Toward Multimodal Emotion Recognition in E-Learning Environments

This paper presents a framework (FILTWAM) for real time emotion recognition in e-learning by using webcams. FILTWAM (Framework for Improving Learning Through Webcams And Microphones) offers timely and relevant feedback based upon learner's facial expressions and verbalizations. FILTWAM's facial expression software module has been developed and tested in a proof of concept study. The main goal of this study was to validate the use of webcam data for a real-time and adequate interpretation of facial expressions into extracted emotional states. The software was calibrated with ten test persons. They received the same computer-based tasks in which each of them were requested a hundred times to mimic specific facial expressions. All sessions were recorded on video. For the validation of the face emotion recognition software, two experts annotated and rated participants' recorded behaviours. Expert findings were contrasted with the software results and showed an overall value of Kappa of 0.77. An overall accuracy of our software based on the requested emotions and the recognized emotions is 72%. Whereas existing software only allows not-real time, discontinuous and obtrusive facial detection, our software allows to continuously and unobtrusively monitor learners' behaviours and converts these behaviours directly into emotional states. This paves the way for enhancing the quality and efficacy of e-learning by including the learner's emotional states.

Keywords: E-learning; human-computer interaction; multimodal emotion recognition, real-time face emotion recognition, webcam

Introduction

During the last decade, several new technologies have been adopted by e-learning specialists for enhancing the effectiveness, efficiency and attractiveness of e-learning (Anaraki, 2004). Nowadays, learners are often used to the web-based delivery of e-learning content and Web 2.0 affordances when communicating, working and learning together with their peers in distributed (a)synchronous settings (Ebner, 2007). More

personalized and ubiquitous learning environments have become common (Cheng, Sun, Kansen, Huang, & He, 2005). However, recent developments of input devices (such as webcams) for interacting with such environments are still underexploited. Such devices firstly offer opportunities for more natural interactions with the e-learning applications. Secondly, they offer better ways for gathering affective user data, as they do not interfere with the learning like questionnaires often do. This is because of their unobtrusive and continuously nature of data gathering. Existing methods for gathering affective user data, like physiological sensors and questionnaires, are either obtrusive or discontinuous. They can hamper learning as well as issues in its suitability for e-learning (Feidakis, Daradoumis, & Caballe, 2011; Sarrafzadeh, Alexander, Dadgostar, Fan, & Bigdeli, 2008). Previous software primarily dealt with offline emotion recognition that cause post-processing of the learner's data. They have a couple of limitations that mainly restrict their application context and might impede their accuracy. The application context is restricted by the fact that such software can only manage a small set of expressions from frontal view faces without facial hair, glasses provided that there is constant illumination. Furthermore, the software requires post-processing steps for analysing videos and images and cannot analyse extracted facial information on different time scales (Pantic, Sebe, Cohn, & Huang, 2005). In addition, their accuracy might also be impeded as this software used no databases for authentic emotions. In our research we will investigate the opportunities of a webcam for continuously online and unobtrusive gathering of affective user data in an e-learning context. In this, we also aim to increase the accuracy of face emotion recognition software by implementation of our facial expression module of FILTWAM. In addition, we will stretch the application context of our software.

It is commonly acknowledged that emotions are an important factor in any learning process, since it influences information processing, memory and performance (Pekrun, 1992). Also, feedback based on emotional states may enhance the learners' awareness of their own behaviour. This may be of relevance in communication skills training and the training of other soft skills. Hence, automated emotion detection as explained in this paper may compensate for the limited number of trainers that are available for online training of communication skills in compare to face-to-face situations (Hager, Hager, & Halliday, 2006). Also other areas of e-learning can benefit from affective user data since emotional states are relevant for more domains and objectives (Bachiller, Hernandez, & Sastre, 2010).

Emotion and e-learning

An important factor in the success of human teaching is the capability of a teacher to recognise and respond to the affective states of students. A human teacher may adjust his/her teaching strategy by observing the emotions, facial expressions, and body movement of the students. In e-learning, just as with conventional classroom learning, it is not only about learning (cognition) but also about the (inter)dependency between cognition and emotion which is mediated by the social learning context (teacher, students, learning material). Using emotions in software systems for e-learning would considerably increase performance if the software could adapt to the emotional state of the learner (Sarrafzadeh, Alexander, Dadgostar, Fan, & Bigdeli, 2008). In e-learning, the limited availability of a teacher has driven an increasing number of studies on affective computing. Affective computing could be remedy of gathering affective user data by assigning computers the human-like capabilities of interpretation and generation of affect features (Jianhua, Tieniu, & RosalindW, 2005).

One study (Feidakis, Daradoumis, & Caballe, 2011) showed how to measure emotions specifically for intelligent tutoring systems (ITS). They categorized emotion measurement tools into three areas: psychological, physiological, and motor-behavioural. Psychological tools are self-reporting tools for capturing the subjective experience of emotions of users. Physiological tools comprise sensors that capture an individual's physiological responses. Motor-behaviour tools use special software to measure behavioural movements captured by PC cameras, mouse or keyboard. These tools require experience and objectivity from the user. Many practical applications would considerably increase performance if they could adjust to the emotional state of the user. In this way, when equipped with affective computing module, an ITS can be turned into an affective tutoring system (ATS). And so, a computer application is able to recognize users' facial emotions and can improve its feedback to learners without involvement a human teacher. There is a growing body of research on ATS which stresses the importance of our approach using facial expressions for deriving emotions (Sarrafzadeh, Alexander, Dadgostar, Fan, & Bigdeli, 2008; Ben Ammar, Neji, Alimi, & Gouardères, 2010). Sebe (2009) reports that the most informative channel for computer awareness of emotions, is through facial expressions.

In this study, which is an extension of our previous studies (Bahreini, Nadolski, Qi, & Westera, 2012; Bahreini, Nadolski, & Westera, 2012), we describe the practical application and the first evaluation results for the face emotion recognition part of FILTWAM framework. FILTWAM uses webcams and microphones to interpret the emotional state of people during their interactions with an e-learning environment. It can trigger timely feedback based upon learner's facial expressions and verbalizations. It is capable of discerning the following emotions: sadness, anger, disgust, fear, happiness, surprise, and neutral.

FILTWAM basically offers software with a human-machine interface for the real time interpretation of emotion that can be applied in e-learning. Our software is an extension of FaceTracker software (Saragih, Lucey, & Cohn, 2010) and it is capable of determining any kind of faces even when some parts of the face are covered. We developed facial emotion recognition, facial emotion classification parts of the software, and created a dataset of facial emotions. Our tool, which is able to recognize, interpret, and simulate human emotions, is built upon existing research (Chibelushi & Bourel, 2003; Ekman & Friesen, 1978). It interprets the emotional state of a user in e-learning environment and provides appropriate feedbacks accordingly. Linking two modalities into a single system for affective computing analysis is not new and has been studied before (Chen, 2000; Zeng, Pantic, Roisman, & Huang, 2009). A recent review study 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 from 72% to 85%.

Although our framework allows for both facial and vocal mood detections, we will restrict ourselves to facial mood detection and provide empirical data for this. In this paper we propose 1) an unobtrusive approach with 2) an objective method that can be verified by researchers, 3) which requires inexpensive and ubiquitous equipment (webcam), and 4) which offers interactive software with user-friendly interface. In this paper, section 2 introduces the FILTWAM framework and its face emotion recognition part. The method for the study of the developed software is described in section 3. Results are presented and discussed in section 4. Discussion, findings, and suggestions for future work are described in section 5. Section 6 explains the conclusion of this research.

The FILTWAM framework

The FILTWAM framework encompasses five functional layers and a number of sub-components within the layers. The five layers are introduced as the: 1) Learner, 2) Device, 3) Network, 4) Application, and 5) Data. Figure 1 illustrates the framework.

Figure 1..

Learner layer

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

Device layer

The device reflects the learner's workstation, whether part of a personal computer, a laptop, or a smart device, and it includes a webcam and microphone for collecting user data.

Network layer

The network uses Internet to broadcast a live stream of the learner and to receive the real-time data of the learner.

Application layer

The application layer is the most important part of FILTWAM. It consists of e-learning environment and several sub-components. The e-learning environment uses a webcam and the face emotion recognition technology to facilitate the learning process for the learner. It contains three sub-components named: the affective computing tool, the rule engine, and the web server.

Affective computing tool

It is the heart of FILTWAM. It processes the facial behaviour and voice intonations data of the learner. It consists of a component for emotion recognition from facial features and voice intonation. In this paper we confine ourselves to the facial emotion detection based on the webcam stream

Emotion recognition from facial features

This component extracts facial features from faces 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 (Ekman & Friesen, 1978). Our facial emotion classification component supports 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 (Saragih, Lucey, & Cohn, 2010). It compares the classified emotions with existing emotions in the facial emotion dataset and trains the dataset using a number of learners' faces.

Rule engine

The 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 e-learning component complies with a specific rule-based didactical approach for the training of the learners.

Web service

The web service component transmits the feedback and training content to the learner. At this stage, the learner can receive a feedback based on his/her facial emotion.

Data layer

The data layer is the physical storage of the emotions. It encompasses the facial emotion dataset, which reflects the intelligent capital of the system. Its records provide a statistical reference for emotion detection.

Method and the proof of concept

Our hypothesis is that data gathered via webcam and microphone can be reliably used to unobtrusively infer learners' emotional states. Such emotional states' measurements would allow for the provision of useful feedback for learning during online training of communication skills or any other adaptive or personalized interventions that would enhance the quality and efficacy of e-learning. This study investigates the hypothesis

and acts as a proof of concept for such communication training.

Participants

An email was sent out to employees from the Centre for Learning Sciences and Technologies (CELSTEC) at the Open University of the Netherlands to recruit the participants for this study. The e-mail mentioned the estimated time investment for enrolling in the study. Ten participants, all employees from CELSTEC (8 male, 2 female; age M=42, SD=10.6), volunteered to participate in study. By signing an agreement form, the participants allowed us to capture their facial expressions and voice intonations, and to use their data anonymously for future research. We assured the participants that their raw data will not be available to the public, will not be used for commercial or similar purposes, and will not be available to third parties. The participants were invited to test the software; no specific background knowledge was requested. They were told that participation within the study might help them to become more aware of their emotions while they were communicating through a webcam and a microphone with our software.

Design

Five consecutive tasks were given to the participants. Participants were asked to expose seven basic face expressions. Totally, hundred face expressions were requested for all five tasks together. The participants were requested to mimic all the hundred emotions once. At the moment, we offer very limited learner support (just a straight forward simple feedback (red/green signal)) to inform the learner whether our current prototype of the software detects the same 'emotion' as the participant was asked to 'mimic'. For the validation of the software, it is important to know whether its detection is correct.

For the learners it is important that they can trust that the feedback is correct (so 'green' if the intended emotion is correctly shown or 'red' if otherwise).

The learning goal of the current study is to let the participants become more aware of their emotions. The first task was meant to train the database of the affective computing software. In the second task participants were asked to mimic the emotion that was presented on the image shown to them. There were 35 images presented subsequently through PowerPoint slides; the participant paced the slides. Each image illustrated a single emotion. All seven basic face expressions were five times present with the following order: happy, sad, surprise, fear, disgust, angry, neutral, happy, et cetera. In the third task, participants were requested to mimic the seven face expressions twice: 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, The fourth task presented 14 slides with the text transcript (both sender and receiver) taken from a good-news conversation.

The text transcript also included instructions what facial expression 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 show the accompanying facial expression. The fifth task with 30 slides was similar to 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 Open University of The Netherlands (OUNL) training course (Lang & van der Molen, 2008) and a communication book (Van der Molen & Gramsbergen-Hoogland, 2005). With task 1, there is no learning for the participants, while at other tasks they could easily understand their emotions simultaneously while looking at the feedbacks.

Test environment/Measurement instrument emotions

Participants performed individually on a single Mac computer. 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 interactions with mouse and keyboard on the computer screen. Moreover, another 1080HD external camera was used for recording the sessions for future usage on a separate computer. 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 facial expressions of the participants and record the complete session. Raters for validating our affective computing software used the recorded video. Figure 2 shows a screen shot of a session for one of the tasks.

Figure 2..

Gathering participants' opinions

A self-developed online Google questionnaire collected participants' opinion, whether the learning goal was achieved, and to report their self-assurance. All opinions were collected using items on a 7- point Likert scale format (1=completely disagree, 7=completely agree). Participants' opinions about their tasks were gathered for: 1) difficulty to mimic the requested emotions, 2) quality of the given feedback 3) clarity of the instructions 4) its attractiveness, and 5) their concentration. Participants' self-assurance was measured by their two 7-point Likert scale items 1) being able to mimic the requested emotions and 2) being able to act.

Procedure

Each participant signed the agreement form before his or her session of the study was started. They individually performed all five tasks in a single session of about 20 minutes. The session was conducted in a completely silent room with good lighting condition. The moderator of the session was present in the room, but did not intervene. All 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 our software. After the session, each participant filled out an online Google questionnaire to gather participants' opinions about their learning and the setup of the study.

Validation

Two raters who analysed the recorded video streams carried out validation of the software output. Two raters, both associate professors at the psychology department of Open University of the Netherlands, were invited to individually rate the emotions of the participants' in the recorded video streams. Both raters are familiar and skilled with using the Facial Action Coding System. Raters overall task was to rate the captured video file streams for facial emotion recognition of the participants.

Firstly, they received an instruction package for doing individual ratings of participants' emotions in one video stream. Secondly, both raters participated in a training session together with the main researcher where ratings of this first participant were discussed to identify possibly issues with the rating task and to improve common understanding of the rating categories. Thirdly, raters resumed their individual ratings of participants' emotions in the nine remaining video streams. Fourthly, they participated

in a negotiation session together with the main researcher where all ratings were discussed to check whether negotiation about dissimilar ratings could lead to similar ratings or to sustained disagreement. Finally, the final ratings resulting from the negotiation session were taken as input for the data analysis.

The data of the training session were also included in the final analysis. The raters received: 1) a laptop, 2) a user manual, 3) an instruction guide on how to use ELAN, which is a professional tool for making of complex annotations on video and audio resources, and 4) an excel file with ten data sheets; each of which represented the participants information, such as name and surname.

Results

In this section we report the outcomes of the study. We will first present the agreement between requested emotions and the emotions as recognized by the software. Next we will present the results of the expert raters. Finally we will contrast the software outputs and the raters' judgments.

Software

Table 1 shows the requested emotions of participants contrasted with software recognition results. These numbers are taken from all 1000 emotions (10 test persons displaying 100 emotions each) including the cases that one or more of the rates judged that the test person was unable to mimic the requested emotion correctly. Each requested emotion is separated in two rows that intersect with the recognized emotions by the software. Our software has the highest recognition rate for the neutral expression (77.2%) and the lowest recognition rate for the fear expression (50%) (See Table 1).

Please note that the obtained differences between software and requested emotions are not necessarily software faults but could also indicate that participants

were sometimes unable to mimic the requested emotions. The software had in particular problems to distinguish surprise from neutral. Error rates are typically between 1% and 14%. The software confused 11.3% of the neutral emotions as surprise and confused 12.5% of surprise as neutral.

Table 1. Requested emotions and recognized emotions by the software – These numbers are taken from all 1000 emotions including 'unable to mimic' by the participants.

		Recognized Emotion by the Software							
		Happy	Sad	Surprise	Fear	Disgust	Angry	Neutral	Total
Requested Emotions	Happy	88	2	6	6	10	1	7	120
		73.4%	1.7%	5%	5%	8.3%	0.8%	5.8%	100%
	Sad	0	46	5	8	10	9	12	90
		0%	51.1%	5.6%	8.9%	11.1%	10%	13.3%	100%
	Surprise	0	0	60	6	4	0	10	80
		0%	0%	75%	7.5%	5%	0%	12.5%	100%
	Fear	0	7	6	40	11	7	9	80
		0%	8.8%	7.5%	50%	13.8%	8.7%	11.2%	100%
	Disgust	3	5	1	6	63	8	4	90
		3.3%	5.6%	1.1%	6.7%	70%	8.9%	4.4%	100%
	Angry	0	2	2	3	12	59	2	80
		0%	2.5%	2.5%	3.7%	15%	73.8%	2.5%	100%
	Neutral	4	15	52	19	10	5	355	460
		0.9%	3.2%	11.3%	4.1%	2.2%	1.1%	77.2%	100%
Total		95	77	132	88	120	89	399	1000

The rows from Table 1 show that all seven basic emotions have different distributions for being confused as of the other emotions. In other words, they have different discrimination rates. Apart from neutral, the emotion that shows best discrimination from other emotions is surprise, as surprise has a high score of 75% and is not confused with happy, sad, and angry. The most difficult emotion is fear, which scores only 50% and is easily confused with disgust 13.8%, angry 8.7%, sad 8.8% and neutral 11.2%, respectively. This is in accordance with Murthy (2009) and Zhang

(1999), who found that the most difficult emotion to mimic accurately is fear and this emotion is processed differently from other basic facial emotions. Moreover, Murthy (2009) also states that the three emotions sad, disgust, and angry are difficult to distinguish from each other and are therefore often wrongly classified.

According to the raters' analysis results, Table 2 specifies that the participants were able to mimic the requested emotion in 69.4% of the occurrences. In 200 occurrences (20%) there was disagreement between raters. In 10.6% of the cases the raters agreed that participants were unable to mimic requested emotions (106 times). Participants are best at mimicking neutral (87.4%) and worst at mimicking fear (21.3%). According to Murthy (2009), people indeed have most difficulties at mimicking fear.

Table 2: Raters' agreements and disagreements about 1000 mimicked emotions.

	Happy	Sad	Surprise	Fear	Disgust	Angry	Neutral	Total
Raters agree:	102	24	50	17	47	52	402	694
Able to mimic	85%	26.7%	62.5%	21.3%	52.2%	65%	87.4%	69.4%
Raters disagree:	16	31	22	24	34	22	51	200
Able/unable to	13.3%	34.4%	27.5%	30%	37.8%	27.5%	11.1%	20%
Raters agree:	2	35	8	39	9	6	7	106
Unable to mimic	1.7%	38.9%	10%	48.7%	10%	7.5%	1.5%	10.6%
								100%

Table 3 shows the requested emotions of participants contrasted with software recognition results. But the difference with Table 1 is that we removed both the 'unable to mimic' records and the records on which the raters disagreed from the dataset. We therefore, re-calculated the results of each emotion separately and in total.

Table 3: Requested emotions and recognized emotions by the software – These numbers are taken by the raters from 694 emotions of the participants that were able to mimic the requested emotions.

	Recognized Emotion by the Software							Total

		Happy	Sad	Surprise	Fear	Disgust	Angry	Neutral	
Requested Emotions	Happy	78	2	5	4	9	1	3	102
		76.5%	2%	4.9%	3.9%	8.8%	1%	2.9%	100%
	Sad	0	13	2	4	2	2	1	24
		0%	54.2%	8.3%	16.7%	8.3%	8.3%	4.2%	100%
	Surprise	0	0	41	2	2	0	5	50
		0%	0%	82%	4%	4%	0%	10%	100%
	Fear	0	0	2	11	3	0	1	17
		0%	0%	11.7%	64.7%	17.7%	0%	5.9%	100%
	Disgust	1	1	0	3	35	7	0	47
		2.1%	2.1%	0%	6.4%	74.5%	14.9%	0%	100%
	Angry	0	1	2	2	8	38	1	52
		0%	2%	3.8%	3.8%	15.4%	73.1%	1.9%	100%
	Neutral	3	9	43	16	8	4	319	402
		0.7%	2.2%	10.7%	4%	2%	1%	79.4%	100%
Total		82	26	95	42	67	52	330	694

In 306 out of 1000 cases at least one of the raters has indicated that the participants were ‘unable to mimic’ the requested emotions properly. We only summed occurrences when both raters agreed to observe that displayed emotion was the same as the requested emotion’ is delivered. The result show positive changes when the ‘unable to mimic’ emotions were removed. All emotions except angry emotion move toward positive changes. For example, happy is changed from 73.4% to 76.5%, surprise from 75% to 82%, and neutral from 77.2% to 79.4% (compare Table 1 and Table 3). The achieved overall accuracy of the software between the requested emotions and the recognized emotions assuming uniform distribution of emotions is the average of the diagonal: 72% (based on Table 3).

Results of the raters for recognizing emotions

Hereafter, we describe how the raters detected participants' emotions from their recorded video streams. The disagreement between the raters, which was 34% before

the negotiation session, was reduced to 22% at the end of the negotiation session. In order to determine consistency among raters we performed the cross tabulation between the raters and also interrater reliability analysis using the Kappa statistic approach. We calculated and presented the Kappa value for the original ratings before negotiation. We have 1000 displayed emotions (see Table 1) whose recognition is rated by two raters as being one of the seven basic emotions. The cross tabulation data are given in Table 4. Each recognized emotion by the rater 1 is separated in two rows that intersect with the recognized emotions by the rater 2. The first row indicates the number of occurrences of the recognized emotion and the second row displays the percentage of each recognized emotion.

Table 4: Rater1 * Rater2 Cross tabulation – All 1000 emotions are rated by both raters.

		Rater2							
		Happy	Sad	Surprise	Fear	Disgust	Angry	Neutral	Total
Rater1	Happy	106	0	1	1	1	0	8	117
		90.6%	0%	0.9%	0.9%	0.9%	0%	6.7%	100%
	Sad	0	32	0	1	3	8	16	60
		0%	53.3%	0%	1.7%	5%	13.3%	26.7%	100%
	Surprise	9	0	57	8	2	1	30	107
		8.4%	0%	53.3%	7.5%	1.9%	0.9%	28%	100%
	Fear	0	0	16	23	14	0	5	58
		0%	0%	27.6%	39.7%	24.1%	0%	8.6%	100%
	Disgust	0	3	2	2	58	8	12	85
		0%	3.5%	2.4%	2.4%	68.2%	9.4%	14.1%	100%
	Angry	1	6	1	1	6	69	10	94
		1.1%	6.4%	1.1%	1.1%	6.4%	73.4%	10.5%	100%
	Neutral	6	4	5	0	1	7	456	479
		1.3%	0.8%	1%	0%	0.2%	1.5%	95.2%	100%
Total		122	45	82	36	85	93	537	1000

Cross tabulation analysis between the raters indicates that the neutral expression has the highest agreement (95.2%) and the fear expression has the lowest agreement

between them (39.7%) (Table 4). According to Murthy (2009), people have more difficulty in recognizing fear facial expression and this could be the reason that the most confused expression is fear among the raters to recognize. Sad is the next confused category, which is recognized as neutral 26.7%. Analyzing of the Kappa statistic underlines the agreement among the raters. The result with 95% confidence among the raters reveals that the interrater reliability of the raters was calculated to be Kappa = 0.715 ($p < 0.001$). Therefore a substantial agreement among raters is obtained based on Landis and Koch interpretation of Kappa values (Landis & Koch, 1977).

Results of contrasting the software outputs and the raters' ratings

Using the raters agreement about the displayed emotions as a reference we report the reliability analysis of our software-based emotion recognition using 95% confidence intervals and $p < 0.001$ in Table 5. It shows the Kappa value of each emotion and the overall Kappa value amongst raters, and the software derived from 694 emotions. This number (694) is used as both raters agreed that the participants were able to mimic the requested emotions (see Table 2).

An analysis of the Kappa values for each emotion reveals that most agreement is for the emotion-category happy (Kappa = 0.887, $p < .001$) followed by neutral 0.818 followed by angry 0.806, disgust 0.704, sad 0.664, surprise 0.644, and finally fear 0.495.

Table 5: The overall Kappa of 694 occurrences and the Kappa value of each emotion among raters and software.

Name of emotion	Happy	Sad	Surprise	Fear	Disgust	Angry	Neutral
Kappa value	0.887	0.664	0.644	0.495	0.704	0.806	0.818
Overall Kappa				0.77			

The result with 95% confidence among the raters and the software reveals that the interrater reliability of them was calculated to be Kappa = 0.77 ($p < 0.001$). Therefore a substantial agreement among the raters and the software is obtained based on Landis and Koch interpretation of Kappa values (Landis & Koch, 1977). We should state that this Kappa value (0.77) is calculated based upon the raters' opinions and the software's results; however the overall accuracy of our software (0.72) is calculated based upon the requested emotions and the recognized emotions.

Participants opinions results

The Google-questionnaire indicated that 8 of 10 participants felt that the feedback supported them to learn and mimic the emotions. The feedback also helped them to become more aware of their emotions. The result of the online questionnaire indicates that all tasks seem moderately difficult. The feedback and the clear instructions were totally helpful. All the tasks were interesting for the participants to do. The concentration factor indicated no distraction during performance. The self-assurance factor was less for tasks 1 and 2 as compared to the other tasks. It was easy to realize that the participants did not regard themselves as actors.

Ethical implications

In this study and in the implementation of our software, learning analytics and users' privacy including making the current participants' data or future users' data available to public without their prior permission are serious issues that we are aware of the consequences. Therefore we used a protected data model for our learning analytics that is described in (Greller & Drachsler, 2012).

Discussion

This study contrasted the requested emotions of participants with software recognition results for the face emotion recognition part of FILTWAM. We used two human raters for a reference. This study showed a substantial agreement between the raters and the software with overall Kappa value 0.77, while including only the cases of full agreement between human raters (694 emotions are considered). The kappa value of 0.77 indicates that the software quite accurately establishes the users' emotions.

The best recognized emotion is surprise 82% followed by neutral 79.4%, happy 76.5%, disgust 74.5%, angry 73.1%, fear 64.7%, and sad 54.2%. Here also the result shows that the most intensive emotions are ranked higher than the less intensive emotions except the neutral emotion. In the 306 cases where one or both raters indicated that our participants were unable to mimic emotions, the participants had problems mimicking sad 66 followed by fear 63, neutral 58, disgust 43, surprise 30, angry 28, and happy 18 times. This is in agreement with Murthy (2009) and Zhang (1999), who found that the most difficult emotion to mimic accurately is fear and this emotion is processed differently from other basic facial emotions. Moreover, our data analysis confirms Murthy (2009) finding, in which was stated that the three emotions sad, disgust, and angry are difficult to distinguish from each other and are therefore often wrongly classified. The overall accuracy of our software based on the requested emotions and the recognized emotions is 72%.

Anger and disgust share many similar facial actions (Ekman & Friesen, 1978) and that is probably the reason why they are two common confused emotions in our Table 1 and Table 3. In the 90 cases of disgust in Table 1 and 47 cases of disgust in Table 3 where the requested emotions and the recognized emotions by the software are displayed, 8 and 7 cases are recognized as angry, respectively. In the 80 cases of angry

in Table 1 and 52 cases of angry in Table 3 where the requested emotions and the recognized emotions by the software are displayed, 12 and 8 cases are recognized as disgust, respectively.

Non-actors were selected for our study. A previous study by Krahmer and Swerts has shown that using actors, although they evidently have better acting skills than layman, will not lead to more realistic (i.e., authentic, spontaneous) expressions (Krahmer & Swerts, 2011). However, as youngsters and older adults are not equally good in mimicking different basic emotions (e.g., older adults are less good in mimicking sadness and happiness than youngsters, but older adults mimic disgust better than youngsters), it is acknowledged that the sample of test persons might influence the findings of the software accuracy (Huhnel, Fölster, Werheid, & Hess, 2014). In our study we used medium-aged adults. It could be that this sample of medium-aged adults can cope for the strength and weaknesses of both older adults and youngsters but this has not been investigated. No gender differences in mimicry for both younger male and female participant have been reported by (Huhnel, Fölster, Werheid, & Hess, 2014). Nevertheless, because there might be gender differences in older age, upcoming research would comprise older adults.

There have been several improvements in the accuracy of the developed emotion recognition software. Bettadapura (2012) reports accuracies for existing expression recognition software solutions ranging from 55% till 98% since 2001. Our software is capable of the unobtrusive and real time detection and categorization of emotions. In 306 cases (30.6%) our participants were unable to mimic the requested emotions, but all appreciated our software for being very easy and straightforward to use. We managed to fulfil our basic requirements of 1) an unobtrusive approach with, 2) inexpensive and

ubiquitous equipment (webcam), and 3) that offers interactive software with user-friendly interface.

It is expected that the rate of correct software emotion recognition can be further improved when the face emotion recognition part of FILTWAM is combined with the voice emotion recognition part which would offer an even more interesting avenue for applying emotion recognition in e-learning (Sebe, 2009). Indeed, the FILTWAM framework is prepared for including multimodal data.

Conclusion

This paper introduced a new framework called FILTWAM to continuously and unobtrusively monitor learners' behaviour during e-learning and to interpret this into emotional states. FILTWAM aims to improve learning using webcams and microphones as input devices and exploits multimodal emotion recognition of learners during e-learning while linking emotion detection to adapted learning activities. We continue Sebe's (2009) approach to combine both visual and audio information for classification to improve the accuracy of detecting one or more basic emotions. FILTWAM anticipates the increased importance of affective user states and cognitive states in pedagogical scaffolding. Our new approach supports the usage of unobtrusive consumer equipment, which is portable and easy to use. Although we have considered only seven basic emotions in this study, our software can be easily extent for more emotions. The outcomes of FILTWAM could influence different groups' best interests in a virtual setting. For example, a doctor/patient model that is investigated in (Alepis &Virvou, 2011) for affective e-learning in medical education and an instructor/learner model that is investigated in (Ben Ammar, Neji, Alimi, & Gouardères, 2010), are two cases that may take advantage of this framework. When learners use our approach they will be supported in improving their communication skills. It will be done by becoming more

aware of their non-verbal behaviour during their conversations (e.g. during their delivery of good news or bad news). The feedback of our software will provide this personalised support. In this, the future development of the voice emotion recognition module, the integration of the face emotion recognition and the voice emotion recognition modules, and handling these two modules in an online e-learning environment are consecutive steps in achieving FILTWAM's full potential for e-learning.

Background

Acknowledgments

We thank our colleagues who participate in the face emotion recognition proof of concept study. We also thank the raters for their help in rating and analysing the data sets. We thank Jason Saragih for permission to develop the affective computing software based on his FaceTracker software (Saragih, Lucey, & Cohn, 2010).

Support

Notes on contributors

References

Alepis, E., & Virvou, M. (2011). Automatic generation of emotions in tutoring agents for affective e-learning in medical education. *Expert Systems with Applications*, 38 (8), 9840-9847.

Anaraki, F. (2004). Developing an Effective and Efficient eLearning Platform. *International Journal of The Computer, the Internet and Management*, 12 (2), 57-63.

Bachiller, C., Hernandez, C., & Sastre, J. (2010). Collaborative learning, research and science promotion in a multidisciplinary scenario: information and communications technology and music. *Proceedings of the International Conference on Engineering Education* (pp. 1-8). Gliwice, Poland.

Bahreini, K., Nadolski, R., Qi, W., & Westera, W. (2012, October 4–5). FILTWAM - A framework for online game-based communication skills training - Using webcams and microphones for enhan- cing learner support. In P. Felicia (Ed.),

The 6th European conference on games based learning (ECGBL) (pp. 39–48). Cork: Academic Publishing International Limited Reading.

Bahreini, K., Nadolski, R., & Westera, W. (2012, October 29–31). FILTWAM - A framework for online affective computing in serious games. In A. De Gloria & S. de Freitas (Eds.), The 4th International Conference on Games and Virtual Worlds for Serious Applications (VS- GAMES'12). Procedia Computer Science (Vol. 15, pp. 45–52), Genoa, Italy. Amsterdam: Curran Associates.

Ben Ammar, M., Neji, M., Alimi, A. M., & Gouardères, G. (2010). The Affective Tutoring System. *Expert Systems with Applications*, 37 (4), 3013-3023.

Bettadapura, V. (2012). Face Expression Recognition and Analysis: The State of the Art. *Journal of CoRR*, abs/1203.6722.

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*.

Cheng, Z., Sun, S., Kansen, M., Huang, T., & He, A. (2005). A personalized ubiquitous education support environment by comparing learning instructional requirement with learner's behavior. *19th International Conference on Advanced Information Networking and Applications (AINA)*, 2, 567-573.

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

Ebner, M. (2007). E-Learning 2.0 = e-Learning 1.0 + Web 2.0?. *The Second International Conference on Availability, Reliability and Security (ARES)*, 1235-1239.

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

Feidakis, M., Daradoumis, T., & Caballe, S. (2011). Emotion Measurement in Intelligent Tutoring Systems: What, When and How to Measure. *Third International Conference on Intelligent Networking and Collaborative Systems*, 807-812.

Greller, W., & Drachsler, H. (2012). Translating Learning into Numbers: A Generic Framework for Learning Analytics. *Educational Technology & Society*, 15 (3), 42–57.

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

Huhnel, I., Fölster, M., Werheid, K., & Hess, U. (2014). Empathic reactions of younger and older adults: No age related decline in affective responding. *Journal of Experimental Social Psychology*, 50, 136-143.

Krahmer, E., & Swerts, M. (2011). Audiovisual Expression of Emotions in Communication. *Philips Research Book* Series. Springer Netherlands, 12, 85-106.

Landis, J. R., & Koch, G. G. (1977), The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.

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

Murthy, G. R. S., Jadon, R. S. (2009). Effectiveness of Eigenspaces for facial expression recognition. *International Journal of Computer Theory and Engineering*, 1 (5), 638-642.

Pantic, M., Sebe, N., Cohn, J. F., & Huang, T. (2005). Affective Multimodal Human-computer Interaction. *Proceedings of the 13th Annual ACM International Conference on Multimedia*, 5, 669-676.

Pekrun, R. (1992). The impact of emotions on learning and achievement: towards a theory of cognitive/motivational mediators. *Journal of Applied Psychology*, 41, 359–376.

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

Sarrafzadeh, A., Alexander, S., Dadgostar, F., Fan, C., & Bigdeli, A. (2008). How do you know that I don't understand?" A look at the future of intelligent tutoring systems. *Computers in Human Behavior*, 24(4), 1342-1363.

Sebe, N. (2009). Multimodal Interfaces: Challenges and Perspectives. *Journal of Ambient Intelligence and Smart Environments*, 1(1), 23-30.

Jianhua, T., Tieniu, T., & RosalindW, P. (2005). Affective Computing: A Review. *Affective Computing and Intelligent Interaction, Springer Berlin Heidelberg*, 3784, 981-995.

Van der Molen, H.T., & Gramsbergen-Hoogland, Y.H. (2005). Communication in Organizations: Basic Skills and Conversation Models. *Psychology Press*, New York.

Zhang, Z. (1999). Feature-Based Facial Expression Recognition: Sensitivity Analysis and Experiment with a Multi-Layer Perceptron. *International Journal of Pattern Recognition Artificial Intelligence*, 13 (6), 893-911.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31 (1), 39–58.

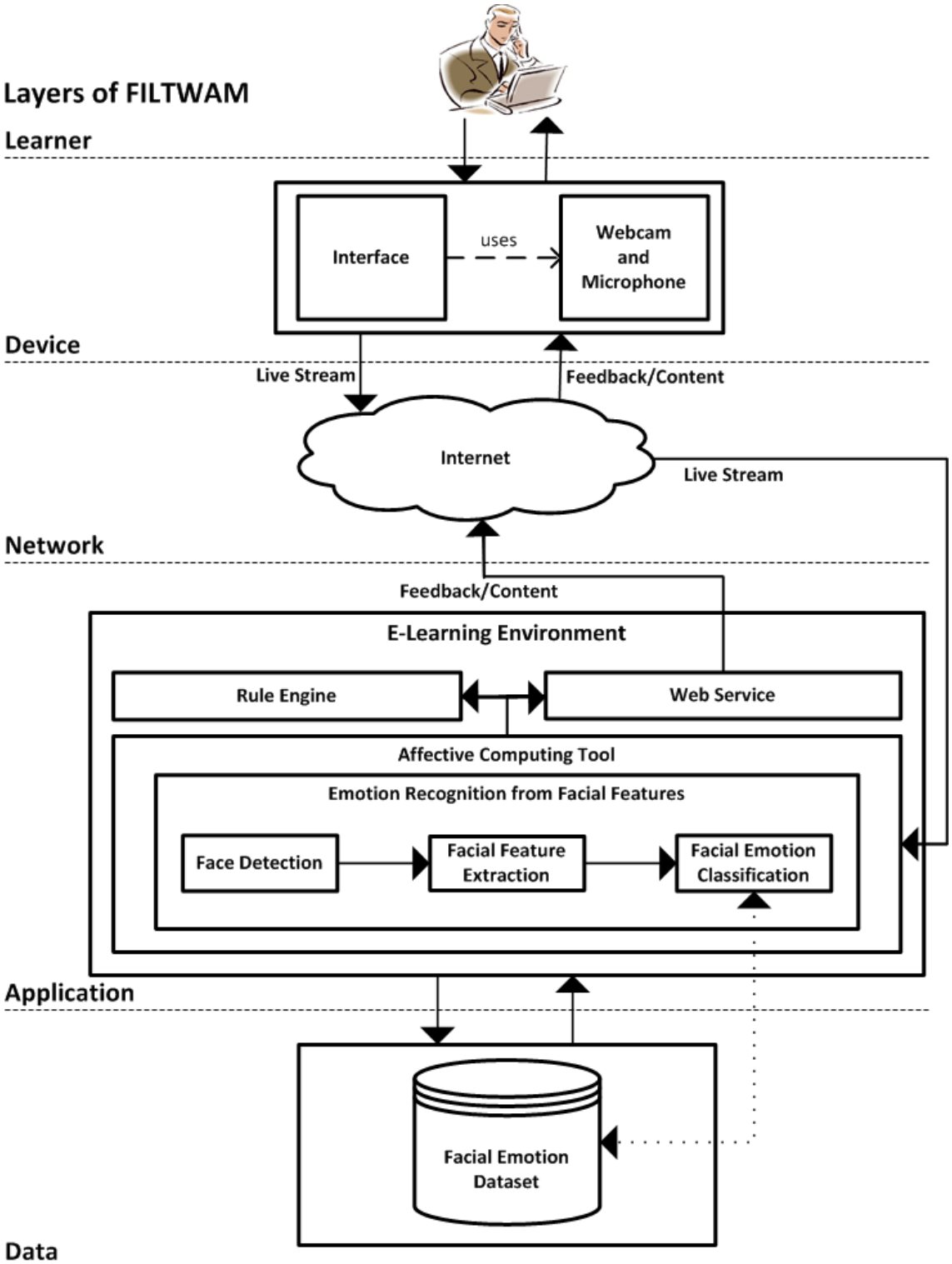
Layers of FILTWAM
Learner
Interface
uses
Webcam and Microphone
Device
Live Stream
Feedback/Content
Internet
Live Stream
Network
Feedback/Content
E-Learning Environment
Rule Engine
Web Service
Affective Computing Tool
Emotion Recognition from Facial Features
Face Detection
Facial Feature Extraction
Facial Emotion Classification
Application
Facial Emotion Dataset
Data

Fig. 1.

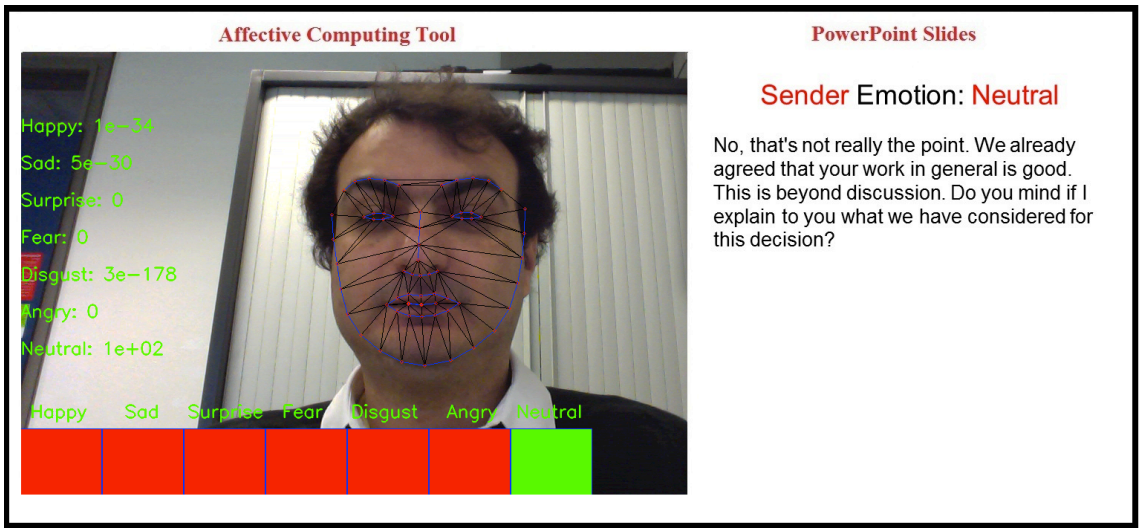
Affective Computing Tool
PowerPoint Slides
Happy: 1e-34
Sad: 5e-30
Surprise: 0
Fear: 0
Disgust: 3e-178
Angry: 0
Neutral: 1e+02
Happy
Sad
Surprise
Fear
Disgust
Angry
Neutral
Sender Emotion: Neutral
No, that's not really the point. We already agreed that your work in general is good. This is beyond discussion. Do you mind if I explain to you what we have considered for this decision?

Fig. 2.