms are unable to provide accurate results in the absence of one or more modalities.

To this end, this paper aims to propose a method for multimodal emotion recognition that effectively deals with modality absence in game-related emotion recognition tasks. More specifically, the main contributions of this work are as follows: We propose a multimodal fusion architecture that uses stacked generalization on augmented noisy datasets and provides enhanced accuracy as well as robustness in the absence of one of the input modalities. Moreover, we designed a list of body actions and facial expressions commonly encountered in a typical game, eliciting emotions based on Ekman's discrete categorization theory. Based on this list of emotions, a bimodal database was created using Microsoft's Kinect sensor, containing feature vectors extracted from users' facial expressions and body gestures. Finally, the

efficacy of the proposed method was examined in an offline comparative setting against mono-modal, as well as early-fusion algorithms.

The remaining of this paper is organized as follows: In Section II, the method of stacked generalization for multimodal fusion is described. Section III is dedicated to the construction of the bimodal database for game-elicited emotions. Finally, the experimental results are presented in Section IV, while conclusions are drawn in Section V.

II. Methodology

A. Facial Expression Analysis

Facial motion plays a major role in expressing emotions and conveying messages. The analysis of facial expressions for emotion recognition requires the extraction of appropriate facial features and consequent recognition of the user's emotional state that can be robust to facial expression variations among different users. Features extracted by applying facial expression analysis techniques can range from simply geo-locating and calculating actual anthropometric measurements, to summarizing an entire group of feature-group elements under a single emotional category, such as happiness or surprise. Fig. 1 represents the overall structure of facial expression analysis features. Using sophisticated and well-trained shape and landmark tracking techniques, specific facial feature points can be identified and located for every consecutive frame obtained by a camera-like sensor. Early forms of low level data processing can then be applied to identify and track muscle activity into specific Action Units (AUs). These AUs can be seen as a form of mid-level representation of the raw data.

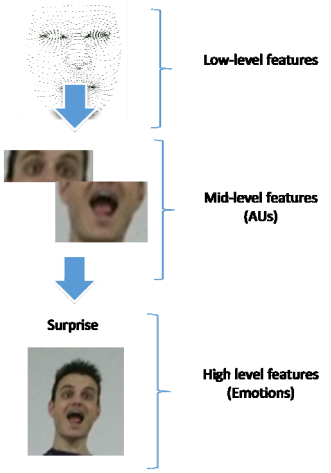

Fig. 1. Facial expression analysis features' structure.

In this respect, we follow the approach described in [12] in which landmark processing leads to low-level facial features describing the three most expressive regions of the human face: the upper component, the middle component and the lower component. Subsequently, we distinct our extracted AU features in two categories, mainly upper face and lower face AUs [13], [14]. In order to extract the aforementioned set of AUs, we followed a similar approach as [13], which incorporates the feature tracking capabilities offered by a dense-ASM tracking framework. More specifically, we employ two three-layer neural networks with one hidden layer to recognize AUs through a number of parameters defined by low-level features extracted for the upper and lower face regions. The ultimate goal of identifying and extracting AUs is to classify expressions under a certain emotion category [15]. We further concatenate the two neural network posteriors in a unified representation, and train an extra layer on top of them as shown in the left part of Fig.4.

B. Body Motion Analysis

The majority of state of the art emotion recognition frameworks capitalize mainly on facial expression or voice analysis, however, research in the field of experimental and developmental psychology has shown that body movements, body postures, or the quantity or quality of movement behavior in general, can also help us differentiate between emotions [16], [17]. To this end, we decided to extract a number of 3D body features, which are deeply inspired by psychological literature, as proposed in [18].

The 3D body movement features are extracted from joint-oriented skeleton tracking using the depth information provided by Kinect sensor. More specifically, the extracted features are classified into the following broad categories: i) kinematic related features: kinetic energy, velocity and acceleration, ii) spatial extent related features: bounding box, density and index of contraction, iii) smoothness related features: curvature and smoothness index, iv) symmetry related features: wrists, elbows, knees and feet symmetry, v) leaning related features: forward and backward leaning of a torso and head as well as right and left leaning and vi) distance related features: distances between hands, distance between hand and head as well as hand and torso. An example of the kinetic energy measurement during the play of the "Path of Trust" prosocial game [19], is demonstrated in Fig. 2.

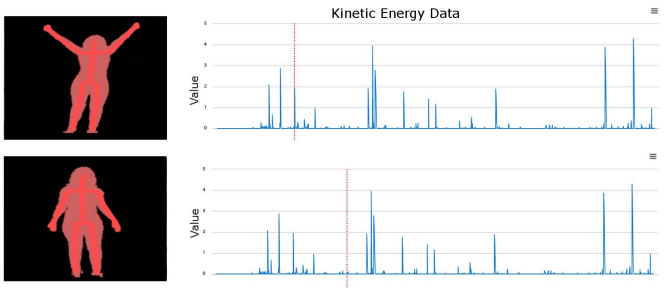

Fig. 2. Kinetic energy data measurement using the Kinect sensor during "Path of Trust" gameplay. The continuous blue line indicates Kinetic energy measurements over time through the entire session. The dotted red vertical line indicates the current frame. Top image depicts density calculation when user's body spatial extent is increased through the extension of the hands. Bottom image shows the corresponding measurement when the student's body is contracted.

For the combination of different set of features, we propose a two-layered network in which we have stacked seven NNs, six at the first layer and one at the second layer. Each layer is trained separately, starting from base layer and moving up to the second, with no feedback from the higher

layer to the lower layer. Each NN of the first layer receives as input the features of a different group of features. Then, the output probabilities of the first layer are fed as input to the second one and a separate NN is trained. The output probabilities of the second layer constitute the classification result of the body motion analysis mono-modal classifier as shown in the right part of Fig.4.

C. Dynamic Fusion of Input Modalities

The multimodal fusion process is responsible for measuring the affective state using a series of visual cues.The number of available sensors, and the total number of features that can be extracted throughout the duration of a single gameplay session, determine the choice of the appropriate level of feature abstraction, leading to robust and reliable decisions in the fusion process. In a similar way to mono-modal expression recognition, the fusion algorithms (different fusion schemes are presented for comparison reasons in the experimental results) will process either low-level feature group or high-level features in order to reach a decision on the player's affective state. Thus, facial expression data could be directly fed to our fusion algorithms, as they contain the required information in each frame. On the other hand, extracting body motion analysis features that can be fused along with the facial expression data is a challenging task. Furthermore, body motion analysis data are crucial in generating multi-modal data in gameplay environments where players' facial analysis data are noisy or even missing. Therefore, we propose a multimodal fusion architecture that uses stacked generalization on augmented noisy datasets and provides enhanced accuracy as well as robustness in the absence of one of the input modalities.

Early and late fusion approaches mainly differ in the way they combine the results from feature extraction on the various modalities. The latter fuses mono-modal decisions into a multimodal semantic representation rather than a multimodal feature representation. As a result, the fusion of decisions becomes easier, while reflecting the individual strength of modalities. Moreover, the late fusion approaches are able to draw a conclusion even when some modalities are not presented in the fusion process, which is hard to achieve in the early fusion approach. In addition, late fusion schemes offer flexibility, in a way that different computational models could be used to different modalities.

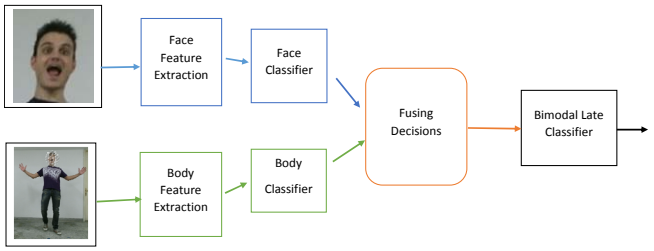

Fig. 3. Bi-modal late fusion scheme.

Recent studies have shown that, deep learning networks can be applied at feature level as well as at decision level, being trained directly on raw data or decisions accordingly. In this direction, we employ a late fusion scheme, where each intermediate classifier is trained to provide a local decision. In terms of affect, local classifiers return a confidence as a probability in the range of [0, 1] in a set of predefined classes. The local decisions are then combined into a single semantic representation, which is further analyzed to provide the final decision about the task. The aforementioned scheme for late fusion is illustrated in Fig.3.

In the proposed stacked generalization approach, we follow the late fusion scheme described above. Given a sequence of Kinect's data streams, we extract feature vectors from users' facial expressions as well as from body gestures. Then, the extracted feature vectors are fed to separate unimodal classifiers. After learning each NN model, the posteriors of the hidden variables can then be used as a new representation for the data. By adopting unified representations of the data, we can learn high-order correlations across modalities. Deeper networks with fewer hidden variables can provide simpler, more descriptive model. Therefore, we consider training an NN over the pre-trained layers for each modality, as motivated by deep learning methods. We stack NNs and train them layer-wise by starting at the base layer and moving up. This is a directed model since there is no feedback from higher layers to the lower layers, as shown in the lower part of Fig.4. This layer-wise architecture improves performance, while avoiding overfitting. In the experimental results the proposed method is compared against mono-modal, as well as early-fusion (SVM, ANN) and late fusion (Linear Weighted) algorithms.

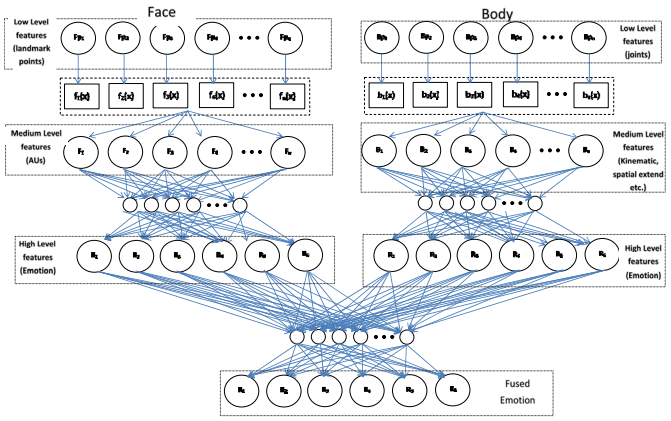

Fig. 4. The proposed architecture of the multimodal fusion method.

III. RGB-D Bimodal Database

In order to evaluate the performance of our methodology, we created a dataset with Microsoft Kinect recording of people performing 5 basic emotions (Anger, Fear, Happiness, Sadness, Surprise), which are commonly encountered in a typical game. We define a "neutral" category, to classify all frames where there is no motion indicating any of the distinct remaining emotion classes. Body movements and face expression were selected based on the literature described in previous sections. Dataset contains 450 videos from 15 subjects. Each video begins with a close to neutral expression

and proceeds to a peak expression. The total duration of its video is 3 seconds. Subjects were shown a short video with the aforementioned movements and afterwards they were asked to perform each movement according to their personal style, 5 times, in front of a Kinect sensor. The emotion label refers to what expression was requested rather than what may actually have been performed. Fig. 5 showcases a selection of these movements.

Fig. 5. Dataset Movements expressing five emotions

Inspired by the work of [20] we propose the training of the fusion model using an augmented noisy dataset with additional samples that have only a single-modality as input. In practice, we added samples that have neutral state for one of the input modalities (e.g., face) and original values for the other input modality (e.g., body). Thus, a third of the training data contains only facial expressions, while another third contains only body gestures, and the last one has both face and body information.

IV. Experimental Results

For the evaluation of facial expression recognition method, we used the baseline set of the Facial Expression Recognition and Analysis (FERA) challenge [21], which is a designated subset of the GEMEP database [22]. The dataset, which was made publicly available as part of the challenge, was created to measure the detection rates of AUs. For comparison reasons, we formalized our results against the baseline method of the FERA 2011 challenge in order to see how our neural networks performs on an entirely different dataset. The results of the AU detection, measured using F1-score for direct comparison of our approach against the FERA 2011 baseline method [21] and the corresponding reported results of a naïve AU detector, are depicted in Fig. 6.

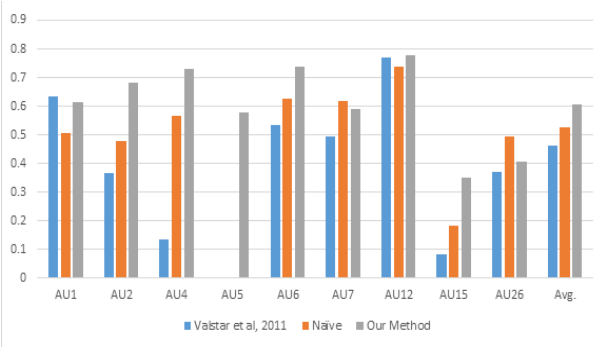

Fig. 6. Comparison of the proposed AU detection method against the FERA 2011 baseline method and a naive AU detector. For the comparison of the methods we use the F1-score metric in nine most common AUs. Our approach (grey) outperforms FERA 2011 baseline method(blue) and a naive AU detector (orange) in almost all tested AUs.

Similarly, in order to evaluate the performance of body mono-modal classifier, we created a dataset containing Kinect recordings of body movements, which express the 5 basic emotions that are likely to appear in a gameplay scenario. The proposed deep learning network classifier outperformed a number of state of the art classifiers with a recognition rate of 93%. The detailed results of these comparative study are presented in Fig.7.

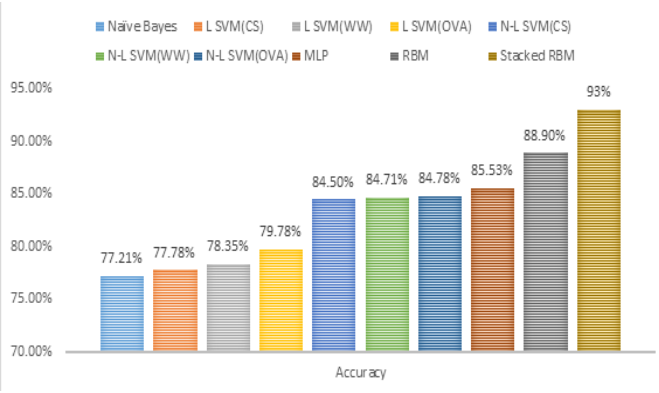

Fig. 7. Comparison of the proposed body motion analysis algorithm for emotion recognition against a number of state of the art classifiers.

Finally, for the evaluation of the proposed multimodal affective state recognition method, we examined whether the proposed fusion algorithm performs better than the intermediate mono-modal classifiers as well as a number of different multimodal approaches. More specifically, we compared the performance of the proposed algorithm against the recognition rates of mono-modal classifiers (both face and body) as well as against the recognition rates of various multimodal schemes (e.g., Linear Weighted, Non Linear SVM and Shallow MLP). As shown in Fig. 8, the proposed model outperforms all other methods, both mono-modal and multimodal (early and late fusion approaches), with a recognition rate of 98.3%. As we can see, the two mono-modal classifiers provide high recognition rates similar to those of early fusion algorithms, i.e., non Linear SVM (N-L SVM) and Shallow NN, while the proposed fusion method outperforms the linear weighted-based late fusion approach, with an improvement of 6,7%.

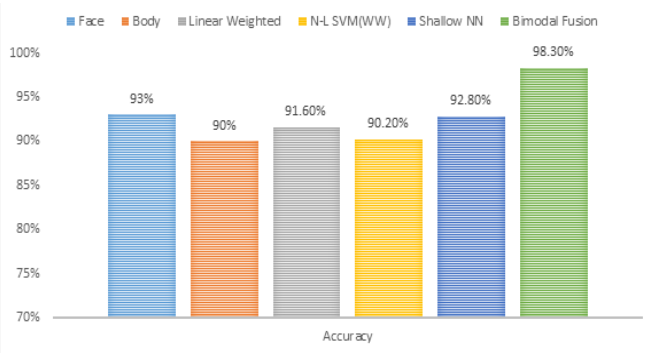

Fig. 8. Comparison of the proposed fusion algorithm against face and body mono-modal classifiers, two early fusion approaches (N-L SVM and Shallow NN) and a late fusion multimodal method (Linear Weighted).

This performance is mainly due to the fact that fusing the output of facial expression and body motion analysis classifiers improves the separability of classes, especially when mono-modal data are noisy. Fig. 9 shows a characteristic case of bi-modal emotion recognition. As we can see, although the two mono-modal classifiers provide relatively low classification confidence, the final output of the system reduces significantly the uncertainty after the fusion of the two modalities.

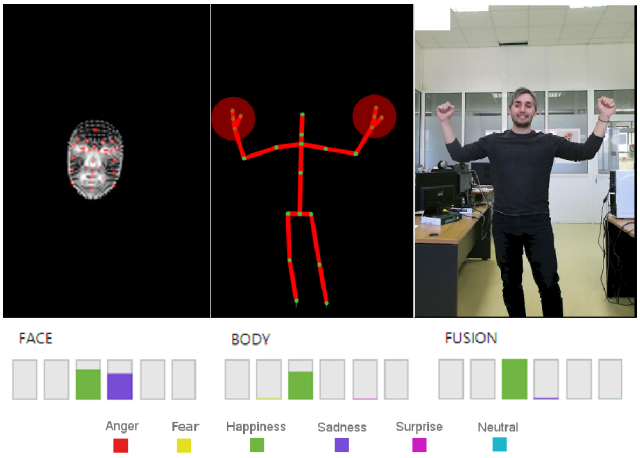

Fig. 9. A pictorial example of multimodal affective state recognition. The fusion algorithm reduces the uncertainty, although facial expression analysis provides high probabilites for two emotion classes (happiness and sadness).

V. Conclusions

In this study, we have presented a complete method for emotion recognition using multimodal visual cues that are likely to appear in a gameplay scenario. Experimental results have confirmed that fusing multimodal information cues can provide better recognition rates than their respective monomodal classifiers. The proposed multimodal fusion architecture uses stacked generalization on augmented noisy datasets and provides enhanced accuracy as well as robustness in the absence of one of the input modalities. In the future, we aim to apply our method to non-acted emotion expressions recordings, as well as to combine the affective information with game data in order to measure the player's engagement during gameplay.

Acknowledgment

The research leading to this work has received funding from the EU Horizon 2020 Framework Programme under grant agreement no. 644204 (ProsocialLearn project).

References

[1] D. Novak, A. Nagle, and R. Riener, “Linking recognition accuracy and user experience in an affective feedback loop”, IEEE Transactions on Affective Computing, pp. 1–1, 2014.

[2] M. Csikszentmihalyi, "Flow : the psychology of optimal experience", New York: Harper & Row, 1990.

[3] H. Yoon, S-W. Park, Y-K. Lee, Y-H. Jang, “Emotion recognition of serious game players using a simple brain computer interface,” IEEE International Conference on ICT Convergence (ICTC), pp. 783 – 786, , 2013.

[4] M.C. Caschera, A. D’Ulizia, F. Ferri, P. Grifoni, "Multimodal interaction in gaming" OTM 2013 Workshops, vol. 8186, pp. 694–703. Springer, Heidelberg, 2013.

[5] R. Plutchik, "Emotion: Theory, research, and experience" vol. 1. Theories of emotion 1, New York: Academic, 1980.

[6] P. Ekman and W. Friesen, "Facial Action Coding System", Palo Alto, CA: Consulting Psychologist, 1978.

[7] J. A. Russell and A. Mehrabian, “Evidence for a three-factor theory of emotions,” J. Res. Personality, vol. 11, no. 3, pp. 273–294, 1977.

[8] Z. Zeng, M. Pantic, G.I. Roisman, and T.H. 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.

[9] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, “Multimodal fusion for multimedia analysis: a survey,” Multimedia Systems, vol. 16, no. 6, pp. 345–379, Apr. 2010

[10] L. Kessous, G. Castellano, and G. Caridakis, “Multimodal Emotion Recognition in Speech-Based Interaction Using Facial Expression, Body Gesture and Acoustic Analysis,” J. Multimodal User Interfaces, vol. 3, no.1, pp. 33-48, 2010.

[11] Z. Guendil, Z. Lachiri, C. Maaoui, and A. Pruski, “Emotion recognition from physiological signals using fusion of wavelet based features,” 7th International Conference on Modelling, Identification and Control (ICMIC), pp. 1-6, 2015.

[12] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging", Affective Computing, IEEE Transactions on, *3*(1), pp. 42-55. vol. 2., pp.68-73, 2012.

[13] Y. L. Tian, T. Kanade, J.F. Cohn, "Recognizing action units for facial expression analysis", Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23(2), 97-115, 2001.

[14] M.F. Valstar, M. Mehu, B. Jiang, M. Pantic, K. Scherer, "Meta-analysis of the first facial expression recognition challenge" Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 42(4), 966-979, 2012.

[15] M. Pantic, and L.J.Rothkrantz, "Automatic analysis of facial expressions: The state of the art", Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(12), 1424-1445, 2000.

[16] P. Ekman, "Differential communication of affect by head and body cues" Journal of personality and social psychology 2, 5, 726, 1965.

[17] H. G. Wallbott, "Bodily expression of emotion", European Journal of Social Psychology 28, 879–896, 1998.

[18] S. Piana, A. Stagliano, A. Camurri, and F. Odone, "A set of full-body movement features for emotion recognition to help children affected by autism spectrum condition", In IDGEI International Workshop, 2013.

[19] K. Apostolakis, K. Kaza, A. Psaltis, K. Stefanidis, S. Thermos, K. Dimitropoulos, E. Dimaraki, P. Daras, "Path of Trust: A prosocial co-op game for building up trustworthiness and teamwork", InB Games and Learning Alliance: Fourth International Conference, GALA 2015, Rome, Italy, December 9-11, 2015.

[20] J. Ngiam, et al. "Multimodal deep learning." Proceedings of the 28th international conference on machine learning (ICML-11). 2011.

[21] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge", In Automatic Face & Gesture Recognition and Workshops, IEEE International Conference on, pp. 921-926, 2011.

[22] T. Bänziger, and K. R. Scherer, "Introducing the geneva multimodal emotion portrayal (gemep) corpus", Blueprint for affective computing: A sourcebook, 271-294, 2010.