A MULTI-MODAL VIRTUAL ENVIRONMENT TO TRAIN FOR JOB INTERVIEW

Hamza Hamdi[1,2], Paul Richard[1], Aymeric Suteau[1] and Mehdi Saleh[2]
[1] *Laboratoire d'Ingénierie des Systèmes Automatisés (LISA), Université d'Angers, 62 Avenue ND du Lac, Angers, France*
[2]*I-MAGINER, 8 rue Monteil, Nantes, France*

Keywords: Virtual reality, Human-computer interaction, Emotion recognition, Affective computing, Job interview.

Abstract: This paper presents a multi-modal interactive virtual environment (VE) to train for job interview. The proposed platform aims to train candidates (students, job hunters, etc.) to better master their emotional state and behavioral skills. The candidates will interact with a virtual recruiter represented by an Embodied Conversational Agent (ECA). Both emotional and behavior states will be assessed using human-machine interfaces and biofeedback sensors. Contextual questions will be asked by the ECA to measure the technical skills of the candidates. Collected data will be processed in real-time by a behavioral engine to allow a realistic multi-modal dialogue between the ECA and the candidate. This work represents a socio-technological rupture opening the way to new possibilities in different areas such as professional or medical applications.

1 INTRODUCTION

Emotion modeling is a topic in computer science since about twenty years (Picard, 1995). This interest has grown by using breakthroughs coming from neurophysiology and psychology which have established a fine connection between emotion, rationality and decision making (Damasio, 1994). Regarding its expressive dimension, emotions constitute a privileged support in order to model Embodied Conversational Agents (ECAs). Because of their "embodiment", these agents are able to communicate through spoken language but also through gestures and facial expressions. Another interesting but challenging topic is real-time multi-modal communication between ECAs and humans. Different systems have been developed in the last decade (Helmut et al., 2005). However, none of these systems allow realistic immersive multi-modal emotion-based dialogue between an ECA and a human.

In this paper, we describe a multi-modal interactive virtual environment (VE) to train for job interview. The proposed platform aims to train candidates (students, job hunters, etc.) to better master their emotional state and behavioral skills. The candidates will interact with an Embodied Conversational Agent (ECA). The first goal is to measure the emotional and behavior states of the candidates using different human-machine interfaces such as Brain Computer Interfaces (BCI), eye tracking systems, and biosensors. Collected data will allow to analyze nonverbal part of the communication such as posture, facial expressions, gestures and emotions.

In the next section, we survey the related work concerning the classification, the recognition and the modeling of human emotions. In section three we will present the platform architecture, the human-machine interfaces and the proposed multi-modal approach for emotion recognition. The paper ends by a conclusion that provides some ideas for future work.

2 RELATED WORK

2.1 Classification of Emotions

An emotion can be defined as a "hypothetical construct indicating a reaction process of an organization to a significant event" (Scherer, 2000). Emotions are now recognized as involving other components such as cognitive and physiological changes, trends in the

Hamdi H., Richard P., Suteau A. and Saleh M..
A MULTI-MODAL VIRTUAL ENVIRONMENT TO TRAIN FOR JOB INTERVIEW.
DOI: 10.5220/0003401805510556
In *Proceedings of the 1st International Conference on Pervasive and Embedded Computing and Communication Systems* (SIMIE-2011), pages 551-556
ISBN: 978-989-8425-48-5

action (e.g. run away) and motor expressions. Each of these components has various functions. Darwin postulated the existence of a finite number of emotions present in all cultures and having a function of adaptation (Darwin, 1872). This postulate was subsequently confirmed by Ekman which divided the emotions into two classes: the primary emotions (joy, sadness, anger, fear, disgust, surprise) which are natural responses to a given stimuli and ensure the survival of the species. The second class involves emotions that evoke a mental image which correlates with the memory of a primary emotion (Ekman, 1999).

Emotions can be represented by discrete categories (e.g. ”anger”) or defined by continuous dimensions such as ”Valence”, ”Activation”, or ”Dominance”. The ”Valence” dimension allows to describe the ”negative - positive” axis. The ”Activation” dimension allows to describe the ”not very active - very active” axis. The ”Dominance” axis is used to represent the feeling of control. These three dimensions were combined into a space called PAD (Pleasure, Arousal, Dominance) originally defined by Mehrabian (Mehrabian, 1996).

2.2 Emotion recognition

2.2.1 Facial Expression Recognition

Facial expressions provide important information about emotions. Therefore, several approaches based on facial expression recognition have been proposed to classify human emotional states (Pantic and Rothkrantz, 2003). The features used are typically based on local spatial position or displacement of specific points and face regions. Tian (Tian et al., 2000) has attempted to recognize Action Units (AU), developed by Ekman and Friesen in 1978 (Ekman and Friesen, 1978) using permanent and transient features of the face and lips, the nasolabial fold and wrinkles. Geometric models were used to locate the forms and appearances of these characteristics. They reached 96% of precision. Hammal proposed an approach based on the combination of two models for segmentation of emotions and dynamic recognition of facial expressions (Hammal and Massot, 2010). For a complete review of recent emotion recognition systems based on facial expression the readers are referred to (Calvo and D’Mello, 2010).

2.2.2 Speech Recognition

Recognition of the emotional state is a topic of growing interest in the domain of speech analysis. Several approaches that aim to recognize the emotions from speech have been reported (Pantic and Rothkrantz, 2003) (Scherer, 2003) (Calvo and D’Mello, 2010). Most researchers employed global prosodic devices in order to ensure acoustic selection of emotional recognition. Statistics on the expression level are calculated, (e.g. mean, standard deviation, maximum and minimum height and contour of the energy in expressions). Roy and Pentland classified the emotions by using a Fisher linear classifier (Roy and Pentland, 1996). Using short sentences, they have recognized two kinds of emotions: approval and disapproval. They obtained a precision going from 65% to 88%.

2.2.3 Emotion Recognition from Physiological Signals

The analysis of physiological signals is another possible approach for emotion recognition (Healey and Picard, 2000) (Picard et al., 2001). Several types of physiological signals can be used to recognize emotions. For example, heart rate, skin conductance, muscle activity (EMG), temperature variations of the skin, variation of blood pressure are signals regularly used in this context (Lisetti and Nasoz, 2004) (Villon, 2007). Each signal is usually studied in conjunction with other ones.

2.3 Multi-modal Emotion Recognition

Multi-modal emotion recognition requires the fusion of collected data. Physiological signals are then mixed with other signals collected through human-machine interfaces such as video or infrared cameras (gestures, etc.), microphones (speech), brain computer interfaces (BCIs) (Lisetti and Nasoz, 2004) (Sebe et al., 2005). Multi-modal information fusion may be performed at different levels. Usually the three following levels are considered (see Fig. 1): Signal level, Feature level, and Decision or Conceptual level.

Fusing information at the signal level actually means to mix two or more, generally electrical, signals. The signals are fused before extracting features required by the decision maker. This method is not possible on signals issued from different modalities because of the difference regarding their nature. However, this method can improve accuracy by using multiple sensors for a single modality (Paleari and Lisetti, 2006) (fig. 1. a).

Fusing information at the feature level means to mix together the features issued from different signal processors. Features extracted from each modality are fused before being passed to the decision maker module. This is the fusion technique used by humans while combining information from different modalities (Pantic and Rothkrantz, 2003). The input signals of the decision maker have to be synchronized. Not all signals are present at the same time, but could be dephased (fig. 1.b).

Combining information at the conceptual level does not mean mixing together features or signals but directly the extracted semantic information. The decision from these modalities is then combined using each apriori rule or machine learning technique. In this technique, it is possible to extend each modality independently before finally putting together the various decisions (fig. 1.c).

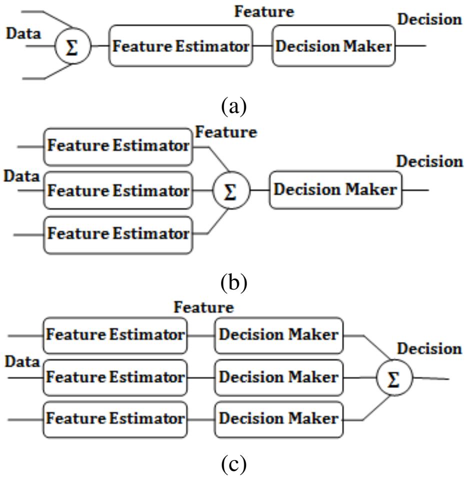
Decision level fusion of multi-modal information is preferred by most researchers. Busso (Busso et al., 2004) compared the feature level and the decision level fusion techniques, observing that the overall performance of the two approaches is the same.

Our goal is to propose a model allowing to analyze various signals and to build a real-time emotion detection system based on multi-modal fusion. We aim to identify the six universal emotions proposed by Ekman and Friesen (Ekman and Friesen, 1978) (anger, disgust, fear, happiness, sadness and joy), to which we add despise, stress, concentration and excitation.

3 SYSTEM OVERVIEW

3.1 System Architecture

The proposed platform aims to support real-time simulation (fig. 2) that allow emotion-based face-to-face dialogue between an ECA (the virtual recruiter) and a human (candidate). Different virtual environments (office, lobby, bar, etc.) can be selected. In addition, the selected ECA can have specific personality and behavior (gentle, aggressive, passive, etc.)



(a)

(b)

(c)

Figure 1: Three levels for multimodal fusion (Sharma et al., 1998): a) data or signal level, b) feature level, c) decision level.

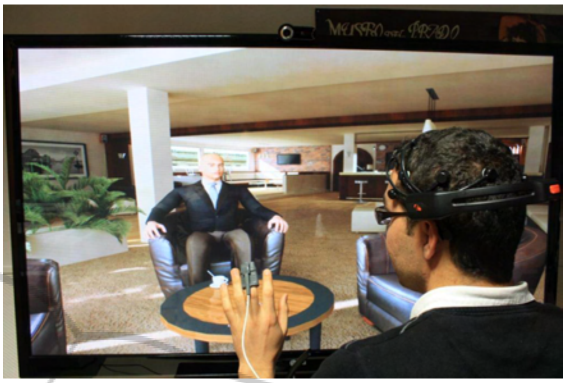


Figure 2: Snapshot of interview simulation.

The ECA is able to communicate using gestures and facial expressions. The interview follows a script including a predefined number of topics to be addressed during the interview. In order to make the simulation more realistic and less predictable, the ECA will adapt the predefined scenario depending on the emotional state and current behavior of the candidate. The proposed platform is based on a specific architecture extracted from the SAIBA (Situation, Agent, Intention, Behavior, Animation) (Vilhjalmsson et al., 2007) framework (see fig. 3):

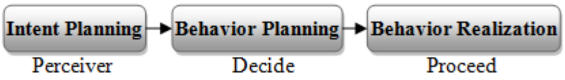


Figure 3: Basic concept of SAIBA Framework.

The SAIBA model is split into three main modules, from a high-level point of view:

- The perception module (Intent Planning), allowing the ECA to obtain information about its environment and its interlocutor,
- The decision module (Behavior Planning), which chooses the best reaction according to what the ECA perceives as well as other parameters do such as its memory,
- The action module or rendering (Behavior Realization) that generates the behavior and the sentences selected by the previous module.

The system architecture is not limited to an ECA, these three modules are integrated into more interconnected modules (see fig. 4):

AI module understands the decision engine of the ECA (the Behavior planning), but also AI from the

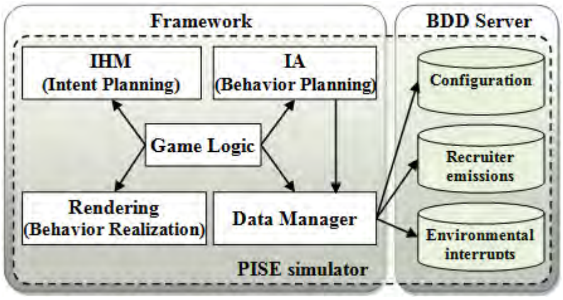


Figure 4: System architecture.

environment. Similarly, the Graphical and Audio rendering engine contains the rendering engine of our virtual recruiter, but also the management of 3D environment display and User Graphical Interfaces (display of the connection screen, menus, etc.).

3.2 Human Computer Interfaces

Human-machine interfaces have two main objectives. The first one consists in collecting data related to the behavioral and emotional states of the candidate. These data will be directly used to control the reaction of the ECA during the interview session. The main challenge is to identify and classify the behavior and the emotional state of a candidate in order to make them interpretable by the ECA. The second objective is related to the ergonomic aspect of the simulation. Thus human-machine interfaces have to be non-intrusive in order to enable a high level of immersion and not constraint the user movements. The following interfaces and sensors have been selected:

≻ Brain computer interface: EPOC (fig. 5 (a));

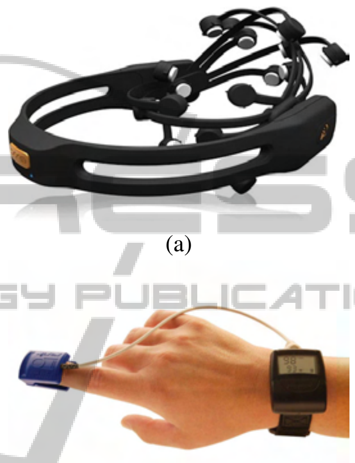
≻ Biofeedback sensor: Nonin (fig. 5 (b));

≻ Microphone and webcam.

3.2.1 Modalities for Emotion Recognition

Human beings are able to communicate and express emotions through various channels involving facial expressions, speech, static and dynamic gestures (Picard, 1995). Different signals and input modalities have been considered:

1. Physiological Signals:
   - Facial Electromyography (EMG) : facial expression,
   - Electrocardiogram (ECG) : heartbeat-related information,
   - Electroencephalography (EEG) : discharge of neurons in the brain,
   - The galvanic skin response (GSR) : electrical resistance of the skin.
2. Speech: the user's emotional state is estimated through speech analysis (pitch, tone, speed).
3. Text: the user's emotional state is estimated through textual content.
4. Gestures: the user's emotional state is estimated through static and dynamic gestures.



(a)

(b)

Figure 5: Human-machine interfaces: (a) Emotiv EPOC, (b) Nonin oximeter.

3.2.2 Annotation and Standardization of Emotions

Emotion Mark-up Language (EmotionML) uses XML for data and annotation information. XML (Extensible Mark-up Language) is a general-purpose specification for creating custom mark-up language. XML includes a set of rules for encoding documents in machine-readable format. It allows the user to define the mark-up elements and facilitate the information systems in sharing structured data. Luneski and Bamidis proposed an XML annotation technique for the emotional data (Luneski and Bamidis, 2007). Wang also proposed the ecgML (ECG Mark-up Language) to annotate ECG data during the acquisition (Wang et al., 2003).

The objective of EmotionML is to transcribe the expression of emotions using XML language. It is able to represent several types of emotions (basic emotions, secondary emotions and the combination of several emotions). It makes it possible to describe the nature of emotions like intensity. This language is not specific to a model or an approach, it is rather simple and could be used to define beacons such as (<category>, <appraisals>, <dimensions>). EmotionML will be implemented as a communication protocol in our framework.

3.3 Framework for Emotion Recognition

Our approach takes into account the multi-modal nature of emotions and is therefore based on multiple sensors and human-machine interfaces (cameras, biofeedback sensors, microphones, brain computer interfaces, etc.). A diagram of the proposed framework is given in fig. 6. The framework has the following significant modules:

1. Feature extraction,
2. Feature selection/reduction,
3. Classification,
4. Decision fusion : Analyzer,
5. Information annotation.

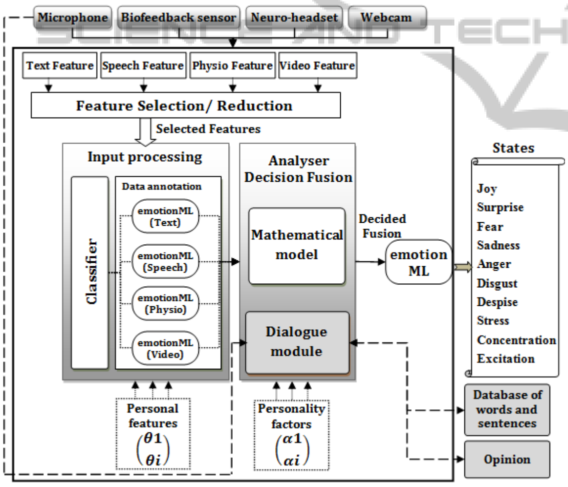


Figure 6: Proposed model for emotion recognition.

Features of the collected data/signals will be extracted from all the channels and the most suitable ones will be selected. Most data/signals are not prepared i.e. are not identified and compared to the behavioral state to which they correspond. Thus, a classifier will accept the selected features and will proceed to the classification of individual input modalities. All information related to classification results for the given stimuli and subject will be annotated/stored using EmotionML.

The decision-fusion module (analyzer) will take the decisions for all the individual channels and perform the data fusion to estimate the candidate emotional state. Thus, the analyzer will integrate them in a mathematical model (equation 1). In fact, assuming that the emotional state at time t depends on the emotional state at time $t-1$, we will sum up pondered data at two successive moments for each human-machine interfaces related to a given emotional state. Finally, the resulted emotions will be stored using EmotionML.

$$\begin{pmatrix} f_1 \\ \dots \\ f_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} (\beta \cdot (\alpha_i \cdot x_i))_t + ((1-\beta) \cdot (\alpha_i \cdot x_i))_{t-1} \\ \dots \\ \sum_{i=1}^{n} (\beta \cdot (\alpha_i \cdot x_i))_t + ((1-\beta) \cdot (\alpha_i \cdot x_i))_{t-1} \end{pmatrix} \tag{1}$$

Where f_i : emotional state i (joy, surprise, fear, sadness, anger, disgust, despise, stress, concentration, excitation).

n: represents the set of human-machine interfaces related to the emotional state i.

α_i: coefficient to be determined by experiments.

$$\sum_{i=1}^{n} \alpha_i = 1$$

β_i: user-defined coefficient.

x_i : data annotation, filtered from input devices.

4 CONCLUSIONS

We presented a multi-modal immersive interactive virtual environment (VE) to train for job interview. The proposed platform aims to train candidates (students, job hunters, etc.) to better master their emotional state and behavioral skills. An Embodied Conversational Agent (ECA) will be used to enable real-time immersive multi-modal and emotion-based simulations. In order to assess the emotional state of the candidates, different human-machine interfaces and bio-sensors have been proposed. In the near future we will carry out some experiments to calibrate the HMIs and identify the signals they provide for different emotional situations. This work opens the way to new possibilities in different areas such as professional or medical applications, and contributes to the democratization of new human-machine interfaces and techniques for affective human-computer communication and interaction.

REFERENCES

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C., Kazemzadeh, A., Lee, S., Neumann, U., and Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *6th international conference on Multimodal interfaces*, pages 205–211, New York, NY, USA. ACM.

Calvo, R. A. and D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transaction on Affective Computing*, 1:18–37.

Damasio, A. (1994). *L'Erreur de Descartes. La raison des motions*. Odile Jacob.

Darwin, C. (1872). *The expression of emotion in man and animal*. University of Chicago Press (reprinted in 1965), Chicago.

Ekman, P. (1999). *Basic emotions*, pages 301–320. Sussex U.K.: John Wiley and Sons, Ltd, New York.

Ekman, P. and Friesen, W. V. (1978). *Facial Action Coding System: A Technique for Measurement of Facial Movement*. Consulting Psychologists Press Palo Alto, California.

Hammal, Z. and Massot, C. (2010). Holistic and feature-based information towards dynamic multi-expressions recognition. In *VISAPP 2010. International Conference on Computer Vision Theory and Applications*, volume 2, pages 300–309.

Healey, J. and Picard, R. W. (2000). Smartcar: Detecting driver stress. In *In Proceedings of ICPR'00*, pages 218–221, Barcelona, Spain.

Helmut, P., Junichiro, M., and Mitsuru, I. (2005). Recognizing, modeling, and responding to users' affective states. In *User Modeling*, pages 60–69.

Lisetti, C. and Nasoz, F. (2004). Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP J. Appl. Signal Process*, 2004:1672–1687.

Luneski, A. and Bamidis, P. D. (2007). Towards an emotion specification method: Representing emotional physiological signals. *Computer-Based Medical Systems, IEEE Symposium on*, 0:363–370.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292.

Paleari, M. and Lisetti, C. L. (2006). Toward multimodal fusion of affective cues. In *Proceedings of the 1st ACM international workshop on Human-Centered Multimedia*, pages 99–108, New York, NY, USA. ACM.

Pantic, M. and Rothkrantz, L. (2003). Toward an affect-sensitive multimodal human-computer interaction. volume 91, pages 1370–1390. Proceedings of the IEEE.

Picard, R. (1995). *Affective Computing, rapport interne du MIT Media Lab, TR321*. Massachusetts Institute of Technology, Cambridge, USA.

Picard, R., Vyzas, E., and Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191.

Roy, D. and Pentland, A. (1996). Automatic spoken affect classification and analysis. automatic face and gesture recognition. In *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition (FG '96)*, pages 363–367, Washington, DC, USA. IEEE Computer Society.

Scherer, K. R. (2000). *Emotion. in Introduction to Social Psychology: A European perspective*, pages 151–191. Blackwell, Oxford.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(7-8):227–256.

Sebe, N., Cohen, I., and Huang, T. (2005). *Multimodal Emotion Recognition*. World Scientific.

Sharma, R., Pavlovic, V. I., and Huang, T. S. (1998). Toward multimodal human-computer interface. *roceedings of the IEEE*, 86(5):853–869.

Tian, Y., Kanade, T., and Cohn, J. (2000). Recognizing lower face action units for facial expression analysis. pages 484–490. Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00).

Vilhjalmsson, H., Cantelmo, N., Cassell, J., Chafai, N. E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A. N., Pelachaud, C., Ruttkay, Z., Thorisson, K. R., van, H. W., and van der, R. J. W. (2007). The behavior markup language: Recent developments and challenges. In *Intelligent Virtual Agents*, pages 99–111, Berlin. Springer.

Villon, O. (2007). *Modeling affective evaluation of multimedia contents: user models to associate subjective experience, physiological expression and contents description*. PhD thesis, Thesis.

Wang, H., Azuaje, F., Jung, B., and Black, N. (2003). A markup language for electrocardiogram data acquisition and analysis (ecgml). *BMC Medical Informatics and Decision Making*, 3(1):4.