FILTWAM - A Framework for Online Affective Computing in Serious Games

Kiavash Bahreini, Rob Nadolski, and Wim Westera

Centre for Learning Sciences and Technologies (CELSTEC), Open University of the Netherlands (OUNL)
Heerlen, The Netherlands
{kiavash.bahreini, rob.nadolski, wim.westera}@ou.nl

***Abstract*—This paper introduces a Framework for Improving Learning Through Webcams And Microphones (FILTWAM). It proposes an overarching framework comprising conceptual and technical frameworks for enhancing the online communication skills of lifelong learners. Our approach interprets the emotional state of people using webcams and microphones and combines relevant and timely feedback based upon learner's facial expressions and verbalizations (like sadness, anger, disgust, fear, happiness, surprise, and neutral). The feedback generated from the webcams is expected to enhance learner's awareness of their own behavior. Our research enhances flexibility and scalability in contrast with face-to-face trainings and better helps the interests of lifelong learners who prefer to study at their own pace, place and time. Our small-scale proof of concept study exemplifies the practical application of FILTWAM and provides first evaluation results on that. This study will guide future development of software, training materials, and research. It will validate the use of webcam data for a real-time and adequate interpretation of facial expressions into emotional states. Participants' behaviour is recorded on videos so that videos will be replayed, rated, annotated and evaluated by expert observers and contrasted with participants' own opinions in future research.**

Keywords: Communication skills; affective computing; web-based training; lifelong learning; serious gaming

I. INTRODUCTION

Communication skills become more prominent in our knowledge society as they used to be in the past. More jobs require more skilled people with respect to communication skills in our time. This is not only for specific kind of jobs, but it is throughout all jobs [1]. The purpose of this research is to investigate novel training approaches for improving communication skills of everyone who has already finished their formal education. Communication skills are a lifelong affair for all members of our society. Even for more recent educated people with respect to communication skills it is important that they keep on improving their level of communication skills.

Two significant factors influence on learning procedure these days. The shortage of trainers who can provide communication skills for face-to-face situations and inflexible training programs that can force learners to attend to specific courses for face-to-face training. This approach influences freedom of place and time of learners [2]. An alternative for this approach is lifelong learning that requires flexible training programs in which learners can practice a lot on a regular basis to improve their communication skills. As lifelong learners often have a job, they have a lot of things to do, and they are probably not too much inclined in spending time for learning additional things. Games can make the learning process more enjoyable for them. We expect that using technology, affective computing tool, and a web-based training system might be insufficient to encourage people to improve their communication skills. For this purpose, FILTWAM is deployed with a game-based didactical approach. Our framework (FILTWAM) offers a smooth setting for learners to improve their communication skills at their own pace, place, and time, although it is not a replacement of the face-to-face training. They might still require face-to-face meeting with knowledgeable experts at specific points of time. We address communication behaviour rather than communication content, as people mostly do not have problems with the "what" but with the 'how" in expressing their message. We use insights from face-to-face training, game-based learning, lifelong learning, and affective computing. These areas constitute starting points for moving ahead the not yet well-established area of using emotional states for improved learning. Our framework and research is situated within this latter area. An independent web-based training enhances flexibility and scalability in contrast with face-to-face trainings.

FILTWAM uses devices, such as mobile phones, laptops, tablets, for learners' communication and comprises an affective computing tool with combining two modalities into a single system for face and voice emotion recognition. It uses webcams and microphones that continuously and unobtrusively collect learners' data and interpret learners' emotional behaviour into emotional states. The feedback generated from FILTWAM is expected to enhance learner's awareness of their own behaviour as well as to improve the alignment between their expressed behaviour and intended behaviour. The facial emotion recognition detects faces, recognizes seven basic face expressions, and provides adequate feedback to the learners. The affective computing tool is built upon existing research [3, 4, 5, 6, 7, 8, and 9]. It offers face detection, face recognition, and face emotion recognition functions that are not new and have been studied in the past. The basic idea behind the affective computing tool and linking two modalities into a single system for affective computing analysis is also not new and studied in [10, 11, 12, 13, and 14]. A more recent, survey review by Sebe [15] shows that the accuracy of detecting one

or more basic emotions is greatly improved when both visual and audio information are used in classification, leading to accuracy levels between 72% to 85%. This hints at combining both visual and audio data for inferring emotion and providing adequate and timely feedback, exactly the reason why this research proposes combining face expression and voice intonation when triggering support during online communication skills training.

To characterize the novelty of our work, we propose a multimodal framework that in real-time interprets emotional behaviour into emotional states. Furthermore, this is applied in educational settings, more precise for soft-skills training purposes. To our knowledge, these approaches have not yet been integrated in any other frameworks. In this paper, section 2 introduces the overarching framework and its sub-frameworks including the conceptual framework and the technical framework. The affective computing tool is also described in section 2. Method and proof of concept as well as result and validation are explained in section 3. Section 4 discusses the findings and provides suggestions for future work.

II. OVERARCHING FRAMEWORK

A. FILTWAM

The overarching framework aims to improve learners' communication skills by providing timely and adequate feedback to the learner exploiting learners' state data, which are gathered through webcam and microphone, as well as by learner input (like keyboard, mouse) when interacting with the online game-based training materials. It provides a fruitful environment for the learners to improve their communication skills using an online game-based training approach.

B. Conceptual Framework

The conceptual framework encompasses four components: 1) Learner, 2) Device, 3) Affective computing tool, and 4) Rules engine. The two latter components are situated within the game-based communication skills training area (see Figure 1). The learner is a lifelong learner who is positively biased towards the paradigm of informal learning and who prefers to study at his own pace, place and time. Figure 1 illustrates how the component 'device' provides the real-time feedback to the learner. The component 'device' could be a personal computer, a laptop, or a smart device by which the learner interacts with the affective computing tool component. The game-based communication skill training encompasses affective computing tool and rules engine components.

The affective computing tool consists of two components, each of which has its own sub-components. The components are emotion recognition from facial features and emotion recognition from vocal features. Both consist of four sub-components: 1) face/voice detection, 2) facial feature/vocal intonation extraction, 3) facial/vocal emotion classification, and 4) facial/vocal emotion dataset. Compared to most previous frameworks, our approach considers facial recognition and vocal recognition in one single framework and intends to perform the operations in real-time. The latter aspect and its application for soft-skill training purposes characterize the novelty of FILTWAM.

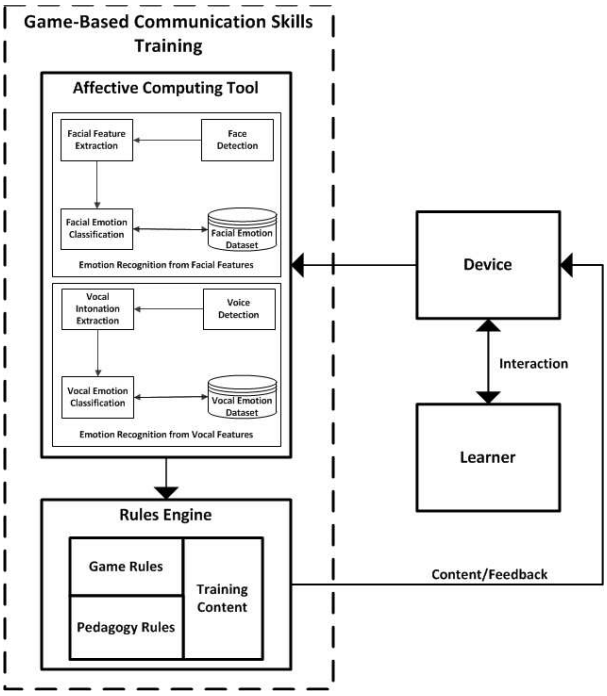

Figure 1. Conceptual framework for online game-based communication skills training.

The process of emotion recognition from facial features starts at face detection component. The face detection is part of face acquisition in contemporary emotion recognition tools [10]. Given a video stream, detecting the learner's face is completed in this component. The variations of the face, poses, angles, and sizes make this step a challenging task. Once the face is detected, the face detection component sends it to the facial feature extraction component to extract 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 extracted automatically. In the past thirty years, most of the emotion classification approaches were introduced by Ekman and focused on classifying the six basic emotions. The facial emotion classification component supports classification of these six basic emotions plus neutral emotion, but can in principle also recognize other or more detailed face expressions when required. This component analyses video sequences and provides images corresponding to each frame to be extracted. First frame of each 25 frames per second is used for this purpose. This component is independent of race, age, gender, hairstyles, glasses, background, and beard. It compares the classified emotions with existing emotions in the facial emotion dataset and trains the dataset using a number of learners' faces. The not yet developed sub components of the emotion recognition from vocal features are exactly the same as the sub components of the emotion recognition from facial features described earlier. The tool then synchronizes and analyses the facial and vocal expressions and transmits the

results to the rules engine component. The output of this component triggers the relevant rules as well as the training content in rules engine component.

The game rules, pedagogy rules, and training content are the significant sub components within the rules engine component. Rules engine is responsible to register and manage all the rules and defines the relationships between different rules. It delivers training contents based upon its defined rules. The pedagogy rules define the instructive methods of instruction. The game rules component filters the data and provides the training content based upon the learners' emotional state. It transmits the generated feedback to the device component to be sent to the learner.

C. Technical Framework

The technical framework contains five layers: 1) Learner layer, 2) Device layer, 3) Network layer, 4) Web server layer, and 5) Data layer. The two first components are described in the conceptual framework. Figure 2 illustrates the technical framework, its layers and components.

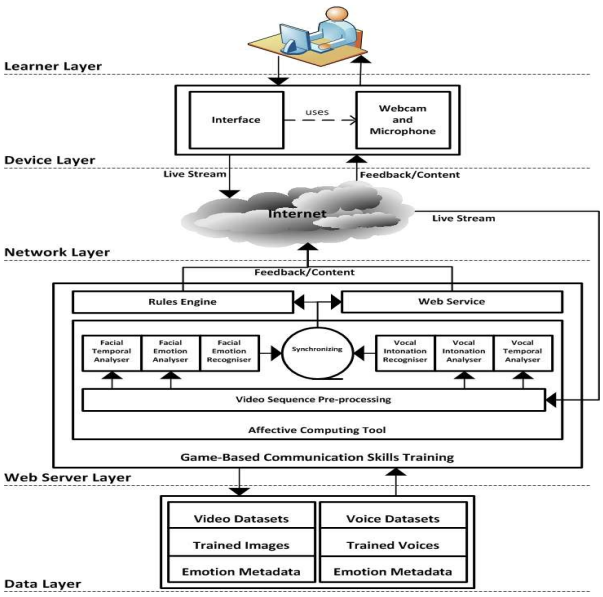

Figure 2. Technical framework of FILTWAM.

A learner opens the GUI of the affective computing tool that uses webcams and microphones. The device provides the video sequence stream of the learner and broadcasts it over the Internet. The provided stream feeds into the video sequence pre-processing component to split it frame by frame. It calculates a time-sequence of the emotions within a period of time and recognizes an emotion within a particular frame. The facial emotion analyser and the vocal intonation analyser take each split frame and analyse the related emotion/intonation within the particular frame. The outputs of these components send to the facial emotion/vocal intonation recognizer components, respectively. The facial temporal/vocal analyser components analyses the video sequence stream and provides timely feedback. These components call the facial temporal/vocal intonation analyser components for each frame and calculate summation of all the different amounts of time of emotions. The combined feedback results are synchronized and transmitted to the learner through the rules engine or the web service components. The recognition process is not complete unless the video/voice datasets, trained images/voices, and emotion metadata of both face and voice components are created in data layer.

III. METHOD AND PROOF OF CONCEPT

A. Participants

Sixteen participants, all employees from the Centre for Learning Sciences and Technologies (CELSTEC) of Open University of the Netherlands volunteered to participate in this experiment. They were asked to participate in the experiment, which helped them to be more aware of their emotions while they were communicating through a webcam and microphone. The experiment was individually conducted.

B. Design

We developed a simple feedback mechanism in our tool with red/green signals to inform the learner whether the software detects the same 'emotion' as the participant was asked to 'mimic'.

C. Tasks

Five consecutive tasks were given to the participants: 1) train the database of the affective computing software by exposing seven basic face expressions, 2) mimic the emotion that was presented through PowerPoint slides. There were 35 images presented after each other; each image illustrated a single emotion with the following order: happy, sad, surprise, fear, disgust, angry, neutral, happy, sad, …), 3) mimic the seven face expressions two times with the following order: angry, disgust, fear, happy, neutral, sad, surprise, angry,…, 4) read and speak aloud the sender 'slides' of transcript taken from a good-news conversation, 5) as in task 4, but in this case the text transcript was taken from a bad-news conversation. The transcripts and instructions for tasks 4 and 5 were taken from an existing OUNL training course [16] and a communication book [17].

D. Test environment/measurement instruments

All tasks were performed on a single Mac machine. The Mac screen was separated in two panels, left and right. The participants could watch their facial expressions in the affective computing software at the left panel, while they were performing the tasks using a PowerPoint file in the right panel. An integrated webcam 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 used the webcam to capture and recognize the participants' emotions, while Silverback usability testing software version 2.0 used the external camera to capture and record the complete experimental session. The recorded video files on the windows machine will be used for our future video

analysis purposes. Figure 3 demonstrates an output of the software and an experimental session for Task 5.

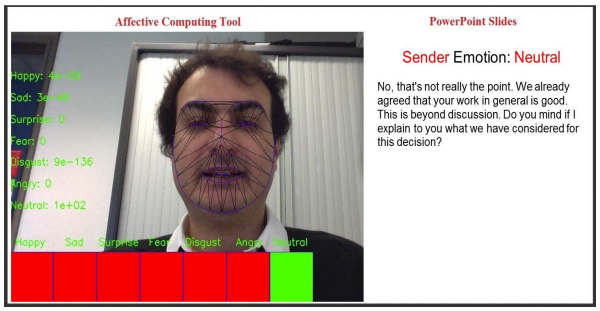

Figure 3. A participant in task 5 and the affective computing software during the experimental session.

E. *Procedure*

The participants invited to participate in this experiment by an email. They performed each individual session in about 20 minutes. They sat in a completely silent room with good lighting condition. The moderator of the session presented in the room, but with no intervention. Each participant was asked to do task one up till five in a row and during one individual session preceded by a short instruction at the beginning of each task. Participants were asked to show mild and not too intense expressions while mimicking the emotions. At the end of a session the participants stated their viewpoints about the software, the problems they encountered, and the given tasks.

F. *Result and validation*

In this paper we report the validation of the software for the third task. Table 1 shows the results of the requested emotions from participants and compares the results with recognized emotions by the software. The obtained false results are not produced just because of the software malfunctioning, but in many cases the participants were also unable to mimic the requested emotions.

TABLE I. VALIDATION RESULT FOR TASK 3

Requested Emotion \ Recognized Emotion	Happy	Sad	Surprise	Fear	Disgust	Angry	Neutral	Total
Happy	71.875	3.125			18.75		6.25	100
Sad		31.25	3.125	12.5	25	6.25	21.875	100
Surprise		3.125	71.875	9.375	9.375		6.25	100
Fear		6.25	18.75	46.875	3.125	3.125	21.875	100
Disgust	6.25	3.125			62.5	15.625	12.5	100
Angry		9.375		9.375	28.125	40.625	12.5	100
Neutral		3.125	6.25	6.25	9.375	6.25	68.75	100

TABLE II. TABLE 2. VALIDATION RESULT FOR TASK 4 (HALF OF THE PARTICIPANTS ARE REPORTED)

Req \ Recognized Emotion	Happy	Sad	Surprise	Fear	Disgust	Angry	Neutral	Total
Happy	73.34	------	20	3.33	3.33	------	------	100
Neutral	------	10.39	10.39	6.5	------	------	72.72	100

IV. CONCLUSION

We propose a multimodal framework that in real-time interprets emotional behaviour into emotional states, is applied in educational settings, and is more precise for soft-skills training purposes. The results showed that the majority of the participants were able to accurately use the software; however they were not fully aware of their emotions to mimic. The participants did not have any problems to mimic happy and neutral emotions, but they had a lot of problems to mimic other emotions. Almost all participants forgot how they trained the software in initial stage; therefore there is a need for a new feedback solution to facilitate this process. A new version of the software for face emotion recognition will be developed, whereas voice emotion recognition functionality will be gradually further developed. The recorded videos will be replayed, rated, annotated and evaluated by expert observers and contrasted with participants' own opinions in future.

ACKNOWLEDGMENT

We thank Jason Saragih for permission to develop our software based on his face tracker software [18].

REFERENCES

[1] C. P. Brantley and M. G. Miller, Effective Communication for Colleges, Thomson Higher Education, 2008.

[2] P. J. Hager, P. Hager, and J. Halliday, "Recovering Informal Learning: Wisdom, Judgment And Community", Springer, 2006.

[3] S. Avidan and M. Butman, "Blind vision", European Conference on Computer Vision, vol. 3953, 2006, pp. 1–13.

[4] S. Bashyal and G.K. Venayagamoorthy, "Recognition of facial expressions using Gabor wavelets and learning vector quantization", Engineering Applications of Artificial Intelligence, 2008.

[5] C. C. Chibelushi and F. Bourel, "Facial expression recognition: a brief tutorial overview", Available Online in Compendium of Computer Vision, 2003.

[6] P. Ekman and W. V. Friesen, "Facial Action Coding System: Investigator's Guide", Consulting Psychologists Press, 1978.

[7] T. Kanade, "Picture processing system by computer complex and recognition of human faces", PhD thesis, Kyoto University, Japan, 1973.

[8] S. Z. Li and A. K. Jain, Handbook of Face Recognition Second Edition, ISBN 978-0-85729-931-4, Springer-Verlag, London, 2011.

[9] P. Petta, C. Pelachaud, and R. Cowie, Emotion-Oriented Systems, The Humaine Handbook, ISBN 978-3-642-15183-5, Springer-Verlag, Berlin, 2011.

[10] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots", Robotics and Autonomous Systems, 42(3-4), 2003, pp. 143-166.

[11] L. S. Chen, "PhD thesis", Joint Processing of Audio-visual Information for the Recognition of Emotional Expressions in Human-computer Interaction, University of Illinois at Urbana-Champaign, 2000.

[12] N. Sebe, I. I. Cohen, T. Gevers, and T. S. Huang, "Emotion recognition based on joint visual and audio cues", International Conference on Pattern Recognition, pp 1136-1139, Hong Kong, 2006.

[13] M. Song, J. Bu, C. Chen, and N. Li, "Audio-visual based emotion recognition: A new approach", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Volume 2, 2004.

[14] R. Subramanian, J. Staiano, K. Kalimeri, N. Sebe, and F. Pianesi, "Putting the Pieces Together: Multimodal Analysis of Social Attention in Meetings", ACM Multimedia, Firenze, Italy, 2010.

[15] N. Sebe, "Multimodal Interfaces: Challenges and Perspectives", Journal of Ambient Intelligence and Smart Environments, January, Vol. 1, No. 1, pp 23-30, 2009.

[16] G. Lang and H. T. van der Molen, “Psychologische gespreksvoering”, Open University of the Netherlands, Heerlen, The Netherlands, 2008.

[17] H. T. van der Molen and Y. H. Gramsbergen-Hoogland, “Communication in Organizations: Basic Skills and Conversation Models”, ISBN 978-1-84169-556-3, Psychology Press, New York, 2005.

[18] J. Saragih, S. Lucey, and J. Cohn, “Deformable Model Fitting by Regularized Landmark Mean-Shifts”, International Journal of Computer Vision (IJCV), 2010.