Improved Multimodal Emotion Recognition for Better Game-Based Learning

Kiavash Bahreini[1], Rob Nadolski[1], and Wim Westera[1]
[1] Welten Institute, Research Centre for Learning, Teaching and Technology, Faculty of Psychology and Educational Sciences , Open University of the Netherlands
Valkenburgerweg 177, 6419 AT Heerlen, The Netherlands
{kiavash.bahreini, rob.nadolski, wim.westera}@ou.nl

Abstract. This paper introduces the integration of the face emotion recognition part and the voice emotion recognition part of our FILTWAM framework that uses webcams and microphones. This framework enables real-time multimodal emotion recognition of learners during game-based learning for triggering feedback towards improved learning. The main goal of this study is to validate the integration of webcam and microphone data for a real-time and adequate interpretation of facial and vocal expressions into emotional states where the software modules are calibrated with end users. This integration aims to improve timely and relevant feedback, which is expected to increase learners' awareness of their own behavior. Twelve test persons received the same computer-based tasks in which they were requested to mimic specific facial and vocal expressions. Each test person mimicked 80 emotions, which led to a dataset of 960 emotions. All sessions were recorded on video. An overall accuracy of Kappa value based on the requested emotions, expert opinions, and the recognized emotions is 0.61, of the face emotion recognition software is 0.76, and of the voice emotion recognition software is 0.58. A multimodal fusion between the software modules can increase the accuracy to 78%. In contrast with existing software our software modules allow real-time, continuously and unobtrusively monitoring of learners' face expressions and voice intonations and convert these into emotional states. This inclusion of learner's emotional states paves the way for more effective, efficient and enjoyable game-based learning.

Keywords: Game-based learning, human-computer interaction, multimodal emotion recognition, real-time emotion recognition, affective computing, webcam, microphone.

1 Introduction

Recent technologies have been adopted by e-learning experts for improving the efficiency, effectiveness and enjoyableness of e-learning [1]. Currently, learners are habitually accustomed to the web-based delivery of e-learning content when communicating, working and learning together with their peers in distributed (a)synchronous settings [2]. It is broadly acknowledged that emotions are important in all learning activities, as they affect information processing, memory usage and performance [3].

This study is part of our research [4, 5, 6] that aims at improving multimodal emotion recognition for better online game-based learning but that can also be applied in e-learning. Game-based learning has several advantages: 1) it is a didactical approach that looks to be in-line with the learners' interests [7], 2) can be very effective for skills training [8], 3) encouraging [9], and 4) it is very fashionable most recently [7]. Learners' vocal and visual input to the technology can be used via a game-approach for enhancing their learning experience. This learning experience therefore becomes more informal, though not pure entertainment. A delicate balance between game play and learning is important [10]. Especially the training of recurrent skills might benefit from a game-based approach, as these require frequent practice where individuals remain to be motivated [11]. Furthermore, as skills training are time-consuming, technology could alleviate the workload of trainers and might also lead to improved face-to-face training as trainers can focus on the training of non-recurrent skills [12]. To accomplish this purpose, the here described and tested software technology from FILTWAM will ultimately be combined with a game-based didactical approach. FILTWAM will be combined with EMERGO for mainly practical reasons. EMERGO is an in-house developed and tested methodology, and open source toolkit for the development and delivery of serious games [13].

FILTWAM uses webcams and microphones to interpret the emotional state of people during their interactions with a game-based learning environment via an affective computing tool. It triggers timely feedback based upon learner's facial expressions and verbalizations. It is designed for distinguishing the following emotions: happiness, sadness, surprise, fear, disgust, anger, and neutral. It mainly offers software with a human-machine interface for the real time interpretation of emotion that can be applied in game-based learning and e-learning. The study is a follow up of our previous studies [4, 5, 6]. It aims at extending our game-based learning setting for multimodal emotion recognition. For this, the affective computing tool is composed of face and voice emotion recognition modules. The affective computing tool represents the development of a software system, which is able to recognize and interpret human emotions. The affective computing tool of FILTWAM is built upon existing research [14, 15, 16, 17]. Linking two modalities (face expression and voice intonation) into a single system for affective computing analysis is not new and has been studied before [18, 19, 20, 21]. A review study by [22] shows that the accuracy of detecting one or more basic emotions is significantly improved when both visual and audio information are used in classification, leading to accuracy levels between 72% and 85%.

Although digital learning is widely used true interaction with digital learning artifacts, it is still limited despite the recent developments of input devices. Webcams and microphones not only offer opportunities for more natural interactions with digital learning artifacts (like serious games) but also offer ways of unobtrusively gathering affective user data during learning process.

We performed one of the most widely used linear fusion method in this study that has been reported in [23]. We propose 1) an unobtrusive approach that supports 2) a real-time interpretation of emotion with 3) an objective method that can be verified by

researchers, which requires 4) inexpensive and ubiquitous equipment, and offers 5) two interactive software modules.
In this paper, section 2 introduces the FILTWAM framework and its sub-components. The methodological setup of the validation of the software modules is described in section 3. Results and discussion are presented in section 4. Section 5 discusses the findings and limitations of this study and proposes future improvements.

2 The FILTWAM Framework

The FILTWAM framework includes five layers and a number of sub-components within the layers (see Figure 1). The five layers are introduced as the: 1) Learner, 2) Device, 3) Data, 4) Network, and 5) Application. We used EMERGO in conjunction with FILTWAM in this study; however FILTWAM can also be used with other game-based learning environments.

2.1 Learner Layer

The learner refers to a subject who uses web-based learning materials for personal development or preparing for an exam.

2.2 Device Layer

The device layer is the most important part of FILTWAM. The device reflects the learner's machine, whether part of a personal computer, a laptop, or a smart device. It includes a webcam and microphone for collecting user data. It contains three sub-components named: the web interface, the EMERGO web service client, and the affective computing tool.

2.2.1 Web Interface

The web interface runs a serious game in the device layer and allows the learner to interact with the game components in the application layer. This component indirectly uses the EMERGO web service client. The web interface will receive the feedback/content through Internet and the game-based learning environment in application layer.

Fig. 1.

2.2.2 EMERGO Web Service Client

The EMERGO web service client uses the affective computing tool; calls the EMERGO web service in the application layer. It reads the affective data and broadcast the live stream including the face emotion recognition data and the voice emotion recognition data through Internet to the EMERGO web service.

2.2.3 Affective Computing Tool

The affective computing tool is the heart of FILTWAM. It processes the facial behavior and vocal intonations data of the learner. It consists of two components for the emotion recognition of both vocal and facial features. The emotion recognition of the vocal features uses the microphone voice streams whereas the emotion recognition of the facial features uses the webcam face streams.

Emotion Recognition from Facial Features

This component extracts facial features from the face and classifies emotions. It includes three sub-components that lead to the recognition and categorization of a specific emotion.

Face Detection

The process of emotion recognition from facial features starts at the face detection component. But we do not necessarily want to recognize the particular face; instead we intend to detect a face and to recognize its facial emotions.

Facial Feature Extraction

Once the face is detected, the facial feature extraction component extracts a sufficient set of feature points of the learner. These feature points are considered as the significant features of the learner's face and can be automatically extracted.

Facial Emotion Classification

We adhere to a well-known emotion classification approach that has often been used over the past thirty years which focuses on classifying the six basic emotions [15]. Our facial emotion classification component supports the classification of these six basic emotions plus the neutral emotion, but can in principle also recognize other or more detailed face expressions when required. This component analyses video sequences and can extract an image for each frame for its analysis. This component is independent of race, age, gender, hairstyles, glasses, background, or beard and its development is based on the FaceTracker software [24]. During the analysis, one image that already includes a not-yet determined emotion is compared with all already classified images in the dataset. Then this image will be classified as one of the indicated emotions. It compares the classified emotions with existing emotions in the facial emotion dataset and trains the dataset using a number of learners' faces.

Emotion Recognition from Vocal Features

This component extracts vocal intonations from voices and classifies emotions. It includes three sub-components that lead to the recognition and categorization of a specific emotion.

Voice Detection

The process of emotion recognition from vocal intonations starts at the voice detection component. But we do not necessarily want to recognize the particular voice; instead we intend to detect a voice and to recognize its vocal emotions. This component divides the received voice signal into meaningful parts that will be used in voice feature extraction and voice emotion classification components.

Voice Feature Extraction

Once the voice is detected, the voice feature extraction component extracts a sufficient set of features from the voice of the learner. These features are considered as the significant features of the learner's voice and can be automatically extracted.

Voice Emotion Classification

We have used a similar emotion classification approach than with the facial emotion classification. This component analyses the voice stream and can extract a millisecond feature of each voice stream for its analysis. We used the sequential minimal optimization (SMO)[1] classifier of WEKA[2] software, which is a software tool for data mining.

2.3 Data Layer

The data layer is another separated layer within the FILTWAM. It physically stores the facial and the vocal datasets of the emotions. This layer reflects the intelligent capital of the system and provides a statistical reference for the detection of emotions.

2.4 Network Layer

The network layer uses the Internet to broadcast a live stream of the learner and to receive the feedback from the learner.

[1] http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/SMO.html

[2] http://www.cs.waikato.ac.nz/ml/weka

2.5 Application Layer

The application layer is the second most important part of FILTWAM. It consists of the game-based learning environment (e.g., EMERGO) and its two sub-components. The game-based learning environment uses the live stream of the facial and the vocal data of the learner to facilitate the learning process. Its sub-components named: the EMERGO rule engine and the EMERGO web service.

2.5.1 EMERGO Rule Engine

The EMERGO rule engine component manages didactical rules and triggers the relevant rules for providing feedback as well as tuned training content to the learner via the device. The game-based learning environment component complies with a specific rule-based didactical approach for the training of the learners.

2.5.2 EMERGO Web Service

The EMERGO web service component receives emotional data from EMERGO web service client component. It provides the training content and feedback to the learner through EMERGO rule engine component. At this stage, the learner can receive a feedback based on his facial and vocal emotions.

3 Method

Our hypothesis is that data gathered via webcam and microphone can be reliably used to unobtrusively infer learners' emotional states. Real-time multimodal emotion recognition of learners can be used for triggering more personalized feedback towards improved online learning. For example, it can be used during online game-based training of communication skills.

3.1 Participants

An e-mail was sent out to employees from Welten Institute at the Open University Netherlands to recruit the participants for this pilot study. The e-mail mentioned the estimated time investment for enrolling in the study. Twelve participants (7 male, 5 female; age M=42, SD=10) volunteered to participate. By signing an agreement form, the participants allowed us to capture their facial expressions and voice intonations, and to use their data for the study. No specific background knowledge was requested. They were told that they needed to do some tasks in which their input through a microphone and webcam would be used to help them to become more aware of their emotions.

3.2 Design

Participants were asked to expose seven basic face and voice expressions (happy, sad, surprise, fear, disgust, angry, and neutral) in four consecutive tasks. In this way, in total eighty face expressions and voice expressions of each participant were gathered. During this study, we offered very limited feedback to the participant, just the name of the recognized emotion and its prediction accuracy. In this way, the participant was informed whether or not our affective computing software detected the same 'emotion' as he was asked to 'mimic'.

In the first task participants were asked to mimic the face expressions while looking at the webcam, speak aloud and use the voice emotion that was shown on the face of the person that was on the presented image to them. There were 14 images subsequently presented through PowerPoint slides; the participant paced the slides. Each image illustrated a single emotion. All seven basic face expressions were two times presented with the following order: happy, sad, surprise, fear, disgust, angry, neutral, happy, sad, et cetera. In the second task, participants were requested to mimic the face expressions and to speak aloud the seven basic expressions twice: first, through the slides that each presented the keyword of the requested emotion and second, through the slides that each presented the keyword and the picture of the requested face and voice emotion with the following order: angry, disgust, fear, happy, neutral, sad, surprise. In total, 14 PowerPoint slides were used for the second task. For the first and the second task, participants could improvise and use their own texts. The third task presented 16 slides with the text transcript (both sender and receiver) taken from a good-news conversation. The text transcript also included instructions which face and voice expressions should accompany the current text-slide. Here, participants were requested to read and speak aloud the sender text of the 'slides' from the transcript and were asked to deliver the accompanying face and voice expressions. The forth task with 36 slides was similar to task 3, but in this case the text transcript was taken from a bad-news conversation. The transcripts and instructions for tasks 3 and 4 were taken from an existing OUNL training course [25] and a communication book [26].

3.3 Test Environment

All tasks were performed on a single Mac machine. The Mac screen was separated in three panels, top-left, top-right, and bottom. The participants could watch their facial expressions in the face emotion recognition module of the affective computing software at the top-left panel, they could watch their analyzed voice expressions in the voice emotion recognition module of the affective computing software at the top-right panel, while they were performing the tasks using a PowerPoint file in the bottom panel. An integrated webcam with a microphone and a 1080HD external camera were used to capture and record the emotions of the participants as well as their actions on the computer screen. The affective computing software with the face emotion recognition module and the voice emotion recognition module used the webcam and the microphone to capture and recognize the participants' emotions. The Silverback usability testing software version 2.0 used the external camera to capture and record the complete this experimental session. Figure 2 displays an output of both software modules and the PowerPoint slide for Task 3.

Fig 2.

3.4 Gathering Participants' Opinions

A self-developed questionnaire was used to collect participants' opinions after carrying out the requested tasks. All opinions were online collected via a Google form using 34 items on a 7- point Likert-scale with possible scores: 1) completely disagree, 2) disagree, 3) mildly disagree, 4) neither disagree nor agree, 5) mildly agree, 6) agree, and 7) completely agree. Participants' opinions were gathered for: 1) perceived difficulty to mimic the requested emotions in the given tasks, 2) perceived usefulness of the given feedback to mimic the emotions in the given tasks, 3) perceived instructiveness of the instructions for the given tasks, 4) perceived attractiveness of the given tasks, and 5) perceived concentration on the given tasks. Participants were also asked to report their self-assurance on 1) being able to mimic the requested emotions in the given tasks and 2) their acting skills on a similar 7-point Likert scale.

3.5 Procedure

Each participant signed the agreement form before his/her session of the study started. They individually performed all four tasks in a single session of about 30 minutes. The session was conducted in a completely silent room with a good lighting condition. The moderator of the session was present in the room, but did not intervene. All twelve sessions were conducted in two consecutive days. The participants were requested not to talk to each other in between sessions so that they could not influence each other. The moderator gave a short instruction at the beginning of each task. For example, participants were asked to show mild and not too intense expressions while mimicking the emotions. All tasks were recorded and captured by both the face emotion recognition module and the voice emotion recognition module of the affective computing software. After the session, each participant filled out the online questionnaire gathering participants' opinions.

4 Results and Discussion

The main purpose of this study focused on the validation of the software with respect to its multimodal detection of emotions. We asked two raters to analyze the recorded streams and carry out validation of the software modules. We do not report the output of our software modules for all tasks in detail in this study; instead we will first present the agreement between two raters, the requested emotions, and the recognized emotions by the two software modules separately.

4.1 Validation Results of the Software

The Kappa value for the validation of the face emotion recognition module based on the requested emotions, recognized emotion, and the raters' rating is reported in Table 1. The Kappa value for the validation of the voice emotion recognition module based on the requested emotions, recognized emotion, and the raters' rating is reported in Table 2.

Table 1. The Kappa value for the validation results of the face emotion recognition module for all the seven emotion for task 1, task 2, task3, and task 4.

Validation of the Recognized Emotion by the Face Emotion Recognition Software Module							
Happy	Sad	Surprise	Fear	Disgust	Angry	Neutral	Total
0.84	0.66	0.69	0.67	0.66	0.77	0.8	0.76

Analyzing of the Kappa statistic underlines the agreement among the raters, the requested emotions, and the face emotion recognition software module. The result with 95% confidence reveals that the interrater reliability was calculated to be Kappa = 0.76 (p <0.001). Therefore a substantial agreement among them is obtained based on Landis and Koch interpretation of Kappa values [27].

Table 2. The Kappa value for the validation results of the voice emotion recognition module for all the seven emotion for task 1, task 2, task3, and task 4.

Validation of the Recognized Emotion by the Voice Emotion Recognition Software Module							
Happy	Sad	Surprise	Fear	Disgust	Angry	Neutral	Total
0.63	0.50	0.51	0.48	0.41	0.50	0.71	0.58

Analyzing of the Kappa statistic underlines the agreement among the raters, the requested emotions, and the voice emotion recognition software module. The result with 95% confidence reveals that the interrater reliability was calculated to be Kappa = 0.58 (p <0.001). Therefore a moderate agreement is obtained among them.

The Kappa value for the validation of the face and the voice emotion recognition module based on the requested emotions, recognized emotion, and the raters' rating is reported in Table 3.

Table 3. The overall Kappa value for the validation results of the face and the voice emotion recognition software modules for all the seven emotion for task 1, task 2, task3, and task 4.

Validation of the Recognized Emotion by the Face and the Voice Emotion Recognition Software Modules							
Happy	Sad	Surprise	Fear	Disgust	Angry	Neutral	Total
0.68	0.50	0.53	0.50	0.43	0.55	0.73	0.61

Analyzing of the Kappa statistic underlines the agreement among the raters, the requested emotions, the face and the voice emotion recognition software modules. The result with 95% confidence reveals that the interrater reliability was calculated to be Kappa = 0.61 (p <0.001). Therefore a substantial agreement is obtained among them.

4.2 Multimodal Fusion of the Two Software Modules

The overall accuracy of our face emotion recognition software module is 75%, whereas it is 52% for the voice emotion recognition software module. In order to perform multimodal fusion between the two software modules, we selected linear weighted fusion method, which is a type of rule-based fusion and has been widely used [23]. As we used nominal data types (e.g. Happy and Angry) we do not need to normalize weights of different modalities in our linear fusion to combine the information. We performed a general formula $V_{m,} 1 \leq m \leq n$ to be a feature vector acquired from ***m***th software module, such as face and voice. We also let $w_{m,} 1 \leq m \leq n$ be the weight given to the ***m***th software module. Our vectors have the same dimensions and combined by using sum operator through this formula: $V = \frac{1}{n}\sum_{m=1}^{n} w_m \times V_m$. To combine the information of the two software modules we obtained a coefficient using this formula for each emotion category and we applied this coefficient into the validation results. This coefficient leads us to a higher accuracy rate for the combination of the two software modules. In case of differ in recognizing the emotions by the two software modules, the higher accuracy between the two modules prominent in all cases. For example, if the voice emotion recognition module returns disgust and the face emotion recognition module returns surprise, then the feedback 79%f will be given to the user. This indicates that the accuracy of 79% is based upon the recognized emotion by the face emotion recognition software module. Table 4 displays the new accuracy using the multimodal fusion and the rules on how the feedbacks are generated.

Table 4. The overall accuracy using the multimodal fusion between the two software modules.

		Face Emotion Recognition						
		Happy	Sad	Surprise	Fear	Disgust	Angry	Neutral
Voice Emotion Recognition	Happy	89.7%	56%f	79%f	58%f	75%f	83%f	90%f
	Sad	80%f	74.4%	79%f	66%v	75%f	83%f	90%f
	Surprise	80%f	56%f	67.8%	58%f	75%f	83%f	90%f
	Fear	80%f	56%f	79%f	50.3%	75%f	83%f	90%f
	Disgust	80%f	56%f	79%f	58%f	63%	83%f	90%f
	Angry	80%f	63%v	79%f	63%v	75%f	100%	90%f
	Neutral	80%f	61%v	79%f	61%v	75%f	83%f	100%

The uniform distribution of emotions is the average of the diagonal: 78% (based on Table 4).

4.3 Participants' Opinions

Here we report the results of the online questionnaire gathering participants' opinions for various aspects (see Table 5). The results indicate that all tasks were regarded moderately difficult and interesting to do. Participants were satisfied with the clarity of the instructions and thought that the feedback was pretty helpful to them. The self-assurance factor was not high among the participants. The results for the concentration factor indicated that participants experienced no distraction during their performance. It can be easily seen that the participants did not regard themselves as actors.

Table 5. The participants' opinions (n = 12).

		Answers by the Participants							
		1	2	3	4	5	6	7	Total
Questions	Difficulty								100%
	It was easy for me to mimic the requested emotions in the given tasks	----	8%	----	33%	33%	26%	----	
	Feedback								
	The feedback did help me to mimic the emotions in the given tasks	----	----	----	17%	17%	50%	16%	
	Self-assurance								
	I am confident that I was able to mimic the requested emotions in the given tasks	3%	3%	48%	10%	16%	20%	----	
	Instructiveness								
	The instructions for the given tasks were clear to me	----	----	8%	----	25%	42%	25%	
	Attractiveness								
	The given tasks were interesting	----	----	----	----	17%	75%	8%	
	Concentration								
	I could easily focus on the given tasks and was not distracted by other factors	----	----	----	----	16%	46%	38%	
	Acting skills								
	I regard myself as a good actor	----	42%	24%	17%	----	17%	----	

1= Completely disagree, 2= Disagree, 3= Mildly disagree, 4= Neither disagree nor agree, 5= Mildly agree, 6= Agree, and 7= Completely agree

5 Conclusion

The FILTWAM framework aims at real-time interpretation of multimodal emotional behavior into emotional states that can be used for better game-based learning. This study examined a multimodal fusion approach for real-time face emotion recognition and voice emotion recognition modules that are part of the FILTWAM framework. We have also examined the two software modules separately. The overall accuracy of our face emotion recognition software based on the requested emotions and the recognized emotions is 75%. The overall accuracy of our voice emotion recognition software based on the requested emotions and the recognized emotions is 52%. This is in accordance with [28, 29]. Compare to our previous study [5], the accuracy of our voice emotion recognition dataset improved from 22.2% to 50%. The overall accuracy of our multimodal fusion method is 78% for combination of the two software modules, which falls into the same range as reported in [22].

The results of the questionnaire indicate participants' low self-confidence on being a good actor and the self-assurance factor that was not high among the participants. This issue clearly hints towards an improved feedback mechanism that will be dealt in our upcoming study. We only recruit twelve participants from middle age group volunteered to participate in this study. The possible outcome might be different for younger and older participants. FILTWAM forms part of our research that aims at improving multimodal emotion recognition for better online game-based learning. In this, a future study will use it for researching its suitability towards improved skill acquisition in the context of an online game-based training for communication skills using EMERGO [13].

Acknowledgments. We thank our colleagues at Welten Institute of the Open University Netherlands who participated in the integration of the face and voice emotion recognition study. We likewise thank the two raters who helped us to rate the recorded streams. We also thank the Netherlands Laboratory for Lifelong Learning (NELLL) of the Open University Netherlands that sponsors this research.

References

1. Anaraki, F.: Developing an Effective and Efficient eLearning Platform. International Journal of The Computer, the Internet and Management. Vol. 12(2):57-63. (2004)
2. Hrastinski, S.: Asynchronous & synchronous e-learning. Educause Quarterly. Vol. 31(4):51-55. (2008)
3. Pekrun, R.: The impact of emotions on learning and achievement: towards a theory of cognitive/motivational mediators. Journal of Applied Psychology. Vol. 41:359–376. (1992)
4. Bahreini, K., Nadolski, R., Qi, W., Westera, W.: FILTWAM - A Framework for Online Game-based Communication Skills Training - Using Webcams and Microphones for Enhancing Learner Support. In P. Felicia (Ed.), The 6th European Conference on Games Based Learning (ECGBL). Cork, Ireland. p. 39-48. (2012)
5. Bahreini, K., Nadolski, R., Westera, W.: FILTWAM and Voice Emotion Recognition. Games and Learning Alliance (GaLA) Conference 2013. Paris. 23-25 October. (2013)
6. Bahreini, K., Nadolski, R., Westera, W.: FILTWAM - A Framework for Online Affective Computing in Serious Games. The 4th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES'12). Procedia Computer Science. Genoa, Italy. Vol. 15:45-52. (2012)
7. Kelle, S., Sigurðarson, S., Westera, W., Specht, M.: Game-Based Life-Long Learning. In G. D. Magoulas (Ed.), E-Infrastructures and Technologies for Lifelong Learning: Next Generation Environments. Hershey, PA: IGI Global. p. 337-349. (2011)

8. Reeves, B., Read, J.L.: Total engagement: Using games and virtual worlds to change the way people work and business compete. Boston. Harvard Business Press. (2009)
9. Gee, J.P.: What video games have to teach us about learning and literacy. New York: Palgrave Macmillan. (2003)
10. Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., Boyle, J. M.: A systematic literature review of empirical evidence on computer games and serious games. Computers and Education. September. Vol. 59(2):661-686. (2012)
11. Van Merrienboer, J.J.G., Kirschner, P.A.: Ten Steps to complex learning. A systematic approach to four-component instructional design. New York: Routledge. (2007)
12. Hager, P. J., Hager, P., Halliday, J.: Recovering Informal Learning: Wisdom, Judgment And Community. Springer. (2006)
13. Nadolski, R. J., Hummel, H. G. K., Van den Brink, H. J., Hoefakker, R., Slootmaker, A., Kurvers, H., Storm, J.: EMERGO: methodology and toolkit for efficient development of serious games in higher education. Simulations & Gaming. Vol. 39(3):338-352. (2008)
14. Bashyal, S., Venayagamoorthy, G.K.: Recognition of facial expressions using Gabor wavelets and learning vector quantization. Engineering Applications of Artificial Intelligence. (2008)
15. Ekman, P., Friesen, W. V.: Facial Action Coding System: Investigator's Guide. Consulting Psychologists Press. (1978)
16. Kanade, T.: Picture processing system by computer complex and recognition of human faces. PhD thesis. Kyoto University, Japan. (1973)
17. Petta, P., Pelachaud, C., Cowie, R.: Emotion-Oriented Systems. The Humaine Handbook. Springer-Verlag. Berlin. (2011)
18. Chen, L.S.: Joint Processing of Audio-visual Information for the Recognition of Emotional Expressions in Human-computer Interaction. University of Illinois at Urbana-Champaign. PhD thesis. (2000)
19. Sebe, N., Cohen, I. I., Gevers, T., Huang, T. S.: Emotion recognition based on joint visual and audio cues. International Conference on Pattern Recognition. Hong Kong. p. 1136-1139. (2006)
20. Song, M., Bu, J., Chen, C., Li, N.: Audio-visual based emotion recognition: A new approach. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 2. (2004)
21. Zeng, Z., Pantic, M., Roisman, G. I., Huang, T. S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31(1):39–58. (2009)
22. Sebe, N.: Multimodal Interfaces: Challenges and Perspectives. Journal of Ambient Intelligence and Smart Environments. January. Vol. 1(1):23-30. (2009)
23. Atrey, P. K., Hossain, M. A., El Saddik, A., Kankanhalli, M.: Multimodal fusion for multimedia analysis: a survey. Multimedia Systems. Vol. 16(6):345-379. Springer-Verlag. (2010)
24. Saragih, J., Lucey, S., Cohn, J.: Deformable Model Fitting by Regularized Landmark Mean-Shifts. International Journal of Computer Vision (IJCV). (2010)
25. Lang, G., van der Molen, H. T.: Psychologische gespreksvoering. Open University of the Netherlands. Heerlen, The Netherlands. (2008)
26. Van der Molen, H. T., Gramsbergen-Hoogland, Y. H.: Communication in Organizations: Basic Skills and Conversation Models. ISBN 978-1-84169-556-3. Psychology Press, New York. (2005)
27. Landis, J. R., Koch, G. G.: The measurement of observer agreement for categorical data. Biometrics. Vol. 33:159-174. (1977)
28. Vogt, T. André, E. Bee, N.: EmoVoice - A framework for online recognition of emotions from voice. In Proceedings of Workshop on Perception and Interactive Technologies for Speech-Based Systems. (2008)

29. Dai, K., Harriet J. F., MacAuslan, J.: Recognizing emotion in speech using neural networks. Telehealth and Assistive Technologies. p.31-38. (2008)

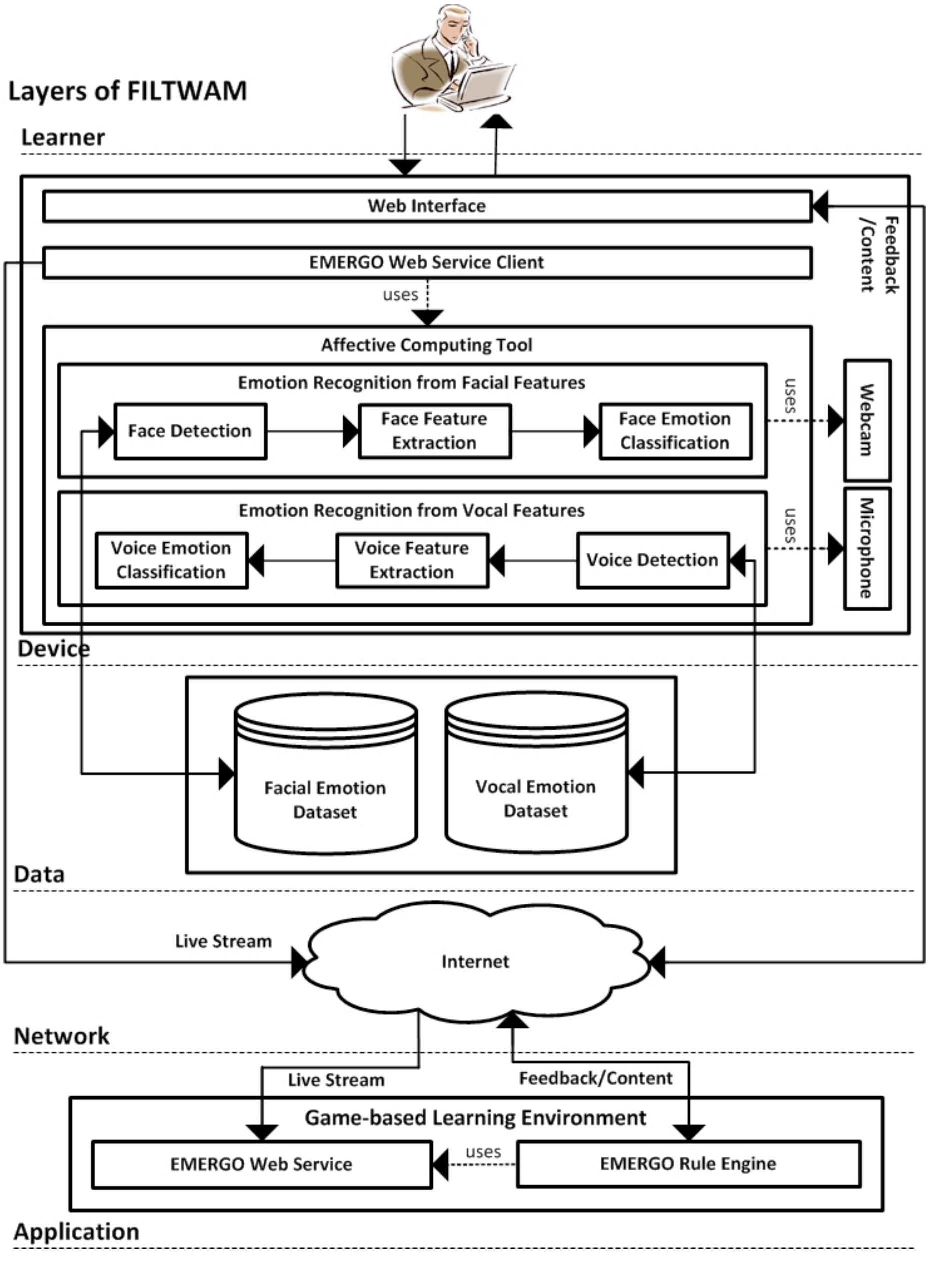
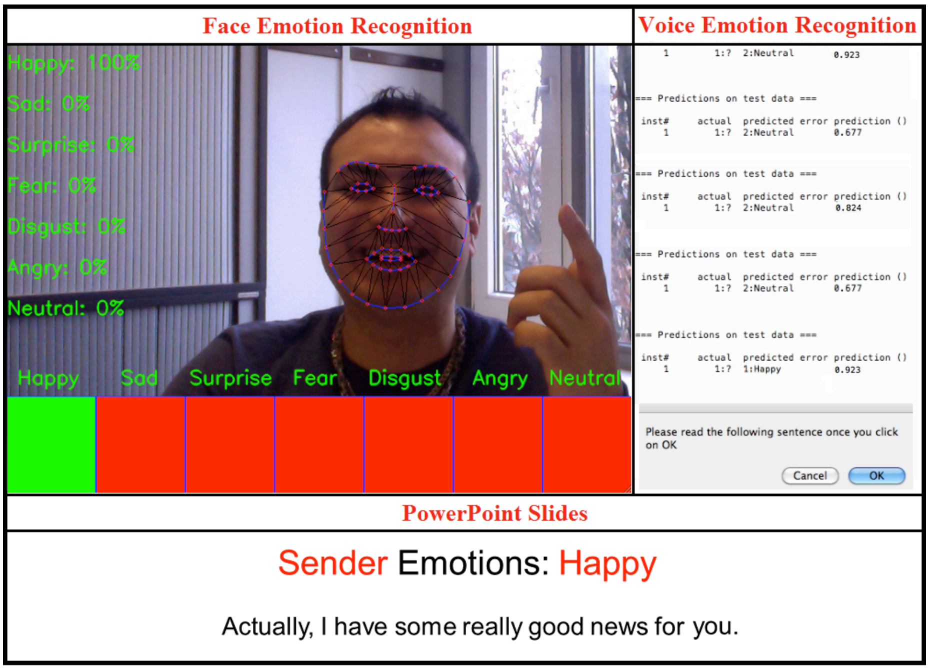


Fig. 2. The FILTWAM framework integrates the face emotion recognition module and the voice emotion recognition module in an online game-based environment. The face emotion recognition sub-component and the voice emotion recognition sub-component have been reported in our previous studies [4, 5, 6]).



Face Emotion Recognition

Voice Emotion Recognition

```
     1         1:?  2:Neutral        0.923


=== Predictions on test data ===

 inst#     actual  predicted error prediction ()
     1         1:?  2:Neutral        0.677

=== Predictions on test data ===

 inst#     actual  predicted error prediction ()
     1         1:?  2:Neutral        0.824


=== Predictions on test data ===

 inst#     actual  predicted error prediction ()
     1         1:?  2:Neutral        0.677


=== Predictions on test data ===

 inst#     actual  predicted error prediction ()
     1         1:?  1:Happy          0.923
```

Please read the following sentence once you click on OK

PowerPoint Slides

Sender Emotions: Happy

Actually, I have some really good news for you.

Fig. 2. The main researcher in task 3, the affective computing software including the face emotion recognition module (top-left) and the voice emotion recognition module (top-right), and the PowerPoint slide (bottom).