International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: https://www.tandfonline.com/loi/hihc20

Data Fusion for Real-time Multimodal Emotion Recognition through Webcams and Microphones in E-Learning

Kiavash Bahreini, Rob Nadolski & Wim Westera

To cite this article: Kiavash Bahreini, Rob Nadolski & Wim Westera (2016) Data Fusion for Real-time Multimodal Emotion Recognition through Webcams and Microphones in E-Learning, International Journal of Human–Computer Interaction, 32:5, 415-430, DOI: 10.1080/10447318.2016.1159799

To link to this article: https://doi.org/10.1080/10447318.2016.1159799

Accepted author version posted online: 02 Mar 2016.
Published online: 28 Mar 2016.

Intl. Journal of Human–Computer Interaction, 32: 415–430, 2016

ISSN: 1044-7318 print / 1532-7590 online
DOI: 10.1080/10447318.2016.1159799

Taylor & Francis
Taylor & Francis Group

Data Fusion for Real-time Multimodal Emotion Recognition through Webcams and Microphones in E-Learning

Kiavash Bahreini, Rob Nadolski, and Wim Westera
Welten Institute, Research Centre for Learning, Teaching and Technology, Faculty of Psychology and Educational Sciences, Open University of the Netherlands, Heerlen, The Netherlands

This article describes the validation study of our software that uses combined webcam and microphone data for real-time, continuous, unobtrusive emotion recognition as part of our FILTWAM framework. FILTWAM aims at deploying a real-time multimodal emotion recognition method for providing more adequate feedback to the learners through an online communication skills training. Herein, timely feedback is needed that reflects on the intended emotions they show and which is also useful to increase learners' awareness of their own behavior. At least, a reliable and valid software interpretation of performed face and voice emotions is needed to warrant such adequate feedback. This validation study therefore calibrates our software. The study uses a multimodal fusion method. Twelve test persons performed computer-based tasks in which they were asked to mimic specific facial and vocal emotions. All test persons' behavior was recorded on video and two raters independently scored the showed emotions, which were contrasted with the software recognition outcomes. A hybrid method for multimodal fusion of our multimodal software shows accuracy between 96.1% and 98.6% for the best-chosen WEKA classifiers over predicted emotions. The software fulfils its requirements of real-time data interpretation and reliable results.

1. INTRODUCTION

Emotions play a significant role in our daily lives. Emotions are manifest in each action of our behaviors (Preeti, 2013). It is generally accepted that emotions are a significant influential factor in the processes of learning, as they affect memory and action (Pekrun, 1992). Being able to demonstrate and understand emotions is important in both face-to-face settings and in computer-mediated communications. Now that people increasingly use modern devices (such as laptops, tablets, and mobile phones) and the Internet to facilitate human–machine and human–human interactions (viz. online communication) (Preeti, 2013), software-based emotion detection is an emerging topic in human–computer interaction research. Emotion recognition will gain relevance in diverse domains, e.g., health, learning, and entertainment, since it allows for adapting the responses of software applications to the end-users' emotional states. Emotion detection could also be applied for computer-based training of soft skills, e.g., communication skills, interview skills, negotiation skills, as it would allow for giving direct feedback to the learners about their emotional appearances. Because of the dynamic and volatile nature of emotions, such applications would often demand a real-time interpretation (Schuller, Lang, & Rigoll, 2002). Unfortunately, most of the current software applications for emotion recognition require offline post-practice analyses of recorded data, which fail to produce real-time results. Scarce real-time methods tend to be based only on a single modality (voice, facial expression, skin resistance, posture, etc.), which restricts their accuracy of emotion recognition considerably (Preeti, 2013; Schuller, Lang, & Rigoll, 2002; Vogt, 2011). In principle, using multimodal data sources would increase the accuracy of emotion detection. However, so far real-time results from multimodal emotion-recognition methods have been highly unreliable and rarely usable for practical application (Grubb, 2013). Schuller and colleagues (Schuller, Lang, & Rigoll, 2002) described that real-time emotion recognition analysis inevitably must be accepted to be lower than offline emotion recognition analysis, and tasks should be limited to very few emotional states.

In this study we present a validation study of our multimodal emotion recognition software system that we have developed and composed of existing software modules for real-time unimodal emotion analysis. For this we have developed and implemented a software architecture framework that is called FILTWAM (Framework for Improving Learning Through Webcams And Microphones). In this framework we have combined two emotion recognition software modules (using face and voice emotion recognition) that we have described in previous research studies (Bahreini, Nadolski, & Westera, 2014; Bahreini, Nadolski, & Westera, 2015). The aim of the research was to offer a real-time and multimodal solution that would increase the accuracy of the combined face and voice emotion

Address correspondence to Kiavash Bahreini, Welten Institute, Research Centre for Learning, Teaching and Technology, Faculty of Psychology and Educational Sciences. Open University of the Netherlands, Valkenburgerweg 177, 6419 AT Heerlen, The Netherlands. E-mail: kiavash.bahreini@ou.nl

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hihc.

recognition software modules. The context of our research is a computer-based training of communication skills. In the sessions, affective learner data were gathered continuously and unobtrusively. Participants in the training sessions were asked to mimic specific facial and vocal emotions while receiving real-time onscreen feedback on their mimicked emotions, based on the software recognition outcomes. We have investigated the following research questions: (1) what is the reliability of multimodal emotion recognition compared to unimodal emotion recognition; and (2) to what extent do the learners appreciate the emotion feedback that was based on our real-time multimodal approach? For the validation and calibration of the emotion detection software, all participants' behaviors were recorded on video and afterwards scored independently by two expert raters.

In this article, we first provide a brief overview of previous research in multimodal emotion detection. Thereafter we present the FILTWAM framework and its multimodal fusion method. In addition, we describe the methodology for the validation study. We will present the results of both the software's accuracy and the users' appreciations of the sessions and finally we will discuss the findings of this study and present the conclusions.

2. RELATED WORKS

Previous research on emotion detection has mainly focused on the so-called unimodal methods as separate sources of data (Buisine et al., 2014; 2009; Nwe, Foo & De Silva, 2003; Zhang, 1999; Vogt, 2011). However, various recent studies (Busso et al., 2004; Chen, 2000; Sebe, Cohen, Gevers, & Huang, 2006; Zeng, Pantic, Roisman, & Huang, 2009) deal with multimodal emotion recognition by combining multiple input data sources (Wagner et al., 2013). Research into multimodal emotion recognition has gained practical relevance in human–computer interaction, social media as well as in learning studies. For example, the impacts of combining kinesthetic learning and facial expression were reported by Gaffary, Eyharabide, Martin, & Ammi (2014) and a multimodal intelligent eye-gaze tracking system was reported by Biswas & Langdon (2015). Furthermore, multimodal emotion recognition and its related technologies could improve learning performance in e-learning context when it combines with affective states of learners and when it provides emotional states of learners through appropriate feedback mechanisms. One example of the multimodal emotion detection and classroom learning was reported in a study by Bosch and his colleagues (Bosch, Chen, D'Mello, Baker, & Shute, 2015). They compared and combined facial expressions and interaction features derived from students' interactions in a serious game. They reported that the unimodal face detections were more accurate than the unimodal interaction detections in the game. Furthermore, they reported that the multimodal approach improved the accuracy of the system to 98%.

An important success factor in the classroom learning is the capability of an instructor to timely recognize and respond to the affective states of their learners. For this, teachers continuously adjust their teaching behaviors by observing and evaluating the behaviors of their learners, including their facial expressions, body movements, and other signals for overt emotions. In e-learning, just as with classroom learning, the dependency and interdependency between cognition and emotion and their relationships, are quite important. The relationships between learners' cognition and emotion are influenced by the electronic learning environment, which mediates the communication between participants (instructor, learners) and contains or refers to learning resources (e.g., photos, audios and videos, and animations). Moreover, the context of learning can also be that a student is only interacting with the e-learning materials, while fellow students and instructors might be irregularly involved too. Software systems for e-learning (e.g., VLE's, PLE's, serious games) could better foster learning if they could adapt the instruction and feedback to their recognized emotional state of the learner (Sarrafzadeh, Alexander, Dadgostar, Fan, & Bigdeli, 2008). The relationship between emotion recognition and e-learning has been studied before (see, for example, D'Mello & Graesser, 2012 for affective learning; Rus, D'Mello, Hu, & Graesser, 2013 for intelligent tutoring systems). An intelligent tutoring system equipped with an affective computing module is an example of aforementioned systems. The affective computing module might be able to recognize learners' facial and vocal emotions. The affective tutoring system can use this module without instructor's involvement for adapting its feedback to the learner taking his emotional state into account. There is a growing body of research on affective tutoring systems, which stresses the importance of our approach using facial and vocal expressions for deriving emotions (Ben Ammar, Neji, Alimi, & Gouarderes, 2010; Sarrafzadeh et al., 2008). In this, our multimodal emotion detection software can be used within intelligent tutoring systems.

Jaimes and Sebe (Jaimes & Sebe, 2007) also showed that the accuracy of detecting one or more basic emotions is greatly improved when both visual and audio information are used in offline data classification. They showed that the multimodal data fusion could raise to accuracy levels from 72% up to 85% if the following conditions are met: (1) clean audio-visual input, such as noise-free data set, closed and fixed microphone, non-occluded portraits; (2) from actors; (3) who speak single words; and (4) who display exaggerated facial expressions of the six basic emotions (happy, sad, surprise, fear, disgust, and anger) (Ekman & Friesen, 1978).

In our study, we aim to improve this accuracy level in an online real-time setting rather than an offline setting. In addition, we want to relax the boundary conditions. From the above-mentioned conditions, we will only follow the first condition with the six basic emotions complemented with the neutral emotion, while neglecting the other three conditions: (1) we will not use actors; (2) test persons will speak sentences rather

than separate words; and (3) test persons will not be required to display exaggerated facial expressions. We follow the study of Busso and colleagues (Busso et al., 2004), which describes two different approaches for combining unimodal systems for data fusion: feature-level fusion, and decision-level fusion.

The feature-level fusion approach includes mixing all the features of each data source into a single file or into a single vector. Different data types of different data sources are combined in this approach. The prepared file will be very huge with mixing of all variables from all the data sources. In contrast with the feature-level fusion, the decision-level fusion approach emphasizes the extraction of the features from each data source separately. It then applies a data classifier algorithm to the features independently. Then, the results are fused over a classified data set. This approach by Busso combines the features coming from different data sources that are needed in our study, but it covers only four basic emotion categories (happiness, sadness, neutral, and anger) in a combined audio and video acted data set (Busso et al., 2004). They compared feature-level and decision-level fusion approaches over this multimodal acted data set.

Busso et al. (2004) reported that the accuracy of detecting four basic emotions is greatly improved from 65% to 89.3% when both visual and audio information are used in offline data classification. It was reported that decision-level fusion provides better results for happiness and sadness emotions, while feature-level fusion delivers superior results for neutral and anger emotions.

Nevertheless, these two fusion approaches are inappropriate for the continuous interpretation of learner's expressions. Chen and colleagues (Chen, Huang, Miyasato, & Nakatsu, 1998) proposed a rule-based approach for multimodal emotion recognition on the six basic emotions. They showed that by combining two modalities of face and voice into a single system, it is possible to achieve higher recognition rates than either modality alone. They proposed a modified algorithm for the rule-based approach and examined the extracted features from both modalities. It is not clear from their study if they followed the real-time approach, though. De Silva and Ng (De Silva & Ng, 2000) proposed a rule-based approach for decision-level multimodal fusion when face expressions are combined with voice expressions. For recognizing six kinds of emotions, they used several statistical techniques and Hidden Markov Models (HMM). They classified anger, fear, sad, dislike, surprise, and happy from facial expressions and speech in a manual way. They only recruited two participants for the recording sessions. Their findings expressed that the face and voice expressions can be combined using a rule-based system to improve the recognition rate. Preeti (2013) proposed a conceptual-level rule-based approach for hybrid multimodal fusion based on both feature-level and decision-level when face and voice expressions are considered simultaneously. She suggested that her approach requires to be implemented in a software application and should be tested in a real situation. Her approach also requires a rule-based engine that has to be included with a lot of rules for multimodal emotion recognition.

All studies mentioned above regard multimodal fusion systems better for emotion recognition than unimodal data sources. However, none of these offered a real-time approach with a reliable acted data set with the neutral emotion and the six basic emotions that have been proposed by Ekman and Friesen (1978). These seven emotions are a de facto standard for studies dealing with emotions in the past thirty years. Likewise, none of these studies on multimodal emotion recognition proposed an accurate software system with capability of real-time interpretation of learner's expressions. Our FILTWAM framework, presented in the next section, offers a hybrid model for real-time, continuous, reliable, and accurate multimodal emotion recognition system that combines two modalities (face and voice) into a single form.

3. THE FILTWAM FRAMEWORK

The FILTWAM framework enables timely feedback to learners during a communication skills training by primarily taking their manifest emotions into account. The learner's emotion data are gathered through a webcam and a microphone when the learner interacts with an e-learning server (Bahreini, Nadolski, Qi, & Westera, 2012a; Bahreini, Nadolski, & Westera, 2012b). The FILTWAM framework includes five layers and a number of components within the layers (see Figure 1). The five layers are introduced as (1) Learner, (2) Device, (3) Data, (4) Network, and (5) Application. In conjunction with FILTWAM in this study, we used EMERGO that is an open source toolkit for the development and delivery of multimedia cases in e-learning environments and it allows users to acquire complex skills. (Nadolski et al., 2008). However, FILTWAM can also be used in other e-learning environments.

3.1. Learner Layer

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

3.2. Device Layer

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

Web Interface

The web interface runs a serious game in the device layer and allows the learner to interact with the game components in the application layer. This component indirectly uses the EMERGO web service client. The web interface will receive the

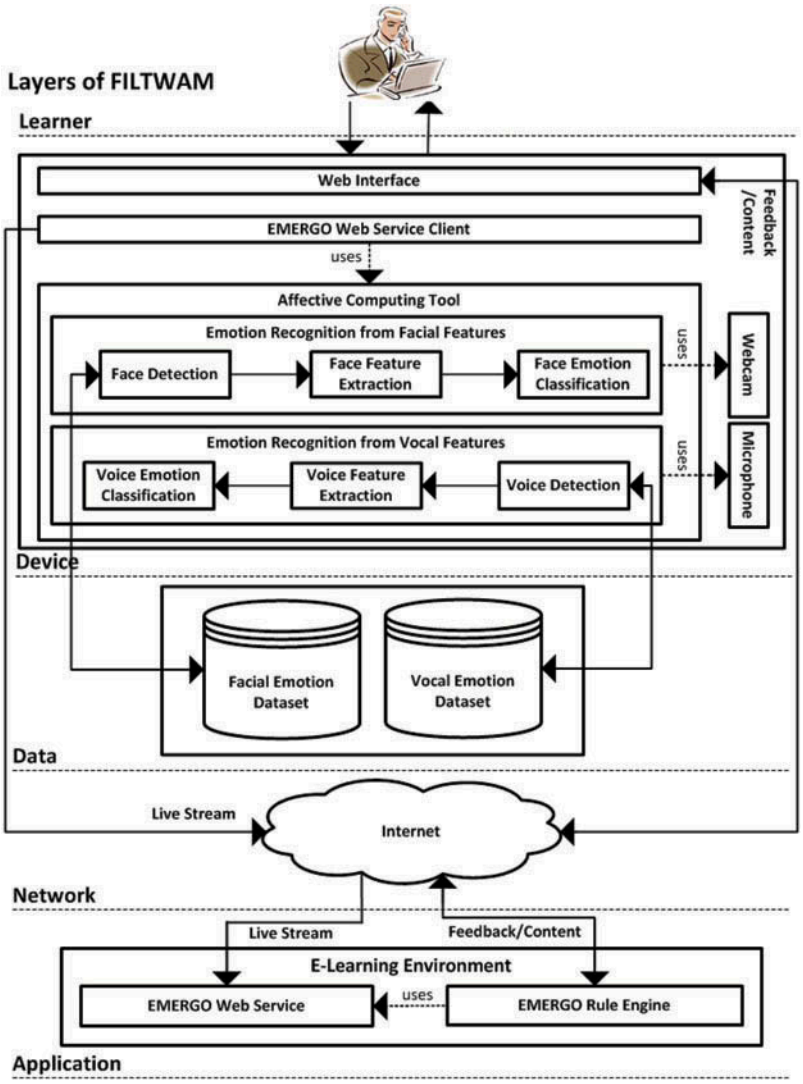

FIG. 1. The FILTWAM framework integrates the face emotion recognition software application and the voice emotion recognition software application in an e-learning environment. The face emotion recognition and the voice emotion recognition components have been reported in our previous studies (Bahreini, Nadolski, & Westera, 2014; Bahreini, Nadolski, & Westera, 2015).

feedback/content through Internet and the game-based learning environment in application layer.

EMERGO Web Service Client

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

Affective Computing Tool

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

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

Face detection. The process of emotion recognition from facial features starts at the face detection component. But we do not necessarily want to recognize the particular face; instead we intend to detect a face and to recognize its facial emotions. The person's face is detected using the Viola–Jones object detection framework in real time (Viola & Jones, 2001; 2002). This framework and its algorithm (cvHaarDetectObjects())

were implemented in Open Source Computer Vision Library (OpenCV), which is an open source software library for computer vision and machine learning.

Face 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. We extract the face features using the Constrained Local Model (CLM) framework, which is an open source framework based on the face tracking and landmark detection algorithms (Cristinacce & Cootes, 2008). This framework is written in C++ and is used in OpenCV to extract the facial landmarks in real time. Moreover, we use stable stochastic optimization strategies like the simplex-based technique described in Cristinacce and Cootes (2004) to extract the facial features. We use the same training set for detecting and tracking the faces as Saragih and his colleagues used (Saragih, Lucey, & Cohn, 2011). We use their training data set based on the CLM strategy and two widely available databases. Their training data set includes more than 600 persons, 3000 images, 66 facial landmarks in each image, 61 connections between each two facial landmarks, and 91 triangles between each three facial landmarks to track the face of a subject in real time. Similarly, we use a training data file from the XM2VTS database (Messer, Matas, Kittler, Luuttin, & Maitre, 1999) and the CMU Pose database (Gross, Matthews, Cohn, Kanade, & Baker, 2008) to extract the facial landmarks.

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 the classification of these six basic emotions plus the neutral emotion, but can in principle also recognize other or more detailed face expressions when required. This component analyzes 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 in face detection and face tracking levels, because its database has been trained using different subjects that met those criteria. Additionally, the development of the component is based on the FaceTracker software (Saragih, Lucey, & Cohn, 2011). However, this component does not recognize any cultural differences in emotion recognition level, because its database has not been trained for this purpose. During the analysis, one image that already includes a not-yet determined emotion is compared with all already classified images in the data set. Then this image will be classified as one of the indicated emotions. It compares the classified emotions with existing emotions in the facial emotion data set and trains the data set using a number of learners' faces. Moreover, we use action units, which are the essential movements of individual muscles over faces and compare them with the emotional labels to classify facial emotion. The final version of our face emotion classification uses the same approach described in the CLM framework with the extended version of the Cohn–Kanade (CK+) database (Lucey, Cohn, Kanade, Saragih, Ambadar, & Matthews, 2010). It uses a trained and tested version of the face emotion data set. We develop our face emotion classification using C and C++ languages in OpenCV version 2.3.1 based on the face-detection and the face-tracking source codes of Saragih and his colleagues (Saragih, Lucey, & Cohn, 2011). Our software classifies emotions based on lips, mouth, nose, eyebrows, eyes, lids, chin, jaw, and cheeks. We use the point distribution model (PDM) to extract the geometric difference on a face from the training set of shapes (Cooper, Cootes, Taylor, & Graham, 1995). Our software uses the facial landmarks in the training data file to interpret noisy and low-contrasted images. The software uses principal component analysis (PCA), which was invented by Karl Pearson (Pearson, 1901) to calculate correlations of movement between groups of facial landmarks among the training data file. The PCA also allows us to convert a set of correlated facial landmarks into a set of linearly uncorrelated facial landmarks. The facial landmark training data file becomes very large, because the number of features each image is extremely large. To overcome a problem that is called "curse of dimensionality," this large feature space is projected into a smaller feature space using the PCA. For allowing real-time feature tracking and feature extraction, we use the CLM and the PDM approaches to recognize facial expressions of each subject.

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

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

Voice feature extraction. Once the voice is detected, the voice feature extraction component extracts a sufficient set of features from the voice of the learner. These features are considered the significant features of the learner's voice and can be automatically extracted. For feature selection, we used the openSMILE software to extract specific features from the input speech streams. This software was developed at Technische Universität München in the scope of the EU-project SEMAINE in 2008. The extracted audio features were processed and stored into an "arff" file using a default configuration file (emobase.conf) to be used in the WEKA software. The openSMILE obtains a large set of features from the input speech signals by default. Such low-level descriptors and functional features are considered to be: duration, intensity, intonation, harmonicity, perturbation, pitch, formants, spectrum, mel-frequency cepstrum coefficients (MFCCs), low-frequency power coefficients (LFPCs), perceptual linear predictive coefficients (PLPs), wavelets, voice quality parameters,

non-linguistic vocalizations, first-order moments, percentiles, zero crossing rate (ZCR), temporal, spectral, extremes, mean, moments, regression, segments, peaks, and onsets. We then condensed the WEKA feature set from 990 features to the 93 features that have been introduced above and are presented in our baseline voice emotion data set. For this, we used an attribute evaluator and a search method in the WEKA software. For reducing the number of uncorrelated variables, we used an orthogonal transformation from the PCA method for speech analysis (Wang, Ling, Zhang, & Tong, 2010). We used a search algorithm (Ranker[1]) to rank the attributes of the extracted features.

Voice emotion classification. This component analyzes the voice stream and can extract a millisecond feature of each voice stream for its analysis. We used the sequential minimal optimization (SMO)[2] classifier of WEKA[3] software, which is a software tool for data mining. The WEKA software uses the generated "arff" file to compare the extracted features with the features within the voice emotion data set to classify the vocal emotion.

3.3. Data Layer

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

3.4. Network Layer

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

3.5. Application Layer

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

E-Learning Environment

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

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

4. FILTWAM AND ITS MULTIMODAL FUSION METHODS

We propose a hybrid fusion method in the WEKA tool that combines the feature-level and decision-level fusion approaches. These two approaches were described and used in (Castellano, Kessous, & Caridakis, 2008). In this study we focus on applying the hybrid fusion method over the multimodal data set that is generated by combination of the face and voice unimodal software modules. This method is based on multimodal emotion data classifier algorithms in the WEKA tool. Figure 2 represents our hybrid method and its steps.

The FILTWAM framework basically offers a hybrid fusion model. In this model, we first follow the decision-level fusion approach that has been explained before and extract the features from each data source separately. The feature analysis, its selection steps, its methodological approaches, and the criteria to select features of the facial expressions and the vocal intonations have been already stated in the FILTWAM framework and its subsections separately. In this model, the first data source, (i.e., the face emotion recognition software) will receive the real-time face expressions from a webcam and will detect the face in a pre-processing step. We then extract the feature points on the detected face. The output of this step goes into the feature analysis and selection step. When the desired features have been selected, the face emotion recognition calls a data classifier algorithm in WEKA. We determine which features should be selected and which data classifier algorithm is called in our software. Next, we make a data set for multimodal emotion integration. Then, we follow the feature-level fusion approach and mix all the features of each data source into a single file and make a data set for multimodal emotion integration. This approach is similarly done for the voice emotion recognition until the data are available to be integrated into one single set. After this, the classifiers for hybrid emotion classification in WEKA are applied over the integrated emotions data set. Then this method will classify the emotional states of the learner.

5. EVALUATION METHODOLOGY

In this study we asked participants to carry out four consecutive tasks that constitute the alpha-release of the communication skills training.

[1] http://weka.sourceforge.net/doc.dev/weka/attributeSelection/Ranker.html

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

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

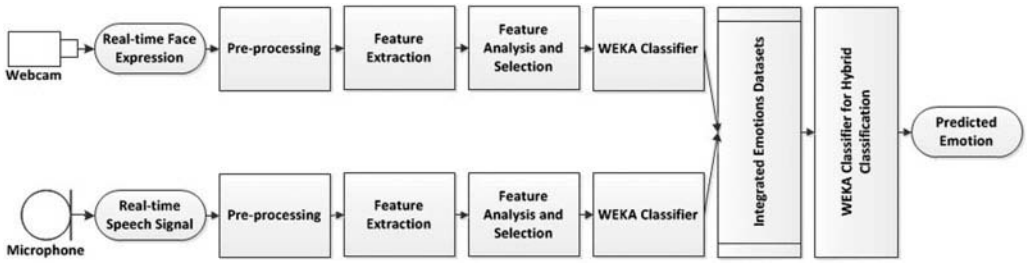

FIG. 2. The hybrid method for data fusion of the combined data sources for face emotion recognition and for voice emotion recognition software modules.

5.1. Participants

Twelve participants, all employees from the Welten Institute (7 male, 5 female; age M = 40, SD = 9) volunteered to participate in the study. Participants were non-actors. The participants were invited to test the multimodal emotion recognition software and take part in the communication skills training. By signing an agreement form, the participants allowed us to record their facial expressions and their vocal intonations. They also allowed us to use their data anonymously for future research. For participating in this experiment, no specific background knowledge was requested.

5.2. Design

Participants were asked to expose the seven basic face and voice expressions (happy, sad, surprise, fear, disgust, anger (Ekman & Friesen, 1978), and neutral) in four consecutive tasks. In this way, in total 80 face expressions and 80 voice expressions of each participant were gathered. During the session, we offered very limited feedback to the participant: the name of the recognized emotion and its prediction accuracy were projected on screen. The participants could watch their own facial expressions at the top-left, the analyzed voice expressions at the top-right, and the PowerPoint sheets with instructions at the bottom of the screen. In this way, the participant was informed whether or not our affective computing software detected the same "emotion" as he or she was asked to mimic. In the first task, the participants were asked to mimic the face expressions while looking at the webcam, speak aloud, and produce the voice emotion that was shown on the image presented to them. There were 14 images subsequently presented through PowerPoint slides; the participant paced the slides. Each image illustrated a single emotion. All seven basic face expressions were presented twice. This task was supposed to help the participants to visualize the real expressions. In the second task, participants were requested to mimic a set of face expressions and to speak aloud the seven basic expressions twice: first, through the slides that each presented the keyword of the requested emotion and second, through the slides that each presented the keyword and a picture example of the emotion. In total, 14 PowerPoint slides were used for the second task. For the first and the second task, participants could improvise and use their own texts. This task was set up to allow the participants to mimic their own expressions and compare the requested emotions with the first tasks. The third task presented 16 slides with the text transcript (both sender and receiver) taken from a good-news conversation. Each slide offered a single text transcript and a requested emotion for both face and voice expressions through a single PowerPoint slide. Here, participants were requested to read and speak aloud the sender text of the "slides" from the transcript and were asked to deliver the accompanying face and voice expressions. This task was set up to provide a real conversation toward a positive result. The fourth task with 36 slides was similar to task 3, but in this case the text transcript was taken from a bad-news conversation. This task was set up to provide a real conversation toward a negative result. The transcripts and instructions for tasks 3 and 4 were taken from an existing OUNL training course (Lang & van der Molen, 2008) and a communication book (Van der Molen & Gramsbergen-Hoogland, 2005). These four tasks were supposed to help the participants to understand and improve their facial and vocal expressions for the seven basic emotions.

5.3. Test Environment

All tasks were performed on a single Mac machine. The Mac screen was divided into three panels: top-left, top-right, and bottom (see Figure 3).

An integrated webcam with a microphone and a 1080HD external camera were used to capture and record the emotions of the participants as well as their actions on the computer screen. The external camera was used for recording the facial and vocal expressions of the participants for future usage (e.g., using by the raters to analyze the participants' expressions) on a separate computer. The affective computing software with the face and voice emotion recognition software modules used webcam and microphone 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 and vocal expressions of the participants and record the complete session.

5.4. Questionnaire and Gathering Participants' Opinions

We have developed an online questionnaire to collect participants' opinion about the multimodal emotion feedback.

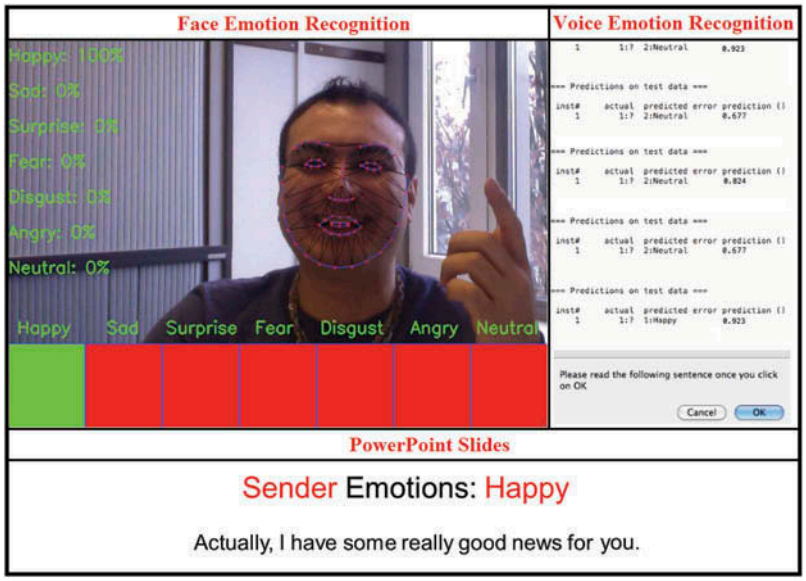

FIG. 3. Screenshot of the main researcher mimicking a task. Task 3 and the affective computing tool including the face emotion recognition software module and the voice emotion recognition software module during the experimental session.

We requested the participants to report their experiences through the questionnaire right after completion of the exercises. All participants' data were collected using items with 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) self-assurance for being able to mimic the requested emotions; (4) clarity of the instructions; (5) the attractiveness of the tasks; (6) their concentration on the given tasks; and (7) their acting skills.

5.5. Procedure

All participants signed the agreement form before his/her session of the study started. They individually performed all four tasks in single sessions of about 30 minutes. The sessions were conducted in a silent room with good lighting conditions. During the session, a moderator was present in the room. The moderator gave a short instruction at the beginning of each task, but did not intervene. The instruction included the request to show mild and not-too-intense expressions while mimicking the emotions after the session. All 12 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.

5.6. Raters

Two expert raters analyzed the recorded video and audio files. First rater is a PhD employee at the Psychology Department of the Open University of the Netherlands and the second rater is a lecturer who also has a psychology background in emotion detection/recognition and works at the Computer and Electrical Engineering Department of IAU University of Tehran. Both raters individually rated the emotions of the participants in the recorded video files. Both raters are familiar and skilled with face, voice, and speech analysis. To determine the accuracy of the emotion recognition system, the raters were asked to categorize and rate the recorded video files of the participants for facial expressions, vocal intonations, and the integration of the two. For supporting the rating process, the raters used the ELAN tool, which is a professional tool for making complex annotations on video and audio resources.

First, the raters received an instruction package for doing ratings of one of the participants' emotions in one video file. Second, both raters participated in a training session where ratings of the participant were discussed to identify possible issues with the rating task and to improve common understanding of the rating categories. Third, raters resumed their individual ratings of participants' emotions in the complete video files. Fourth, they participated in a negotiation session 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 this negotiation session were contrasted with the software results for the further analysis by the main researcher. The data that the raters rated during the training session were also included in the final analysis. The raters received (1) a user manual; (2) 12 video files of all 12 participants; (3) an instruction guide on how to use ELAN; and (4) an Excel file with 12 data sheets; each of which represented the participants' information, such as name and surname.

The raters rated the facial expressions and the vocal intonations of the participants in the form of categorical labels

covering the six basic emotions (happiness, sadness, surprise, fear, disgust, and anger) suggested by Ekman and Friesen (1978), as well as the neutral emotion.

6. EMPIRICAL STUDY OF USER BEHAVIOR AND RESULTS

In this section we report on the results of the study. We first present the comparison of the recognized emotions of the participants by the raters for both modalities. Second, we present the combined comparison of the raters for both modalities. Third, we combine the raters' agreement on both participants' facial and vocal expressions with the multimodal results of the face and voice software modules. Fourth, we report the results of comparing the software outputs and the raters' ratings using WEKA classifiers in our hybrid model. Finally, we will report participants' opinions.

6.1. Results of Raters and Multimodal Software for Recognizing Emotions

Hereafter, we describe how the raters detected participants' emotions from their recorded video files. The disagreement between the raters for the face emotion recognition, which was 21% before the negotiation session, was reduced to 12.5% at the end of the negotiation session. The disagreement between the raters for the voice emotion recognition, which was 27% before the negotiation session, was reduced to 19.2% 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 (κ) statistic approach. We calculated and presented the κ value for the original ratings before negotiation. We have 960 displayed emotions whose recognition is rated and negotiated by two raters as being one of the seven basic emotions. The cross-tabulation data (agreement matrix between the raters) are given in Tables 1 and 2 for the face and voice emotion recognition results, respectively. Each recognized emotion by one rater is separated into two rows that intersect with the recognized emotions by the other rater. The first row indicates the number of occurrences of the recognized emotion and the second row displays the percentage of each recognized emotion.

In Table 1, the cross-tabulation analysis between the raters indicates that the neutral expression has the highest agreement (95.3%). It followed by anger (91.2%), happy (83.2%), disgust (83%), sad (72%), surprise (68%). The fear expression has the lowest agreement between them (56.2%). Our data analysis between the two raters indicates that they experienced some difficulties in distinguishing between "surprise and happy," "fear and surprise," "sad and anger," and "sad and neutral" groups. Indeed, the raters had to correct their recognition rate after the negotiation session mostly in these four groups. The high value of the κ statistic of Table 1 (before negotiation) establishes the agreement among the raters. The result with 95% confidence among the raters reveals that the inter-rater reliability of the raters was calculated to be $\kappa = 0.8$ ($p < 0.001$). Therefore, a substantial agreement among raters is obtained based on Landis and Koch interpretation of κ values (Landis & Koch, 1977).

TABLE 1
Rater1 * Rater2 Cross Tabulation for the Face Emotion Recognition ($\kappa = 0.8$)

	Emotion Recognized by the Software							
	Happy	Sad	Surprise	Fear	Disgust	Anger	Neutral	Total
Requested Emotion								
Happy	84	0	10	1	0	1	5	101
	83.2%	0%	9.9%	1%	0%	1%	4.9%	100%
Sad	1	41	0	0	3	7	5	57
	1.7%	72%	0%	0%	5.2%	12.3%	8.8%	100%
Surprise	11	1	51	4	0	0	8	75
	14.7%	1.4%	68%	5.3%	0%	0%	10.6%	100%
Fear	0	0	9	27	6	2	4	48
	0%	0%	18.8%	56.2%	12.5%	4.2%	8.3%	100%
Disgust	1	0	1	1	54	3	5	65
	1.5%	0%	1.5%	1.5%	83%	4.7%	7.8%	100%
Anger	0	0	1	1	1	44	1	48
	0%	0%	2.2%	2.2%	2.2%	91.2%	2.2%	100%
Neutral	5	8	7	0	3	4	539	566
	0.9%	1.4%	1.2%	0%	0.5%	0.7%	95.3%	100%
Total	102	50	79	34	67	61	567	960

TABLE 2
Rater1 * Rater2 Cross Tabulation for the Voice Emotion Recognition ($\kappa = 0.712$)

	Emotion Recognized by the Software							
	Happy	Sad	Surprise	Fear	Disgust	Anger	Neutral	Total
Requested Emotion								
Happy	112	0	2	1	5	10	12	142
	79%	0%	1.4%	0.7%	3.5%	7%	8.4%	100%
Sad	1	40	0	5	2	0	47	95
	1%	42.1%	0%	5.3%	2.1%	0%	49.5%	100%
Surprise	5	0	45	0	1	1	0	52
	9.6%	0%	86.6%	0%	1.9%	1.9%	0%	100%
Fear	1	3	0	39	1	7	4	55
	1.7%	5.5%	0%	71%	1.8%	12.7%	7.3%	100%
Disgust	1	2	3	2	38	9	0	55
	1.8%	3.6%	5.5%	3.6%	69.1%	16.4%	0%	100%
Anger	1	0	7	1	4	35	0	48
	2%	0%	14.6%	2%	8.4%	73%	0%	100%
Neutral	7	18	0	16	0	5	467	513
	1.3%	3.6%	0%	3.1%	0%	1%	91%	100%
Total	128	63	57	64	51	67	530	960

The result with 95% confidence among the raters in Table 2 reveals that the inter-rater reliability of the raters was calculated to be $\kappa = 0.712$ ($p < 0.001$). Therefore, also for voice emotion recognition a substantial agreement among raters is obtained. From the literature we know that the human recognition accuracy obtained by Nwe, Foo, & De Silva (2003) was 65% and that obtained by Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss (2005) was 80%.

We followed the study of Geertzen (2012) and Hallgren (2012) for inter-rater analysis with multiple raters, and report on the combination of the raters' agreements on facial expressions of the participants and the face emotion recognition software results in Table 3.

The overall value of the κ statistic of 0.644 ($p < 0.001$) reflects a substantial agreement among raters and the face emotion recognition software based on the Landis and Koch interpretation of κ values (Landis & Koch, 1977).

As discussed above, we followed the same approach for the voice expressions and reported the results in Table 4.

The overall value of the κ statistic of 0.533 ($p < 0.001$) reflects a moderate agreement among raters and the voice emotion recognition software based on the Landis and Koch interpretation of κ values (Landis & Koch, 1977).

We take the next step according to the raters' analysis results in order to address the ratings by the two independent raters were used to determine the accuracy of the system in multimodal emotion recognition and to show how the ratings of the raters were similarly used in the same system. We used the combined data set including both the face and voice emotion recognition software results and the raters' analysis results, and removed the occurrences from the data set where the raters mentioned that the participants were unable to mimic the requested emotions (including the exaggerated emotions) and where there was a sustained disagreement between the raters. We only kept the occurrences where four ratings of the face and voice of the two raters were similar (rater 1 rated two times: one time for face and one time for voice, and similarly rater 2 rated two times: one time for face and one time for voice). Using this filtering technique, the multimodal data set indicated that

TABLE 3
The Overall κ Value of 960 Occurrences and the κ Value of Each Emotion Between Two Raters for Facial Expressions of the Participants and the Results of the Face Emotion Recognition Software with 95% Confidence Interval

	Happy	Sad	Surprise	Fear	Disgust	Anger	Neutral
Raters agree:	0.684	0.540	0.471	0.530	0.574	0.666	0.748

Note. Overall $\kappa = 0.644$.

TABLE 4
The Overall κ Value of 960 Occurrences and the κ Value of Each Emotion Between Two Raters for Vocal Intonations of the Participants and the Results of the Voice Emotion Recognition Software with 95% Confidence Interval

	Happy	Sad	Surprise	Fear	Disgust	Anger	Neutral
Raters agree:	0.586	0.321	0.601	0.478	0.431	0.496	0.619

Note. Overall $\kappa = 0.533$.

TABLE 5
The Overall κ Value of 534 Occurrences and the κ Value of Each Emotion Between Raters' Agreements and the Emotion Recognition Software Results

	Happy	Sad	Surprise	Fear	Disgust	Anger	Neutral
Raters agree:	0.912	0.729	0.818	0.776	0.818	0.860	0.886

Note. Overall $\kappa = 0.86$.

the participants were able to mimic the requested emotion in 534 cases (56%) for the facial expressions and for the vocal intonations at the same time. Using the raters' agreement on the multimodal data set 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 κ value of each emotion and the overall κ value amongst raters (rater 1 for face and voice and rater 2 for face and voice) and the face and voice emotion recognition software derived from 534 emotions. This number (534) is used as both raters agreed that the participants were able to mimic the requested emotions. An analysis of the κ values for each emotion reveals that most agreement is for the emotion category of happy ($\kappa = 0.912$, $p < 0.001$) followed by neutral 0.886, anger 0.860, surprise 0.818, disgust 0.818, fear 0.776, and sad 0.729.

Analysis of the κ statistic of Table 5 with 95% confidence yields an inter-rater reliability of the raters, the face, and the voice emotion recognition software modules of $\kappa = 0.86$ ($p < 0.001$). Therefore, an almost perfect agreement is obtained based on Landis and Koch's interpretation of κ values (Landis & Koch, 1977).

6.2. Results of Contrasting the Software Outputs and the Raters' Ratings

In this section we report on the results based on our hybrid method represented in Figure 2. We address how problems such as overfitting have been faced in our data set using the WEKA tool. Overfitting is the problem that appears in producing a classifier that fits the training data too tightly and works well on it, but not on independent test data. In order to solve this issue, we used model selection algorithms to automatically decide which features to keep and which features to leave. We then used a completely separate test set with no instances in common with the training set. We also used cross-validation on our training data to prevent overfitting on our training set too. We used ten-fold cross-validation statistical approach for evaluating and comparing learning algorithms on our integrated data set by dividing the date into two subsets: one subset (10%) is used to train a model and the other (90%) used to validate the model. We then compared the results of the generated confusion matrix of each classifier algorithm in WEKA and reported the best-chosen WEKA classifiers over predicted emotions. In this, all 79 available WEKA classifiers have been applied over the integrated emotional data set and on the predicted emotions for hybrid classification. Among them, top eight classifiers, which showed better prediction results over the emotional states of the learners, have been reported in Table 6. Furthermore, the overall κ based on the raters' analysis result is reported in Table 6. Our approach shows a very accurate and reliable result, leading to accuracy levels from 96.1% to 98.6% for the best-chosen WEKA classifiers over predicted emotions. These classifiers are kind of data mining algorithms that have been implemented in WEKA. They allow for supervised classification[4] in data mining tools like WEKA.

This result indicates that the function classifiers (SMO and Logistic) in WEKA have the highest minimum and maximum accuracies among other classifiers (97% and 98.6%). SMO is a type of support vector classifier and that is the reason that it is fast and accurate enough for data classification. It applies sequential minimal optimization algorithm of Platt (Platt, 1999). Logistic uses a multinomial logistic regression model with a ridge estimator that can be used for building our multimodal data (Le Cessie & van Houwelingen, 1992). AODEsr is a Bayesian classifier that detects uniqueness between two attribute values and removes the general attribute value (Zheng & Geoffrey, 2006). WAODE is also a kind of Bayesian classifier that creates the Weightily Averaged

[4]http://wiki.pentaho.com/display/DATAMINING/Data+Mining+Algorithms+and+Tools+in+Weka

TABLE 6
The Integration Results of the Multimodal Emotion Recognition in the WEKA Data Mining Tool and The Overall κ Value Based on the Raters' Analysis Result

Classifier Type	Classifier Name	Minimum Accuracy	Maximum Accuracy
Bayes	AODEsr	96.4%	98.2%
Bayes	WAODE	96.6%	98.1%
Functions	Logistic	97%	98.6%
Functions	SMO	97%	98.6%
Lazy	LBR	96.1%	96.5%
Lazy	LWL	96.4%	98.1%
Rules	JRip	97.1%	98.5%
Rules	NNge	97.2%	97.8%

Note. Overall $\kappa = 0.86$.

One-Dependence Estimators model (Jiang & Zhang, 2006). Lazy Bayesian Rules Classifier (LBR) is a naive Bayesian classifier (kind of lazy) that provides effective classifier learning. It achieves lower error rates over a range of learning tasks (Zheng & Webb, 2000). Locally Weighted Learning classifier (LWL) is also a lazy classifier that uses an instance-based algorithm to allocate instance weights to the multimodal data (Frank, Hall, & Pfahringer, 2003). JRip is a kind of rules classifier that provides a propositional rule learning method that decreases error rates (Cohen, 1995). Nearest Neighbor Like (NNge) classifier is also a kind of rules classifiers. It uses if–then rules for the data classification (Brent, 1995).

6.3. Results of the Raters for Recognizing Emotions

Table 7 presents the opinion of the participants. The answers to the questionnaire indicated that eight of 12 participants found that it was somewhat easy, easy, or completely easy for them to mimic the requested emotions in the given tasks (see the difficulty of the given tasks). Seven out of 12 mildly agreed, agreed, or completely agreed that the feedback supported them to lead and mimic the emotions. The feedback also helped them to become more aware of their own emotions. The self-assurance scores are about uniformly distributed over the participants. Five out of 12 participants completely disagreed, disagreed, or mildly disagreed that they were able to mimic the requested emotions in the given tasks. This factor supports the relevance of this study, which focuses on the training of acting skills and communication skills. All participants except two agreed that the instructions for the given tasks were clear to them to perform the tasks. All the tasks were completely attractive, attractive, or mildly attractive for the participants to perform. Participants indicated no distraction during performance. None of the participants (except for two) regarded himself as an actor and none had any clear idea about the associated skills.

7. DISCUSSION

This study validated the multimodal emotion recognition software module of FILTWAM by contrasting the software results with ratings from two human experts. We proposed a hybrid fusion method that combines the future-level and decision-level fusion methods. The hybrid method for multimodal fusion of our multimodal software shows accuracies between 96.1% and 98.6% for the best-chosen WEKA classifiers over predicted emotions. In contrast to our previous studies on the unimodal approach of the face emotion recognition (accuracy 72%) and the voice emotion recognition (accuracy 67%), our multimodal approach provides a better accuracy of 98.6% when both modalities are combined in a multimodal data set and likewise are analyzed using the proposed hybrid model. We managed to fulfil our basic requirements of (1) an unobtrusive approach with, (2) inexpensive and ubiquitous equipment (webcam and microphone), that (3) offers real-time and reliable software output that can be customized for and connected to any e-learning environment.

This study showed a substantial agreement between the raters and the multimodal software with regard to the participants' facial and vocal expressions with an overall κ value of 0.761. This κ value indicates that the multimodal software quite successfully uses the participants' facial expressions and vocal intonations for emotion recognition. The best κ value of the recognized emotions among the raters and the multimodal software is neutral 0.837 followed by happy 0.800, anger 0.747, disgust 0.729, fear 0.651, surprise 0.624, and sad 0.538. Here the results show that two of the lesser intensive emotions (neutral and sad) are ranked higher and lower than other emotions. Our data analysis partly falsifies Murthy and Jadon's (2009) finding that the three emotions sad, disgust, and anger are difficult to distinguish from each other and are therefore often wrongly classified. In contrast, our software produces reliable recognition of anger and disgust.

This study showed a substantial agreement between the raters and the face emotion recognition software with regard

TABLE 7
Participants' Opinions

		Answers by the Participants							
		1	2	3	4	5	6	7	Total
Questions									
Difficulty	It was easy for me to mimic the requested emotions in the given tasks	0%	8%	8%	17%	42%	17%	8%	**100%**
Feedback	The feedback did help me to mimic the emotions in the given tasks	0%	0%	16%	17%	33%	17%	17%	
Self-assurance	I am confident that I was able to mimic the requested emotions in the given tasks	8%	17%	8%	25%	17%	17%	8%	
Instructiveness	The instructions for the given tasks were clear to me	0%	0%	8%	8%	25%	34%	25%	
Attractiveness	The given tasks were interesting	0%	0%	0%	0%	33%	50%	17%	
Concentration	I could easily focus on the given tasks and was not distracted by other factors	0%	0%	0%	0%	8%	46%	46%	
Acting skills	I regard myself as a good actor	17%	25%	17%	25%	8%	8%	0%	

Note. 1, completely disagree; 2, disagree; 3, mildly disagree, 4, neither disagree nor agree; 5, mildly agree; 6, agree; 7, completely agree.

to the participants' facial expressions with an overall κ value of 0.644. The best κ value agreement among them is neutral 0.748 followed by happy 0.684, anger 0.666, disgust 0.574, sad 0.540, fear 0.530, and surprise 0.471. This is roughly in agreement with Murthy and Jadon (2009) and Zhang (Zhang, 1999), who found that the most difficult emotions to mimic accurately are fear and sad as these emotions are processed differently from other basic emotions. Moreover, our data analysis confirms Murthy and Jadon's (2009) findings that the neutral and the happy emotions are the easiest emotions to mimic accurately. We have not investigated the issues related to the cultural differences in the judgments of facial expressions of emotion between the raters, the problem that identified by Paul Ekman (Ekman, 1972), and that has been widely investigated by other researchers (e.g., Jack, Garrod, Yub, Caldarac, & Schyns, 2012; Russell, 1994); instead, we have considered the disagreement between the raters for the face emotion recognition in a negotiation session, which was 21% before the negotiation session and was reduced to 12.5% at the end of the negotiation session. This decrease indicates that the influence of cultural differences might be reduced between the two raters after the negotiation session. Moreover, we have not investigated any cultural differences in this study.

However, this study showed a moderate agreement between the raters and the voice emotion recognition software with regard to the participants' vocal intonations with an overall κ value of 0.533. The best κ value agreement among them is neutral 0.619 followed by surprise 0.601, happy 0.586, anger 0.496, fear 0.478, disgust 0.431, and sad 0.321. This is roughly in agreement with Murthy and Jadon (2009) and Zhang (Zhang, 1999), who found that the most difficult emotions to mimic accurately are fear and sad as these emotions are processed differently from other basic emotions. Moreover, our data analysis confirms Murthy and Jadon's (2009) findings that the neutral, the surprise, and the happy emotions are the easiest emotions to mimic accurately.

We invited non-actors for this study in order to avoid extreme emotional expressions that are normally performed by actors. We know that actors might not be able to perform the tasks without exaggeration even though they are instructed. Our assumption was that the participants were comfortable with receiving the feedback during tasks' performance. We know that there would always be some uncertainty for the feedback given to the participants in real situations (e.g., the learner might get interrupted by a person during the tasks performance and this may upset him and consequently affect his performance). With respect to the findings on participants' appreciations of the alpha-release of communication skills training, the participants found that they were not sure if they were able to mimic the requested emotions in the given tasks. They appreciated the emotion feedback as it assisted them to learn and to become more aware of their own emotions. Consequently, being able to

mimic the requested emotions is an important factor that supports the relevance of this study, which focuses on the training of acting skills and communication skills. Therefore, we state that the participant in some cases might not be able to express their natural emotions; however, the two raters as two filters recognized and reported this issue accordingly.

A previous study by Krahmer and Swerts has shown that the use of actors, although they evidently have better acting skills than a layman, will not enhance the realism (i.e., authentic, spontaneous) of 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 multimodal 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 (Huhnel et al., 2014). Nevertheless, because there might be gender differences in older age, upcoming research would comprise older adults.

Finally, one may wonder if the real-time feedback given to the participants during the experimental sessions may have stimulated the participants to adapt their behaviors (i.e., how they act) to the standards exposed by the software, and thereby unwantedly help to raise the observed system's accuracy. In other words, it might be possible that some participants will exaggerate their facial and/or vocal expressions to make the system detect the "correct" emotions. In principle, such internal feedback loop might produce a flattering result for the accuracy. First, however, it should be noted that the two raters expressly removed the exaggerated performances of the participants. Thereby extreme bias is excluded. Second, this multimodal experiment used two independent data sets (face and voice) that were trained beforehand without feedback given to the participants. For establishing a true multimodal data set based on the combined observation (face and voice) of the same persons, we then involved the expert raters. The raters' unimodal outcomes of this study (the new data sets) were then compared with the unimodal outcomes of the fixed data sets. The unimodal accuracy results were very similar, which establishes the validity of the multimodal data set. Overall, we conclude that no influence of the real-time feedback loop on the measured multimodal system accuracy could be established.

8. CONCLUSION

This article described the integration between face and voice emotion recognition software modules covered by the FILTWAM framework. It proposed a hybrid model for multimodal emotion recognition. Hereby FILTWAM may be considered a powerful tool for supporting learning. We continued Sebe's approach (Sebe, 2009) to combine both visual and audio information for classification. We improved the accuracy of the multimodal emotion recognition over detecting one or more basic emotions to 98.6%. Our study has shown that combining two separate modalities into a multimodal approach will improve the accuracy of the software and will provide results that are more reliable. Our approach allows to continuously and unobtrusively monitor learners' behavior during learning activities. It interprets learners' behaviors and converts these into emotional states with high accuracies, in real time, while using domestic devices (webcam, microphone). Hereby FILTWAM may be considered a powerful tool for supporting learning. Moreover, the learners who will use this software in the future will be able to become more aware of their own emotions during tasks' performance. The feedback of our software will assist the learners to obtain this awareness. Although we have considered only seven basic emotions in this study, our software modules can be easily extended for more emotions. The outcomes of FILTWAM could influence different groups' best interests in other settings too.

ACKNOWLEDGMENTS

We thank our colleagues at Welten Institute of the Open University Netherlands who participated in the study of multimodal emotion recognition. We likewise thank the two raters who helped us to rate the recorded video files.

FUNDING

We thank the Netherlands Laboratory for Lifelong Learning (NELLL) of the Open University Netherlands that has sponsored this research.

REFERENCES

Bahreini, K., Nadolski, R., Qi, W., & Westera, W. (2012a). FILTWAM - A framework for online game-based communication skills training - Using webcams and microphones for enhancing learner support. In P. Felicia (Ed.), *The 6th European Conference on Games Based Learning (ECGBL)* (pp. 39–48). Cork, Ireland: Academic Conferences, Ltd.

Bahreini, K., Nadolski, R., & Westera, W. (2012b). 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 (VSGAMES' 12). Procedia Computer Science.15* (pp. 45–52). Genoa, Italy: Elsevier B.V.

Bahreini, K., Nadolski, R., & Westera, W. (2014). Towards multimodal emotion recognition in E-learning environments. *Interactive Learning Environments*. DOI: 10.1080/10494820.2014.908927.

Bahreini, K., Nadolski, R., & Westera, W. (2015). Towards real-time speech emotion recognition for affective E-learning. *Education and Information Technologies*, 1–20. Springer US. DOI: 10.1007/s10639-015-9388-2.

Ben Ammar, M., Neji, M., Alimi, A. M., & Gouarderes, G. (2010). The affective tutoring system. *Expert Systems with Applications*, *37*(4), 3013–3023.

Biswas, P., & Langdon, P. (2015). Multimodal intelligent eye-gaze tracking system. *International Journal of Human–Computer Interaction*, *31*(4), 277–294, DOI: 10.1080/10447318.2014.1001301.

Bosch, N., Chen, H., D'Mello, S., Baker, R., & Shute, V. (2015). Accuracy vs. availability heuristic in multimodal affect detection in the wild. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)* (pp. 267–274). Seattle, WA: ACM.

Brent, M. (1995). *Instance-based learning: Nearest neighbour with generalization*. Hamilton, NZ: University of Waikato, Department of Computer Science.

Buisine, S., Courgeon, M., Charles, A., Clavel, C., Martin, J. C., Tan, N., & Grynszpan, O. (2014). The role of body postures in the recognition of emotions in contextually rich scenarios. *International Journal of Human–Computer Interaction*, *30*(1), 52–62, DOI: 10.1080/10447318.2013.802200.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A database of German emotional speech. *Proceedings of the Inter Speech*, 1517–1520. Lisbon, Portugal.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., . . . Narayanan, S. S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of ACM 6th International Conference on Multimodal Interfaces* (pp. 205–211). New York: ACM.

Castellano, G., Kessous, L., & Caridakis, G. (2008). Emotion recognition through multiple modalities: Face, body gesture, speech, affect and emotion in human–computer interaction. In C. Peter & R. Beale (Eds.), *Lecture notes in computer science 4868* (pp. 92–103). Berlin Heidelberg: Springer.

Chen, L. S., Huang, T. S., Miyasato, T., & Nakatsu, R. (1998). Multimodal human emotion/expression recognition. *Proceedings of the 3rd International Conference on Face and Gesture Recognition*, 366–371.

Chen, L. (2000). Joint processing of audio-visual information for the recognition of emotional expressions in human–computer interaction. PhD thesis. University of Illinois at Urbana–Champaign.

Cohen, W. W. (1995). Fast effective rule induction. *Twelfth International Conference on Machine Learning*, 115–123.

Cooper, D. H., Cootes, T. F., Taylor, C. J., & Graham, J. (1995). Active shape models – Their training and application. *Computer Vision and Image Understanding*, *61*, 38–59.

Cristinacce, D., & Cootes, T. (2004). A comparison of shape constrained facial feature detectors. *IEEE International Conference on Automatic Face and Gesture Recognition (FG'04)*, 375–380.

Cristinacce, D., & Cootes, T. (2008). Automatic feature localisation with constrained local models. *Journal of Pattern Recognition*, *41*(10), 3054–3067.

De Silva, L. C., & Ng, L. C. (2000). Bimodal emotion recognition. *IEEE International Conference on Automatic Face and Gesture Recognition*, 332–335.

D'Mello, S. K., & Graesser, A. C. (2012). AutoTutor and Affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems*, *2*(4), 1–39.

Ekman, P. (1972). Universals and cultural differences in facial expression of emotion. In J. K. Cole (Ed.), *Nebraska Symposium on Motivation* (pp. 207–283). Lincoln, NE: University of Nebraska Press.

Ekman, P., & Friesen, W. V. (1979). *Facial action coding system: Investigator's guide*. Consulting Psychologists Press. https://www.paulekman.com/product/facs-manual/

Frank, E., Hall, M., & Pfahringer, B. (2003). Locally weighted Naive Bayes. *19th Conference in Uncertainty in Artificial Intelligence*, 249–256.

Gaffary, Y., Eyharabide, V., Martin, J. C., & Ammi, M. (2014). The impact of combining kinesthetic and facial expression displays on emotion recognition by users. *International Journal of Human–Computer Interaction*, *30*(11), 904–920, DOI: 10.1080/10447318.2014.941276.

Geertzen, J. (2012). Inter-rater agreement with multiple raters and variables. Retrieved from https://mlnl.net/jg/software/ira/

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2008). Multi-pie. *IEEE International Conference on Automatic Face and Gesture Recognition (FG'08)*, 1–8.

Grubb, C. (2013). Multimodal emotion recognition. Technical Report. Retrieved from http://orzo.union.edu/Archives/SeniorProjects/2013/CS.2013/

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23–34. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/.

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.

Jack, R. E., Garrod, O. G. B., Yub, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, *109*(19), 7241–7244. DOI: 10.1073/pnas.1200155109.

Jaimes, A., & Sebe, N. (2007). Multimodal human–computer interaction: A survey, computer vision and image understanding. *Special Issue on Vision for Human–Computer Interaction*, 108(1–2), 116–134.

Jiang, L., & Zhang, H. (2006). Weightily averaged one-dependence estimators. *Proceedings of the 9th Biennial Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, 970–974.

Krahmer, E., & Swerts, M. (2011). Audio-visual expression of emotions in communication. In *Philips Research Book Series 12* (pp. 85–106). Dordrecht, The Netherlands: Springer.

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*. Heerlen: Open University of the Netherlands.

Le Cessie, S., & van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, *41*(1), 191–201.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kande dataset (CK+): A complete facial expression dataset for action unit and emotion-specified expression. In *Proceedings of the Third IEEE Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010)* (pp. 94–101). San Francisco, CA: IEEE.

Messer, K., Matas, J., Kittler, J., Luuttin, J., & Maitre, G. (1999). XM2VTSDB: The extended M2VTS database. *International Conference of Audio and Video-Based Biometric Person Authentication (AVBPA'99)*, 72–77.

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.

Nadolski, R. J., Hummel, H. G. K., Van den Brink, H. J., Hoefakker, R., Slootmaker, A., Kurvers, H., & Storm, J. (2008). EMERGO: Methodology and toolkit for efficient development of serious games in higher education. *Simulations & Gaming, 39*(3), 338–352. DOI: http://sag.sagepub.com/content/39/3/338.full.pdf+html.

Nwe, T., Foo, S., & De Silva, L. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication*, 41, 603–623.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*(11), 559–572. DOI: 10.1080/14786440109462720.

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.

Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, & A. Smola (Eds.), *Advances in Kernel Methods - Support Vector Learning* (pp. 185–208). Cambridge, MA: MIT Press.

Preeti, K. (2013). Multimodal emotion recognition for enhancing human–computer interaction. PhD dissertation. University of Narsee Monjee, Institute of Management Studies, Department of Computer Engineering. Mumbai, India.

Rus, V., D'Mello, S. K., Hu, X., & Graesser, A. C. (2013). Recent advances in intelligent tutoring systems with conversational dialogue. *AI Magazine*, *34*(3), 42–54.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, *115*, 102–141.

Saragih, J., Lucey, S., & Cohn, J. F. (2011). Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision (IJCV)*, *91*(2), 200–215.

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.

Schuller, B., Lang, M., & Rigoll, G. (2002). Multimodal emotion recognition in audio-visual communication. *IEEE International Conference on Multimedia and Expo, ICME '02*, *1*, 745–748. DOI: 10.1109/ICME.2002.1035889.

Sebe, N., Cohen, I., Gevers, T., & Huang, T. S. (2006). Emotion recognition based on joint visual and audio cues. *18th International Conference on Pattern recognition*, 1136–1139.

Sebe, N. (2009). Multimodal interfaces: Challenges and perspectives. *Journal of Ambient Intelligence and Smart Environments*, *1*(1), 23–30.

Van der Molen, H. T., & Gramsbergen-Hoogland, Y. H. (2005). *Communication in organizations: Basic skills and conversation models*. New York, NY: Psychology Press.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Conference on computer vision and pattern recognition*, I-511-I-518.

Viola, P., & Jones, M. (2002). Robust real-time object detection. *International Journal of Computer Vision*, *57*(2), 137–154.

Vogt, T. (2011). *Real-time automatic emotion recognition from speech: The recognition of emotions from speech in view of real-time applications*. Südwestdeutscher Verlag für Hochschulschriften. ISBN-10: 3838125452.

Wang, S., Ling, X., Zhang, F., Tong, J. (2010). Speech emotion recognition based on principal component analysis and back propagation neural network. In *Proceedings of the 2010 International Conference on Measuring Technology and Mechatronics Automation (ICMTMA '10)*, 03 (pp. 437–440). IEEE Computer Society, Washington, DC, USA.

Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., & Andre, E. (2013). The Social Signal Interpretation (SSI) framework: Multimodal signal processing and recognition in real-time. *Proceedings of the 21st ACM International Conference on Multimedia, MM '13*, 831–834.

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.

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.

Zheng, F., & Geoffrey, I. W. (2006). Efficient lazy elimination for averaged-one dependence estimators. *Proceedings of the Twenty-third International Conference on Machine Learning (ICML 2006)*, 1113–1120.

Zheng, Z., & Webb, G. (2000). Lazy learning of Bayesian rules. *Machine Learning*, *4*(1), 53–84.

ABOUT THE AUTHORS

Kiavash Bahreini is a computer scientist with an interest in affective computing, human–computer interaction, machine learning, real-time applications, data analytics, and e-learning applications; he is a post-doctoral researcher in the Welten Institute, Research Centre for Learning, Teaching and Technology at the Open University of the Netherlands.

Rob Nadolski is an educational technologist with an interest in enhancing learner support facilities, e-learning applications, complex cognitive skills; he is an assistant professor in the Welten Institute, Research Centre for Learning, Teaching and Technology at the Open University of the Netherlands.

Wim Westera is an educational media researcher with an interest in serious gaming and simulation; he is a full professor in the Welten Institute, Research Centre for Learning, Teaching and Technology at the Open University of the Netherlands.