FILTWAM - A Framework for Online Game-based Communication Skills Training - Using Webcams and Microphones for Enhancing Learner Support

Kiavash Bahreini, Rob Nadolski, Wen Qi, and Wim Westera
Centre for Learning Sciences and Technologies, Open University of the Netherlands, Heerlen, The Netherlands
*{kiavash.bahreini,*rob.nadolski,wen.qi,wim.westera}@ou.nl

Abstract: This paper provides an overarching framework embracing conceptual and technical frameworks for improving the online communication skills of lifelong learners. This overarching framework is called FILTWAM (Framework for Improving Learning Through Webcams And Microphones). We propose a novel web-based communication training approach, one which incorporates relevant and timely feedback based upon learner's facial expressions and verbalizations. This data is collected using webcams with their incorporated image and microphones with their sound waves, which can continuously and unobtrusively monitor and interpret learners' emotional behaviour into emotional states. The feedback generated from the webcams 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. Our approach emphasizes communication behaviour rather than communication content, as people mostly do not have problems with the "what" but with the 'how" in expressing their message. For our design of online game-based communication skills trainings, 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. A self-contained game-based training enhances flexibility and scalability, in contrast with face-to-face trainings. Furthermore, game-based training better serve the interests of lifelong learners who prefer to study at their own pace, place and time. In the future we may possibly integrate the generated feedback with EMERGO, which is a game-based toolkit for delivery of multimedia cases.
Finally, we will report on a small-scale proof of concept study that on the one hand exemplifies the practical application of our framework and on the other hand provides first evaluation results on that. This study will guide further development of software and training materials and inform future research. Moreover, it will validate the use of webcam data for a real-time and adequate interpretation of facial expressions into emotional states (like sadness, anger, disgust, fear, happiness, and surprise). For this purpose, participants' behaviour is also recorded on videos so that videos will be replayed, rated, annotated and evaluated by expert observers and contrasted with participants' own opinions.

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

1. Introduction

In the twenty first century, we live in a knowledge society in which communication skills become more important as they used to be in the past. More jobs require more skilled people with respect to communication skills. This is not only for specific kind of jobs, but it is throughout all jobs (Brantley & Miller, 2008). For example, thirty years ago a system engineer could almost work by him or herself and this is not the case anymore in the knowledge society era. There are a lot of people that have been educated in an area that communication was not that important in society as it is today. This clarifies why such people might have a lack of communication skills and need to improve these skills.

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. It is obvious that the job market demands people who already have the necessary communication skills but are also willing to improve such skills. Lifelong learning requires flexible training programs in which learners can practice a lot on a regular basis in order to improve their communication skills. Nowadays, such learners must attend specific courses that use face-to-face approach, turning them into quite inflexible training programs as far as freedom of place and time is concerned. An additional problem is the shortage of trainers that can provide communication skills for face-to-face situations (Hager, Hager, & Halliday, 2006).

We propose an online web-based communication skills training framework (FILTWAM). This is not a replacement of the face-to-face training, but intends to offer a smooth setting for learners to improve their

communication skills at their own pace, place, and time. This online environment offers a lot of flexibility and freedom to learners. However learners might still require face-to-face meeting with knowledgeable experts at specific points of time. Improving communication skills requires timely and adequate feedback that can be provided by FILTWAM. 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. The affective computing tool continuously and unobtrusively collects learners' data using a webcam and a microphone and interprets 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 component detects faces, recognizes seven basic face expressions (happy, sad, surprise, fear, angry, disgust, and neutral), and provides adequate feedback to the learners. The affective computing tool is built upon existing research (Avidan & Butman, 2006; Bashyal & Venayagamoorthy, 2008; Chibelushi & Bourel, 2003; Ekman & Friesen, 1978; Kanade 1973; Li & Jain, 2011; Petta, Pelachaud, & Cowie, 2011). 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 (Chen 2000; Fong, Nourbakhsh, & Dautenhahn, 2003; Sebe et al., 2006; Song et al., 2004; Subramanian et al., 2010). A more recent, survey review by Sebe (2009) 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, precisely the reason why this research proposes combining face expression and voice intonation when triggering support during web-based communication skills training.

This research proposes online game-based learning as an overarching training approach for communication skills. Game-based learning has several advantages: 1) it is a didactical approach that seems to be in-line with the lifelong learners' interests, 2) it is very popular nowadays, and 3) it seems very fruitful to people (Connolly et al., 2012). Various authors anticipate the great opportunities of games (and simulations) in formal as well as in informal education, because of positive effects on learning outcomes (see, e.g. Aldrich, 2009; Amory, 2007; Kiili, 2005; Quinn, 2005). Game-based learning can be very effective for skills training (see e.g., Reeves & Read, 2009) and motivating (Gee, 2003; Gee & Levine, 2009). This didactic approach focusses on learning instead of entertainment, can be quite informal and offers a lot of possibilities to the lifelong learners on video and audio bases. Games can be motivating and interesting for the learners. 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 pure 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. Each gaming environment including our online gaming-based training system is defined by its rules, which need to be managed by a rules engine. The rules engine of our system registers and manages all the rules and defines the relationships between different rules being either game rules or pedagogy rules. It delivers training (assignment and content) based upon such rules. Within FILTWAM, the pedagogy rules define the instructive strategies the methods of instruction (assignments) as well as the content of instruction. At run-time the learner's states need to establish for triggering one or more of such rules. We may use EMERGO as the game-based engine in order to manage the rules and the contents.

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.

2. Overarching framework - 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. The overarching framework comprises a conceptual framework and a technical framework. The interplay of both frameworks is needed to provide the learner with adequate feedback. FILTWAM offers a lot of flexibility and freedom to learners. It comprises an affective computing tool with combining two modalities into a single system for face and voice emotion recognition. The affective computing tool interprets learners' emotional behaviour into emotional states in real-time. At the conceptual level, the rules engine component deals with triggering specific feedback messages (i.e., content of feedback). The game-based training component meets the didactical approach and gaming mechanism to train the learners. The rules engine manages game rules, pedagogy rules, and influences training content. It provides desire feedback based upon the rules. The technical framework is the expanded version of the conceptual framework and provides a blueprint for the realisation of FILTWAM. It represents different layers and the physical components. Complete details of the frameworks and illustrations of the relevant layers and components are described in the next sections.

2.1 Conceptual framework

This conceptual framework meets our hypotheses described earlier. 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). There other two components are depicted in the right side of this Figure. 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 and where the learner is situated in the framework and 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 is the heart of FILTWAM. It 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) the 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.

The process of emotion recognition from facial features starts at face detection component. We do not intent to know whose face it is; instead we intend to recognize facial emotions. The face detection is part of face acquisition in contemporary emotion recognition tools (Li & Jain, 2011). 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 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 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. This sub component has not yet been developed and will be implemented and tested in future research. The tool then synchronizes and analyses the facial and vocal expressions and transmits the results to the rules engine component, which might be EMERGO. 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.

2.2 Technical framework

The technical framework encompasses five layers and a number of sub components within the layers. The five layers are introduced as the: 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.

Suppose that a learner intends to improve his communication skills while interacting with FILTWAM. He first uses a dedicated application (i.e. webcam/microphone software) to send his facial emotions and vocal intonations to the training software. He opens the application interface that uses an embedded webcam and microphone or utilizes external peripheral equipment. He permits the application utilizing his video sequence stream and transmitting his facial emotions and vocal intonation to third layer called network layer. This layer is responsible for digital data transmission of video sequence stream over the network. The device provides the video sequence stream of the learner and broadcasts it over the Internet through the live stream channel between the device layer and the network layer. The provided video sequence stream feeds into the video sequence pre-processing component, which is encompassed by the affective computing tool within the game-based communication skills training system. We analyse the video sequence stream and divide the stream frame by frame. It allows us to train the datasets, to enable finding a time-sequence of the emotions within a period of time, and to recognize an emotion within a particular frame. The learner video sequence stream feeds into this component and is split to image sequence frames to be sent to the facial/vocal temporal analyser and facial emotion and vocal intonation analyser components. 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 results combine in the synchronizing component. It synchronises the facial emotions and the vocal intonations data and transmits them to the rules engine component or the web service component to be transmitted as feedback/content into the network layer.

All of the components are installed in a web server and are situated in web server layer within the technical framework. The process of facial emotion recognition and vocal intonation recognition are not complete unless the video/voice datasets, trained images/voices, and emotion metadata of both face and voice components are created in data layer. After the learner's video stream feeds into the web server layer, the affective computing tool checks whether the split frame that includes the learner emotion has been already stored in the datasets or otherwise stores the emotions when it is necessary. The trained images/voices components will be triggered during the training process. For each facial/vocal emotion a related metadata is provided in the emotion metadata component. Consequently, when the learner facial and vocal emotions are not stored in the datasets, the require steps will be given. The feedback/content of the data layer and the web server layer transmits to the network layer, proceeds to the device layer, and goes through the interface to the learner. At this stage, the learner can receive his facial and vocal emotions feedback.

3. Method and proof of concept

3.1 Participants

Sixteen participants, all employees from the Centre for Learning Sciences and Technologies (CELSTEC) of Open University of the Netherlands (12 male, 4 female; age M=42.5, SD=10.9), volunteered to participate in this experiment. By signing an agreement form, the participants agreed to capture their facial expressions and voice intonations, and they authorized us to use their data for all further research purposes. The participants did not know in advance what tasks they had to perform; no specific background knowledge was requested. 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 with virtual characters in the affective computing software program. The experiment was individually conducted.

3.2 Design

The hypothesis is to show that it is possible to gather data from which we can infer learners' emotional states (with enough confidence) and to show that such emotional states can be exploited in providing feedback. This preliminary experiment investigates the hypothesis and acts as a proof of concept. At the moment, we don't vary on learner support (just a straight forward simple feedback (red/green signal)) to inform the learner whether the software detects the same 'emotion' as the participant was asked to 'mimic'. For the software, it is important to know whether the detection is OK. For the learner it is important to know that the feedback is correct (so 'green' if the emotion is correct or 'red' if the emotion is incorrect).

3.3 Tasks

Five consecutive tasks were given to the participants: 1) train the database of the affective computing software by exposing seven basic face expressions (happy, sad, surprise, fear, disgust, angry, and neutral), 2) mimic the emotion that was presented through PowerPoint slides. There were 35 images presented after each other; each image illustrated a single emotion (each of the seven basic face expressions is mimicked five times with the following order: happy, sad, surprise, fear, disgust, angry, neutral, happy, sad, …), 3) mimic the seven face expressions two times: first, through slides that each presented the keyword of the requested emotion and second, through slides that each presented the keyword and the picture of the requested emotion with the following order: angry, disgust, fear, happy, neutral, sad, surprise, angry,…, 4) slides presented a text transcript (both sender and receiver) taken from a good-news conversation, participants were requested to read and speak aloud the sender 'slides' of transcript. The text transcript also included instructions what facial expression should accompany the current text-slide. The participants were asked to read the sender part of message and show the accompanying facial expression, 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 (Lang & van der Molen, 2008) and a communication book (Van der Molen & Gramsbergen-Hoogland, 2005).

3.4 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. Moreover, another 1080HD external camera was also utilized to send the participants' emotions to a windows machine so as to record their facial expressions for future usage. The affective computing software used the webcam to capture and recognize the participants' emotions, while Silverback usability testing software (screen recording software) version 2.0 used the external camera to capture and record the complete experimental session. Dell webcam central software was also used in windows machine to record the complete facial expression session of the participants. 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 4.

3.5 Procedure

The participants invited to participate in this experiment by an email. They were asked if they intend and like to improve their acting skills. They performed each individual session in about 20 minutes. They sat in a completely silent room with good lighting condition. Each participant was asked to read and sign the form before the session is started. The moderator of the session presented in the room, but with no intervention. All sessions were performed in about 325 minutes.

Each participant was asked to do tasks 1 up till five in a row and during one individual session preceded by a short instruction at the beginning of each task. All tasks were video-taped and being captured by the affective computing software. 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.

3.6 Result and validation

In this paper we report the validation of the software for the third task in which the participants mimicked the emotions that were presented through PowerPoint slides with respect to participants' behaviour. 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 unable to mimic the requested emotions.

4. 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. A proof of concept study of FILTWAM has been conducted and 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 them. They positively mentioned that they could easily fix the wrong emotions when looked at the reflections of their emotions in the software screen. 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. Based upon the tests' results and the participants' feedback, and as the innovative power that FILTWAM can offer for improving communication skills of lifelong learners, a new version of the software for face emotion recognition will be further 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.

5. Acknowledgments

We thank Jason Saragih for permission to develop the affective computing software based on his face tracker software (Saragih, Lucey, & Cohn, 2010).

References

Aldrich, C. (2009) "The complete guide to serious games and simulations", Somerset, NJ: Wiley.

Armory, A. (2007) "Game Object Model Version II: a Theoretical Framework for Educational Game Development", Educational Technology Research and Development 55, pp 51–77.

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

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

Brantley C. P., and Miller M. G. (2008) "Effective Communication for Colleges", Thomson Higher Education.

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

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

Connolly T. M., Boyle E. A., MacArthur E., Hainey T., and Boyle J. M. (2012) "A systematic literature review of empirical evidence on computer games and serious games", Computers and Education, Volume 59, Issue 2, September, pp 661-686.

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

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

Hager P. J., Hager P., and Halliday J. (2006) "Recovering Informal Learning: Wisdom, Judgment And Community", Springer.

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

Kiili, K. (2005) "Digital Game-Based Learning: Towards an Experiential Gaming Model", The Internet and Higher Education 8(1), pp 13-24.

Lang, G., and van der Molen, H.T. (2008) "Psychologische gespreksvoering", Open University of the Netherlands, Heerlen, The Netherlands

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

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

Quinn, C.N. (2005) "Engaging learning. Designing E-Learning Simulation Games", San Francisco, CA: Pfeiffer, John Wiley and Sons, Inc.

Reeves, B., and Read, J.L. (2009) “Total engagement: Using games and virtual worlds to change the way people work and business compete”, Boston, Harvard Business Press.

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

Sebe N. (2009) “Multimodal Interfaces: Challenges and Perspectives”, Journal of Ambient Intelligence and Smart Environments, January, Vol. 1, No. 1, pp 23-30.

Sebe, N., Cohen, I., Gevers, T., and Huang, T.S. (2006) “Emotion recognition based on joint visual and audio cues”, International Conference on Pattern Recognition, pp 1136-1139, Hong Kong.

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

Subramanian R., Staiano J., Kalimeri K., Sebe N., and Pianesi F. (2010) “Putting the Pieces Together: Multimodal Analysis of Social Attention in Meetings”, ACM Multimedia, Firenze, Italy.

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

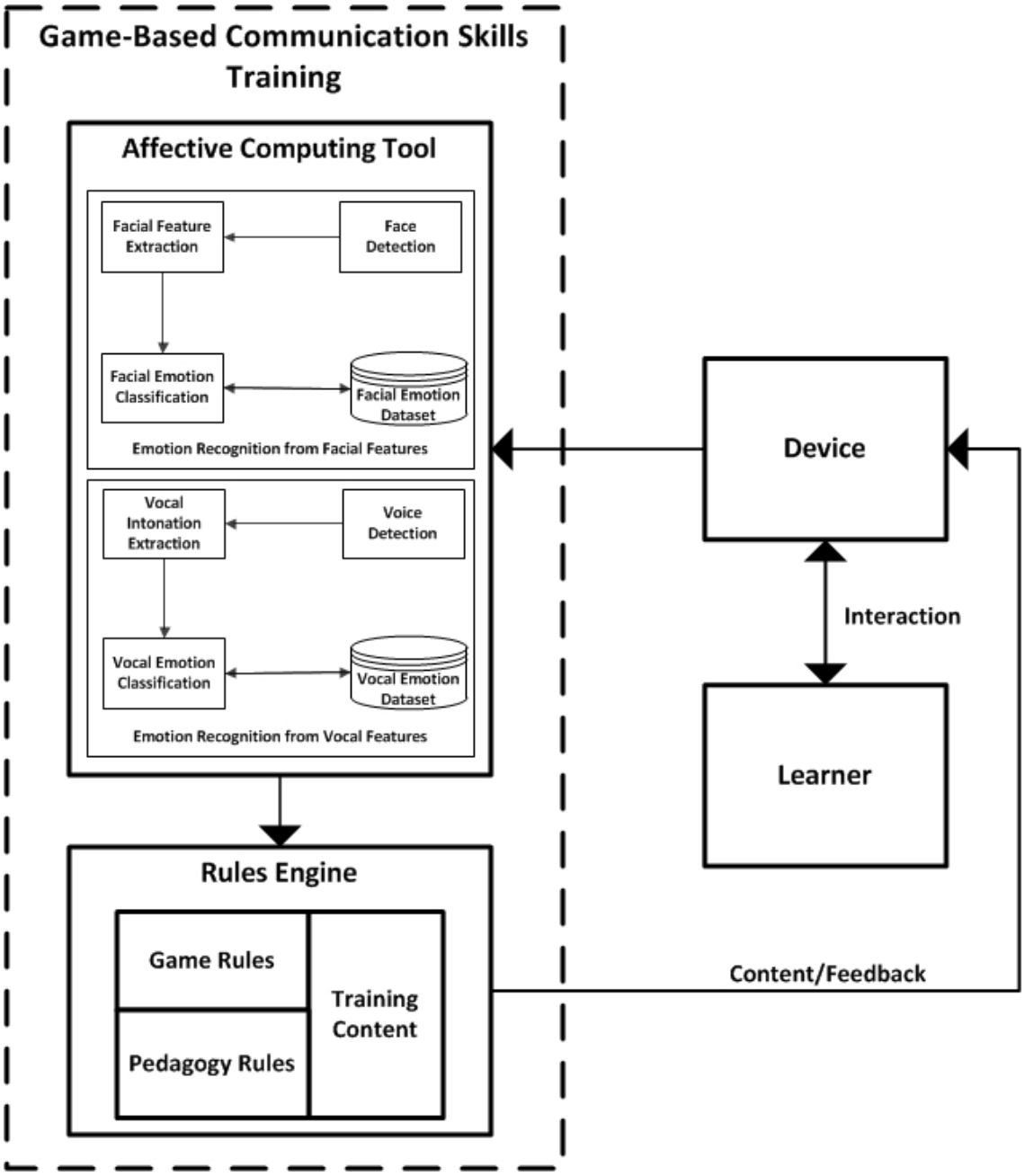

Figure 1: Conceptual framework for online game-based communication skills training

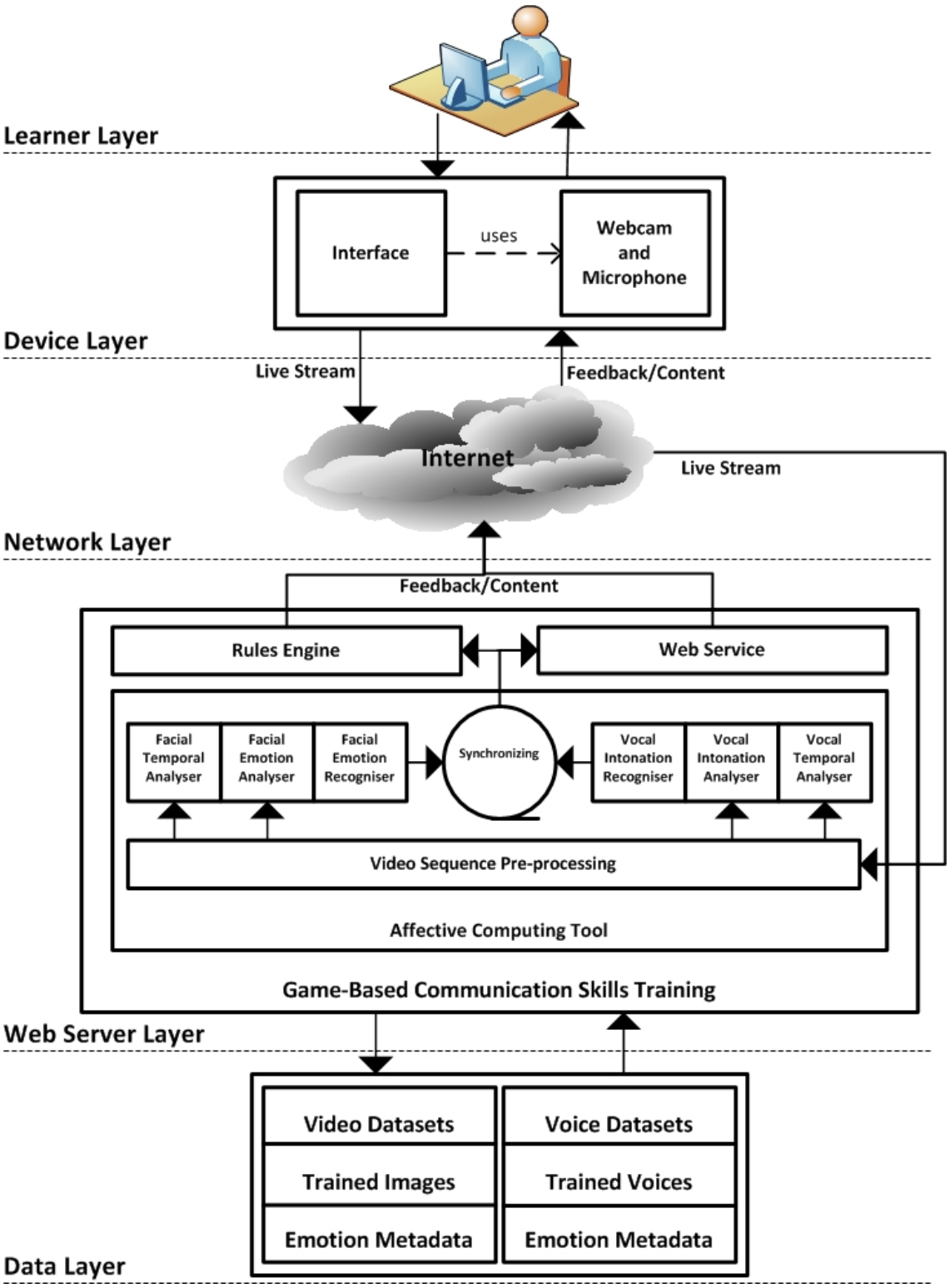

Figure 2: Technical framework of the proposed system

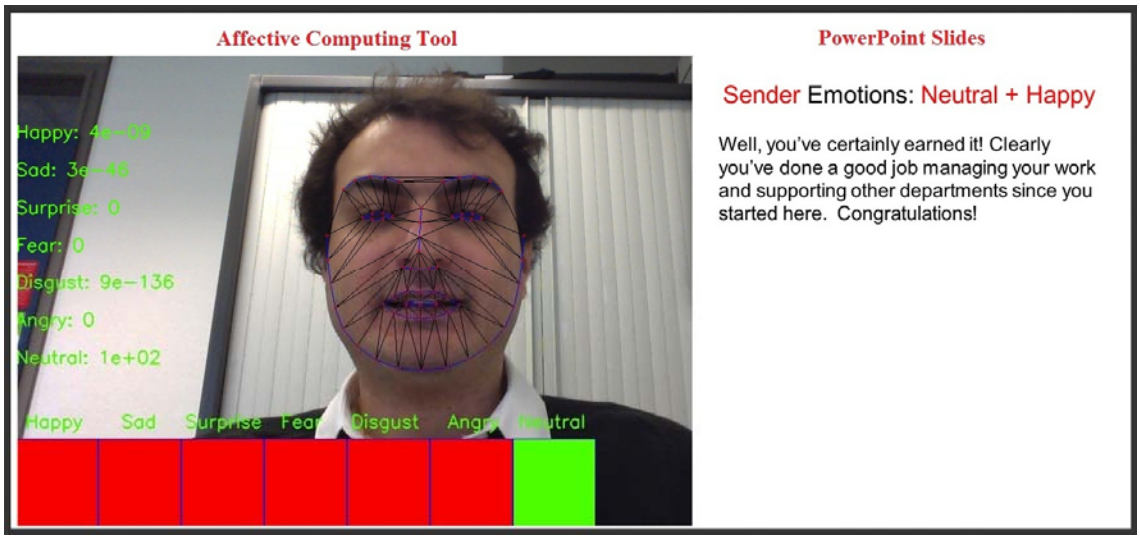

Figure 3: A participant in task 4 and the affective computing software during the experimental session

Table 1: Validation results

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