Emotion Recognition for Affect Aware Video Games

Mariusz Szwoch and Wioleta Szwoch

Gdansk University of Technology
{szwoch,wszwoch}@eti.pg.gda.pl

Abstract. In this paper the idea of affect aware video games is presented. A brief review of automatic multimodal affect recognition of facial expressions and emotions is given. The first result of emotions recognition using depth data as well as prototype affect aware video game are presented.

1 Affect Aware Video Games

Affective computing is one of the active research areas on human behavior studies and has a significant impact in many fields such as healthcare, education, entertainment etc. [2,3]. Affective computing studies how to automatically recognize, interpret, and process human emotions via analyzing available sensory data. Affect-aware applications, as their additional functionality, react to user emotions.

Video games belong to the wide area of entertainment applications. Though, video games are among some of the most natural applications of affect only few of them seek to incorporate their players affective state into the gameplay. Such games can be referred as affective or more properly affect-aware games. Unfortunately, this affect-awareness is usually statically built-in the game at its development stage basing on the assumed model of so called representative player [4]. There are two problems with such attitude. Firstly, each player differs in some way from that averaged model. Secondly, and more important, players affect state can change even radically from session to session making almost impossible to predict the current user emotions at the development stage. That is why the real-time recognition of players affect may become such important for video games industry in the nearest future.

For the last several years only a few truly affect-aware games have been developed, mainly as noncommercial academic projects. As an example, "Feed The Fish" takes a players facial expressions as input and dynamically responds to the player by changing the game elements [5]. The goal of this system is to use human expressions to build a communication channel between the game and players so playing the game can be more enjoyable [5].

Another attempt is to create a universal architecture for affect-aware games. In [1] a Koko library framework is proposed that abstracts an affect model and sensors handling from other game components which are application logic and

R.S. Choraś (ed.), *Image Processing & Communications Challenges 6*,
Advances in Intelligent Systems and Computing 313, DOI: 10.1007/978-3-319-10662-5_28

the game engine. As a proof of concept two sample games had been developed basing on simple physiological and physical signals such as heart rate, skin conductance, or GPS players location.

Nowadays commercial games put a focus on engaging players emotionally through both gameplay techniques and interaction with virtual agents or human players in multiplayer games. But it is quite probable that this situation will change in the nearest future due to development of affect recognition techniques and new input devices and non-intrusive sensors allowing to observe the player in different information channels or modalities. For example, Valve Software, expect affect emotion as essential element of future games, and experimented with biometrics directly by introducing them into a special build of Left 4 Dead 2 [6].

2 Automatic Affect Recognition

Even the best emotional model is able only to predict the current humans affect state, that is why affect recognition is so important in affective software dealing with real users as it try to recognize real not estimated emotions.

There are several affect models developed for different application fields. Some models define continuous emotional space of two or more dimensional coordinate system and place all emotions within this model space. The best known example of this approach is PAD emotional state model defining three dimensions: Pleasure, Arousal, and Dominance, which express respectively, pleasure, intensity and the dominant nature of the emotion [7]. Using this model several different emotion sets may be defined containing 6, 7, or more affect states [8]

Emotion, or affect recognition is a key to create a truly affect-aware video games or other software. It may seem that affect recognition techniques could be applied in applications in any field but, in general, different applications differ in affect model of the user and possible sets of information inputs. For example, health care applications more probably seek for pain or mental illness symptoms allowing to use wired or even intrusive sensors to receive very trusty information. On the other side, affective e-learning and training systems seeks rather for users attention, engagement or boredom and do not demand any other input devices than internet camera and microphone. Affect-aware video games seems similar to e-learning software in the aspect of expected affect states which, however, can be extended by a fear, excitement and other emotions which are not so probable in other applications. Video games, also, may demand some additional controllers and gamers are usually open for new technical solutions that could improve games playability.

So, there are two important assumptions of affect recognition model that should be defined for specific application. The primary is the affect model and the subset of emotions to be recognized. The other is the definition of input devices and sensors to be used.

The goal of automatic emotion recognition is to recognize the current affect state of the user basing on potentially multimodal input signals about the human appearance or physiological parameters as well as other information such as environmental conditions, application mode, the interaction history, etc. The research in this field has been performed since early 2000., attracting the interest of artificial intelligence and computer vision research communities. Most attempts use face as a primary (and most often the only) source of information trying to recognize facial expressions from static images and video recordings [9,10,11]. This attitude proved to be successful in the case of good lighting and exposure conditions. Other research focused on other expressions of human emotions such as speech, gestures, gaze, and others [11]. Also physiological signals are often used especially in health care systems [2].

Numerous research projects proved that multimodal approach gives in general better results, in terms of recognition efficiency, then taking into account single data input [2]. Combination of multiple types of inputs from different modalities or different features over the same modality vastly significantly improves the system classification abilities [12]. In general, there are two main multi-modal information fusion attitudes. In the first attitude, called early, or feature-level fusion, techniques and methods for affect recognition feed a set of features into a commonly used classifier at the feature level. In the other approach, called late, or decision-level fusion, the recognition is based on some kind of decision systems which input is fed by the classification results of monomodal classifiers [1,12]. The decision system can be a voting expert system, decision tree, belief network, and other.

It is also possible to create hybrid systems in which classifiers of the first level can also use multimodal data, e.g. from a single channel like video or physiological signals.

The process of affect recognition does not differ in its essence from other classification systems such as optical character recognition (OCR), face recognition etc. Classification assumes existence of classifier C that decides about belongingness of an object o_i to one of predefined classes x_j. The object being classified is represented by a feature vector $\boldsymbol{y}$ which can contain features of any type extracted from the object (Fig. 1). In most applications a supervised learning is used to create a classifier on the base of labeled set of objects representative for each class x_j. There is a variety of classifiers type from simple to complex ones that can be used in almost each classification task or only in specific applications. Also, many existing classification algorithms may be adapted to new tasks. There are a large number of software libraries that make available many classifiers of different kind. These classifiers should be fed by properly extracted features (Fig. 1). Thus, the crucial matter is to properly select channels to acquire information about the classified objects, and properly preprocess it. Efficient solving of any recognition problem usually requires researchers with knowledge and pragmatic experience in this field. In most cases, the recognition process uses the same general scheme and proceeds through the same stages (Fig. 1). Acquired data usually needs some initial preprocessing (e.g. filtering, noise removal) and optional segmentation (e.g. background removing). From the preprocessed data a set of different features can be extracted. If the number of features is very large only most informative can be selected to form a feature

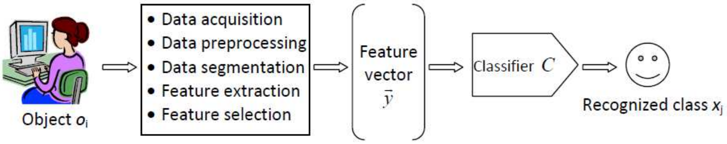


Fig. 1. General recognition process

vector that indirectly characterizes the object. Finally, the classifier C classifies an object into one of possible classes x_j .

2.1 Emotion Recognition from Video Channel

Many affect recognition systems are based on video channel taking into account facial expressions, gestures and posture analysis in mono-modal as well as in multi-modal systems [2,9,10,15]. This situation results from non-intrusive property of cameras and becomes more and more practical with the rapid development of hardware capabilities and computer vision technology. Moreover, vision channel is the most informative, as human beings tend to express their emotions in a visual way.

Automatic analysis of facial expression from video is one of the most common approaches in affective computing [18]. The most extensive work on facial expression analysis is summarized in several survey papers [10,11]. The most advanced algorithms define human face models with characteristic control points and a set of transformations of those points for each recognized facial expression. The active appearance model (AAM) is a statistical approach that models the shape and texture of a target object and has a great success in modeling and recognition of human faces. Traditional AAM framework can fail when the face pose changes as only 2D information is used to model a 3D object. To overcome this limitation, different 3D extensions of AAM are proposed [20] as well as 3D Morphable Models (3DMM) [21]. Both attitudes has a great recognition efficiency of facial expressions and their main limitation is high computing complexity.

Other attitudes use detection of facial characteristic points (e.g. eyes, elbows, nose etc.) using different heuristics and algorithms including texture matching, edge detection, profile matching and other.

Analysis of body posture, gestures, hand and body movements is also an active research area in many computer vision applications. For example, in [16], three types of affective states are recognized using combined sensory information from the face video, the posture sensor and the game being played. In [17] facial expressions are combined with hand gestures for recognition of six prototypical emotions.

2.2 Emotion Recognition from Audio Channel

Apart from facial expression, human emotions can also be indicated by voice intonation and also by spoken words. Voice intonation expressed in its pitch, tone and cadence may be described by a set of parameters calculated for time and spectral domains in subsequent time windows. Some sample parameters include signal amplitude, mean and power value, zero-crossing as well as Melfrequency Cepstral Coefficients (MFCC) which has been widely used in acoustic signal processing as they fully consider the hearing characteristics of human beings [19]. In most case the audio channel is used in multimodal affect recognition systems combined with facial expression analysis [22,8].

2.3 Emotion Recognition from Physiological Signals

Cognitive-motivational-emotive model assumes that human affect expression consists both of emotional state contains both of affect as well as physiological response. Sometimes affect expressions may be better or worse controlled by human but it is very hard or even impossible to control physiological reactions.

There are many physiological signals that can be measured and input to the system as indicator of human emotions. The most important signals are: skin conductance and temperature, blood volume pulse, electromyography, respiratory signal and electrocardiography. All the physiological sensors are non-invasive but sometimes they may be intrusive or not comfortable for users. For evaluating peoples emotions by physiological signals special equipment is needed.

There are two main problems with using of physiological signals in automated emotion recognition. Firstly, most of these signals respond also to other inner and outer factors such as human health and physical condition, temperature, humidity and others.

Secondly, measuring physiological signals often cause some inconvenience to the user what is rather unacceptable in a usual computer usage. Fortunately, a great development of non-intrusive, miniaturized and low power remote sensors has been observed that allows for remotely collecting behavior and physiological data from people. Also several computer hardware manufacturers plan to embed some physiological sensors into the control devices for video games.

2.4 Enhancing Emotion Recognition Using Scene Depth Data

The common problem of algorithms processing visual (both achromatic and color) data are insufficient and uneven lighting conditions. Though, there are many attempts to correct scene luminance in such case, they are not always sufficient enough. Depth sensors allow for acquiring the depth image of a scene. As the most popular depth sensors use non-visual infrared light technology they are generally resistant to common problems of RGB cameras. That is why information from depth sensors seems to be very useful when combined with optical channel or even not.

Depth sensors have become very popular since introducing Microsoft Kinect sensor for Xbox console in 2010. This relatively cheap sensor, accompanied with RGB camera, set of microphones and natural user interface (NUI) library gives to software developers a powerful tool for creation of applications that understandİ human stature, movement, and speech, and in the nearest future will try to recognize user face and emotions as well.

Depth sensors can be used to enhance processing of visual information at different stages of recognition pipeline (Fig. 1). For example, initial test indicate that scene depth information significantly improve image segmentation allowing to extract users hand and body from the background [13]. Also efficiency improvement for face localization and recognition of characteristic face points is expected due to the 3D nature of human face. The theoretically estimated and practically measured Kinect resolution is about 2mm per voxel and seems enough for mentioned applications. There is also a specific problem with processing of depth information which is noise level. Initial tests show that it can be efficiently limited using low pass filtering in time domain with a window of several frames [13].

3 Preliminary Results and Future Plans

The main goal of the research project Emotion Recognition for Affect-Aware Video Games (ERAAVG) is creation of a complete multi-modal affect recognition system to provide automatic and continuous affect recognition of video game players and definition of a complete framework that would enable cooperation between developed affect recognition system and any video game by application programming interface (API) defined within the project. To validate results of the project a prototype affect aware video game will be developed that would adopt its parameters, such as difficulty level, pace or humor basing on the current affect state of a player and challenges in the game.

The first task of the project is to create a multimodal database of emotions recordings that will serve as learning and testing sets for classifiers developed in succeeding steps. The second task of the project is to work out proper preprocessing and segmentation techniques as well as to define the most informative features sets for each information channel. The third task of the project is to create new algorithm for facial expression recognition basing on RGB and depth images. The fourth task of the project is to create separate algorithms for affect recognition basing on data from different channels. In order to validate results of emotion recognition we plan to develop a prototype game with bidirectional interface to affect recognition software.

In the following subsections the preliminary results of ERAAVG project are described.

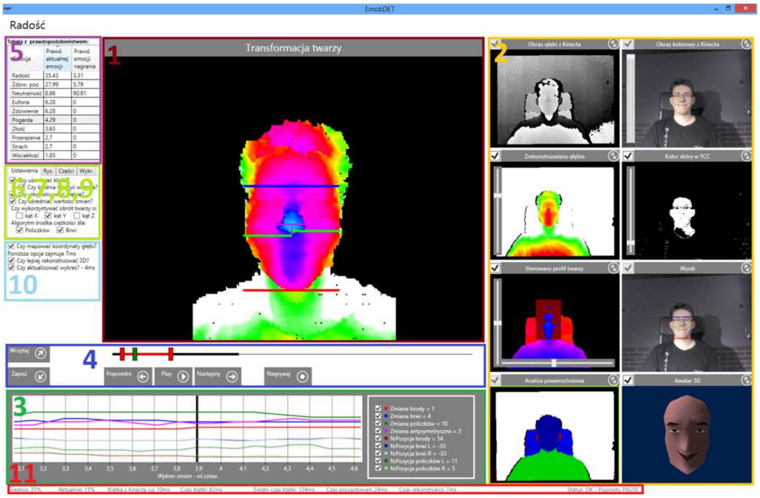
3.1 FEEDB Database

Facial Expressions and Emotions Database (FEEDB) has been created as the first task of the project [13], [14]. This multimodal database contains totally 3200 recordings of emotions that will serve as learning and testing sets for classifiers. There are two parts, or versions, of FEEDB. The first one consists of 1650 recordings of 50 persons posing for 33 different facial expressions and emotions [13] which are stored using Microsoft proprietary XED format. The second version of FEEDB consists of 1550 recordings of 50 persons [14] recorded as AVI video streams. Additionally, the most interesting set of 9 emotions had been selected to be expressed spontaneously. Special video materials had been prepared to provoke tested persons to naturally express the following expressions: surprise, joy, anger, scorn, disgust, sadness, fear, concentration, and excitement.

The database consist of synchronized multimodal recordings with RGB, depth, and audio channels acquired using Microsoft Kinect for PC sensor at a resolution 640x480 pixels at 30 fps. Some of the recordings are accompanied by additional information from physiological signals (heart rate, skin conductance, breath and temperature) acquired using FlexComp Infinity device with appropriate sensor set. Chosen frames of recordings are indexed by defining characteristic points of the face and can be used for learning and testing of the classifier for emotion recognition.

3.2 Face Detection and Emotion Recognition

FEEDB database has been used as a learning and testing dataset in a prototype system for recognition of 9 emotions using recordings of depth channel only. The set of emotions covers neutral, joy, surprise neutral and positive, euphoria, fear, fright, anger, and scorn (Fig. 2). The application locates the face and its characteristic points, then classifies emotions on the base of recognized action units (fundamental movements of face muscles in Facial Action Coding System). The system recognizes emotions in real-time and its average efficiency of the recognition for 25 persons was 50%. Though it is not very high it is a good starting point for further research especially in fusion of ther input channels. The experiment proved that FEEDB can be practically used in face and emotion recognition tasks.



Radość

Tabela z prawdopodobieństwem:

Emocja	Prawd. aktualnej emocji	Prawd. emocji nagrania
Radość	35.43	3.31
Zdziw. poz.	27.99	5.79
Neutralność	8.86	90.91
Euforia	6.28	0
Zdziwienie	6.28	0
Pogarda	4.29	0
Złość	3.63	0
Przerażenie	2.7	0
Strach	2.7	0
Wściekłość	1.85	0

Ustawienia | Rys. | Części | Wykr.

☑ Czy uś[illegible] kl[illegible]?
☑ Czy średnia [illegible]ć w[illegible]a?
☑ Czy [illegible]kstru[illegible]ębie?
☑ Czy uśredniać wartości zmian?
Czy wykorzystywać obrót twarzy o:
☐ kąt X ☑ kąt Y ☐ kąt Z
Algorytm środka ciężkości dla:
☑ Policzków ☑ Brwi

☑ Czy mapować koordynaty głębi?
Poniższa opcja zajmuje 7ms
☑ Czy lepiej rekonstruować 3D?
☑ Czy aktualizować wykres? - 4ms

☑ 🟥 Zmiana brody = 1
☑ 🟦 Zmiana brwi = 4
☑ 🟩 Zmiana policzków = 10
☑ 🟪 Zmiana antysymetryczna = 3
☑ 🟥 N:Pozycja brody = 54
☑ 🟦 N:Pozycja brwi L = -33
☑ ▨ N:Pozycja brwi R = -33
☑ 🟩 N:Pozycja policzków L = 11
☑ 🟩 N:Pozycja policzków R = 5

Średnio: 35% Aktualnie: 15% Klatka z Kinecta co: 10ms Czas klatki: 62ms Średni czas klatki: 124ms Czas przygotowań: 24ms Czas rekonstrukcji: 7ms Status: OK - Poprostu PAUSE

Fig. 2. Application for emotion recognition using depth maps

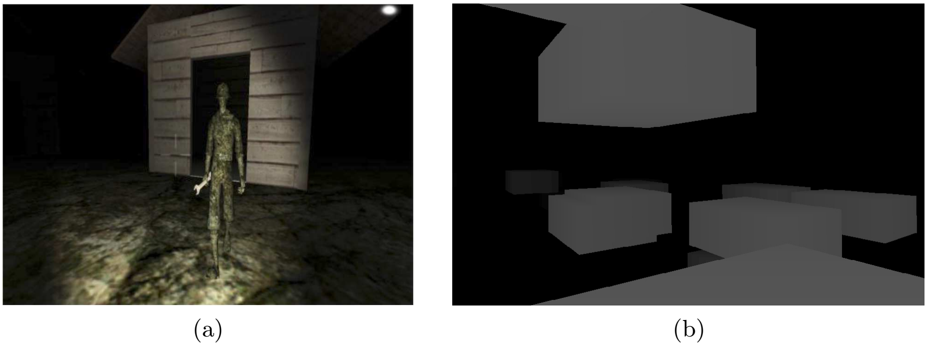


Fig. 3. Sample screens from prototype of an affect-aware video game (a) Zombie land (b) moving platforms

3.3 Affect Aware Video Game

A prototype affect aware video game has also been developed within ERAAVG project using Unity 3D game engine (Fig. 3). The game consists of three levels with different challenges, namely escaping from attacking zombies (Fig. 3a), jumping across moving platforms (Fig. 3b) and finding exit of dynamically reconfiguring maze. The difficulty of each level changes depending on current player state. For example at the first level the user under the stress finds torch batteries more frequently to threaten zombies. In turn, the speed of movings platforms depends on user self-assurance, but this time platforms speeds-up to make the game more difficult. Unfortunately, the game has not been connected with emotion recognition application and predicts users states basing on users behavior only. Nevertheless, initial experiments with several players has resulted with positive opinions as the game was able to surprise the player by dynamic changes.

3.4 Conclusions and Future Works

Affect aware video games seem to be very interesting to players as they engage players emotionally and can dynamically react to players emotional state which was confirmed by a prototype game developed within ERAAVG project. Such video games can have an entertainment character but can also be used in other applications, such as e-learning, psychological training or therapy, and even in marketing systems.

Automatic emotion recognition is a key to create affect-aware software. A prototype emotion recognition system proved that real-time emotion recognition is possible, though fusion of many input channels is needed to receive reliable results. A comprehensive data set, such as FEEDB, is also needed to develop emotion recognition system.

The next task in ERAAVG project is to develop a bidirectional framework that could connect affect recognition application with any affect-aware video game using a specific application programming interface. Further steps cover development of more efficient affect recognition algorithm using multimodal data such as RGB and depth video, audio channel, and physiological sensors. Also other affect aware video games will be developed that could increase players satisfaction from a gameplay.

Acknowledgment. The research leading to these results has received funding from the Polish-Norwegian Research Programme operated by the National Centre for Research and Development under the Norwegian Financial Mechanism 2009-2014 in the frame of Project Contract No Pol-Nor/210629/51/2013.

References

1. Derek, J., Sollenberger, Â., Munindar, P.: Singh: Koko: an architecture for affect-aware games. In: Proceedings of the First International Workshop on Agents for Games and Simulations. Springer (2010)
2. Liu, X., Zhang, L., Yadegar, J.: An intelligent multi-modal affect recognition system for persistent and non-invasive personal health monitoring. In: 2011 IEEE 22nd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), pp. 2163–2167. IEEE (2011)
3. Picard, R.: Affective Computing: From Laughter to IEEE. IEEE Transactions on Affective Computing 1(1) (2010)
4. Adams, E.: Fundamentals of Game Design. New Riders Publishing (2009)
5. Obaid, M., Han, C., Billinghurst, M.: Feed The Fish: An Affect-Aware Game. In: Australasian Conference on Interactive Entertainment 2008, Brisbane, Australia, December 3-4 (2008)
6. Leadbetter, R.: Games will detect your feelings (2011), `http://www.eurogamer.net/articles/digitalfoundryvalve-biometrics-blog-entry`
7. Russell, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. Journal of Research in Personality, 273–294 (1977)
8. Ortony, A., Clore, G.L., Collins, A.: The cognitive structure of emotions. Cambridge University Press, Cambridge (1988)
9. Pantic, M., Rothkrantz, L.J.M.: Automatic analysis of facial expressions: the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 22(12), 1424–1445 (2000)
10. Fasel, B., Luettin, J.: Automatic Facial Expression Analysis: A Survey. Pattern Recognition 36(1), 259–275 (2003)
11. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information. In: Proc. of ACM 6th International Conference on Mutlmodal Interfaces (2004)
12. Gunes, H., Piccardi, M.: Affect Recognition from Face and Body: Early Fusion vs. Late Fusion. In: Proc. IEEE Int. Conf. Systems, Man, and Cybernetics SMC, pp. 3437–3443 (2005)
13. Szwoch, M.: On Facial Expressions and Emotions RGB-D Database, BDAS (2014)
14. Szwoch, M.: FEEDB: a multimodal database of facial expressions and emotions. In: Proc. of the 6th Int. Conf. on Human System Interaction, pp. 524–531 (2013)

15. Pantic, M., Rothkrantz, L.J.M.: Toward an Affect-Sensitive Multimodal Human-Computer Interaction. Proc. of IEEE 91(9), 1370–1390 (2003)
16. Kapoor, A., Picard, R.W., Ivanov, Y.: Probabilistic Combination of Multiple Modalities to Detect Interest. In: Proc. IEEE ICPR (2004)
17. Balomenos, T., Raouzaiou, A., Ioannou, S., Drosopoulos, A., Karpouzis, K., Kollias, S.D.: Emotion Analysis in Man-Machine Interaction Systems. In: Bengio, S., Bourlard, H. (eds.) MLMI 2004. LNCS, vol. 3361, pp. 318–328. Springer, Heidelberg (2005)
18. Ekman, P., Friesen, W.V.: Unmasking the face: a guide to recognizing emotions from facial clues. Prentice-Hall, Imprint Englewood Cliffs (1975)
19. Sigurdsson, S., Petersen, K., Schioler, T.: Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded Music. In: Proc. Int. Conf. Music Inf. Retrieval, pp. 286–289 (2006)
20. Chen, C.W., Wang, C.C.: 3D Active Appearance Model for Aligning Faces in 2D Images. In: International Conference on Intelligent Robots and Systems IROS, pp. 3133–3139 (2008)
21. Xiao, J., Baker, S., Matthews, I., Kanade, T.: Real-Time Combined 2D+3D Active Appearance Models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 535–542 (2004)
22. Chen, L.S., et al.: Emotion recognition from audiovisual information. In: IEEE Second Workshop on Multimedia Signal Processing, December 7-9, pp. 83–88 (1998)