Proceedings of the World Conference on Innovative VR 2009
WINVR09
February 25-26, 2009, Chalon-sur-Saône, France

XXXXXXXXXX

TOWARDS AN ARCHITECTURE MODEL FOR EMOTION RECOGNITION IN INTERACTIVE SYSTEMS: APPLICATION TO A BALLET DANCE SHOW

Alexis Clay
ESTIA Recherche
Bidart, France

Nadine Couture
ESTIA Recherche
Bidart, France

Laurence Nigay
LIG-IMAG
Grenoble, France

ABSTRACT

In the context of the very dynamic and challenging domain of affective computing, we adopt a software engineering point of view on emotion recognition in interactive systems. Our goal is threefold: first, developing an architecture model for emotion recognition. This architecture model emphasizes multimodality and reusability. Second, developing a prototype based on this architecture model. For this prototype we focus on gesture-based emotion recognition. And third, using this prototype for augmenting a ballet dance show.

We hence describe an overview of our work so far, from the design of a flexible and multimodal emotion recognition architecture model, to a presentation of a gesture-based emotion recognition prototype based on this model, to a prototype that augments a ballet stage, taking emotions as inputs.

INTRODUCTION

Whereas interaction among human beings heavily relies on emotional communication, interaction with computers yet lacks the emotional side of communication and is hence unnatural to us. The field of affective computing [18] emerged in the beginning of the nineties to address this issue: giving the computer a way to assess a user's affective state (i.e. be able to decode the affective signals a human might send – emotion as an input) and to provide affective feedback (sending signals that a human could decode – emotion as an output).

Emotion recognition is a young field that does not yet benefit from dedicated models. Our goal is threefold: providing with an extendable, reusable, and multimodal emotion recognition architecture model; building a gesture-based emotion recognition system prototype based on this architecture model; and finally, emotionally augmenting a ballet dance show. As for the last point, we aim at providing a better artistic experience for the audience. Though not being emotion synthesis, this part relies on some emotion-inducing phenomena.

In this paper, we will provide an overview of our work in progress: we will first describe our architecture model and the gesture-based emotion recognition prototype. We will then present the Shadoz application for virtually augmenting a ballet dance show. All of these works are in progress, and we will present the identified limitations of our current system.

1 A GENERIC ARCHITECTURE MODEL FOR EMOTION RECOGNITION

In order to integrate emotion recognition in any interactive system, we define a generic software architecture model. Few generic models exist for emotion recognition [4] [16]. We define a flexible and multimodal architecture model, the "emotion branch". We limit ourselves in the definition of this model to emotion interpretation of physical cues from the user, such as face expressions, vocal characteristics, ANS, gesture and posture. We don't aim at covering each of these modalities but rather provide a model for implementing a multimodal emotion recognition system.

The emotion branch model is a collection of software components functioning as black boxes. Those components can emit data flows and subscribe to data flows from other components, not unlike [1]. Our model allows consideration of any of the emotional communication channels: facial expressions, voice modulations, physiological cues, etc. Its flexibility emphasizes reusability, modifiability, and possibilities of distributing the recognition process onto several processes or machines. In this part, we first present the key components that define the emotion branch and then explain how the emotion branch can be plugged into canonical software architecture for interactive systems.

1.1 The emotion branch

Emotion recognition is typically represented as a three layer process: The first layer allows gathering data from the world with different kinds of sensors (e.g. video camera, microphone, motion capture suit). The second layer extracts some emotional cues and characteristics from the signal. Camurri et al. [4] designed a four layer architecture for a video-based software for gesture-based emotion recognition, separating low-levels and high-level cues extraction into two layers. The third layer interprets those cues as an emotion.

We defined the emotion branch as a sequence of three components: the Capture Component, the Analysis Component, and the Interpretation Component, thus mirroring the three layer conception of emotion recognition (see Figure 1). However, each of these components is composed of many subcomponents; the practical use of the capture, analysis and interpretation component is to gather subcomponents sharing the same conceptual role (that is, capturing the data, analyzing it to extract emotional cues, and interpret those cues).

Figure 1. The emotion branch

1.2 The Capture and Analysis Components

A component is not a monolithic software. Each component can be populated by a colony of software units as for context capture with contextors [9]. The Capture and Analysis Components are built in the same way and comprise four kinds of subcomponents:

1. *Capture Units* (in the Capture Component) and *Features Extractors* (in the Analysis Component) are the task-related subcomponents. Capture Units are the interface with a device that captures information from the outside world (e.g. a motion capture suit capturing the movements of a dancer). The captured signal is the output of the Capture Unit. A Capture Unit is dependant on the device. Within the Analysis Component, a Feature Extractor's task is to analyze an incoming flow of data to extract a characteristic which is sent as an output (e.g. giving the arm expansion state –open or closed – from the motion capture data).
2. *Adaptors* transform a data flow: changing the format, adding some parameters, etc.
3. *Fusion Engines* merge totally and permanently equivalent and/or redundant data flows (according to the definition of redundancy in the CARE properties of multimodality [15])
4. *Synchronization Engines* temporally synchronize the different data flows and handles delays.

Those four kinds of subcomponents can be arranged in any way. Each subcomponent is highly specific, but the distribution into many subcomponents facilitates reusability and modifiability.

1.3 The Interpretation Component

Interpreting emotional cues accurately is complex task that depends on many factors. When designing a computer-based interpretation system, one has to consider four points. The first one is the underlying theory and representation of emotion that will be considered. Many theories and representations exist [21], that will affect the output of the interpretation system: labels of emotion and coordinates in a circumplex model (typically the valence-arousal continuous model) are the most used models in emotion recognition.

The second one is the computational method used. Many methods have been used in emotion recognition, such as Hidden Markov Models[11], decision trees[6], neural networks… etc.

The third point is choosing which cues will be inputs to the interpretation system and therefore be evaluated. The fourth one is setting up the parameters for the chosen computational method, according to the cues to be evaluated. However, in the case of machine learning methods, those parameters are automatically determined through a training session of the algorithm.

The Interpretation Component contains two kinds of subcomponents: Interpreters are interpretation systems such as described above. They subscribe to features flows as an input; their role is to interpret them into an emotion in the desired model (label or valence/arousal coordinates). Fusion Engines are subcomponents regrouping the characteristics of adapters and fusion and synchronization engines in the Capture and Analysis Components. This part of the model is still under refinement.

Multimodal emotion recognition allows better, complementary and more robust results. Performing multimodal emotion recognition can be made in two manners: 1) by training an interpreter on evaluating multimodal cues (e.g., over mouth shape and arm extension) or 2) by choosing or somehow finding a compromise between several interpretations. Our architecture model allows to design a multimodal application based on the first or second method, or a combination of both.

1.4 Integration within interactive system architecture models

The emotion branch can be integrated within canonical software architecture models for interactive systems. In Figure 2, we consider two key software architecture models for interactive systems, namely the ARCH reference model [22] and the agent-based MVC model [12]. For adding the emotion branch within the ARCH model, we apply its branching mechanism as shown in Figure 2.a. For the case of the MVC agent of Figure 2.b, we consider a new facet (i.e., the branch emotion) made of three computational elements.

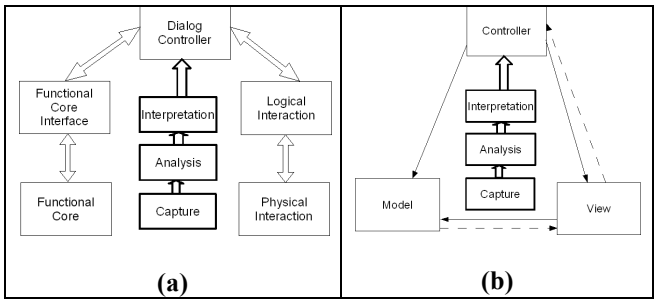

Figure 2. The emotion branch within **(a)** the ARCH model **(b)** the MVC Model.

In Figure 2, the emotion branch is connected to the Dialog Controller of ARCH or to the Controller facet of an MVC agent. This is, however, not always the case. We identified three cases that correspond to different roles that emotion can play in an interactive system.

Case 1: As shown in Figure 2, users' emotion can have a direct impact on the Dialog Controller (DC). The DC has the responsibility for task-level sequencing. Each task or goal of the user corresponds to a thread of dialogue. In this case where the emotion branch is connected to the Dialog Controller, the tasks and their sequence can be modified according to the recognized emotion. For example, in an interactive training system, recognition of sadness or anger of the user (i.e., the learner) could trigger the appearance of a help dialog box about the current exercise. Moreover in our driving simulator [2] as well as in the Multimodal Affective Driver Interfaces [14], alarms are presented according to the current recognized state of the driver, modifying the task-level sequencing and therefore the Dialog Controller.

Case 2: The recognized emotion can be manipulated by the Functional Core branch (i.e., Functional Core Interface and Functional Core components of ARCH) as shown in Figure 3.a. The recognized emotion is therefore a domain object. This is the case in our augmented ballet dance show (section 4) where the recognized emotion conveyed by the dancer is presented to the audience.

Case 3: The detected emotion can have an impact over the Interaction branch as shown in Figure 3.b. For example, a recognized emotion might trigger the change of output modalities (e.g., reducing the frustration of the user). For input interaction, emotion detection could for example imply a dynamic change of the parameters of the speech recognition engine, making it more robust.

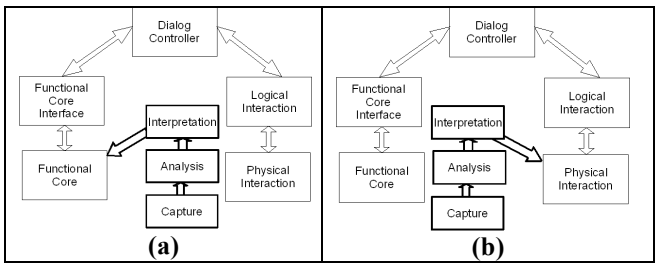

Figure 3. (a) Emotion branch connected to the Functional Core branch. **(b)** Emotion branch connected to the Interaction branch.

Having presented our generic architecture model, we now present how we apply it to the case of emotion recognition based on body movements.

2 EMOTION RECOGNITION FROM BODY MOVEMENTS

2.1 Underlying concepts

For applying our architecture model, we first need to define the considered emotions, the movement characteristics used for recognizing the emotions and the interpretation mechanism.
Many conducted studies both in computer science (e.g. [7]) and psychology (e.g. [10]) identify some emotional movement characteristics. Our study is based on the definition of emotions from [20]. According to this definition, emotions are intense; they do not last long and can change rapidly. They are focused on the triggering stimulus; appraisal of these stimuli may give birth to an emotion which in turn impacts the behavior of the whole body. This definition relates to Ekman's well established definition of the six basic emotions: joy, anger, sadness, surprise, disgust, and fear. Those emotions and their physiological and physical expressions are inherited from an evolutionary process, making them good candidates when considering emotion recognition from body movements.
For recognizing these six basic emotions, as a first step, we consider the movement characteristics defined in [10]. We therefore consider four features: arm position (open/ undetermined/ closed), sagittal direction (forward/ undetermined/ backward), vertical direction (upward/ undetermined/ downward) and speed (slow/ undetermined/ fast). The interpretation mechanism that defines the recognized emotions according to the extracted features is based on rules.

2.2 Implementation

In order to illustrate our "emotion branch" model, we implemented a system composed of three main components as advocated by the emotion branch (Figure 1 and 4) in C++ language using TrollTech's Qt library (http://trolltech.com). The corresponding whole recognition process is undergone at each frame: the software receives position data from the sensors. It then computes the movement characteristics and interprets them to define an emotion for the current frame. The system relies on
Qt's signal/slot functionality for implementing the data flows connecting the different subcomponents.

The capture is based on Xsens Moven motion capture suit and a polhemus liberty 6-DOF sensor. The Moven suit delivers rotation and position coordinates for the whole body in real time. The polhemus sensor is more precise at detecting an absolute position that the Moven. As such, we choose to replace

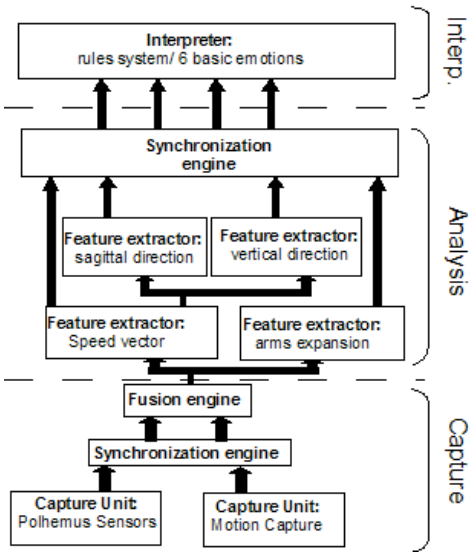

Figure 4. the eMotion architecture.

the user's body absolute position with the polhemus sensor. This is done with the fusion engine (see Figure 4). The two data flows are synchronized before fusion. The resulting flow provides all the information from the Moven suit, but with better accuracy concerning where the user actually is.
The Moven software features the ability to send the coordinates over the network through UDP packets. As such, the Moven software can be considered a Capture Unit in our architecture, deported on another machine (relatively to the rest of the eMotion program).

From the motion data flow, four feature extractors have been developed to form the Analysis Component. We compute the arm expansion (open/closed) and the speed vector of the basin's movement. This vector's norm is computed to get the speed feature (slow/fast), and the vector's value is used to determine the sagittal direction (forward, still, backward) and the vertical direction (downward, still, upward) of the movement. “Arms expansion” and “speed” feature extractors subscribe to the motion data flow that is sent by the Capture Component. “sagittal direction” and “vertical direction” feature extractors subscribe to another feature extractor output flow: they use the speed vector to determine their own outputs. The outputs of those fours feature extractors are then synchronized before being sent to the Interpretation Component.

The interpretation component is based on rules. It subscribes to the four flows of extracted movement characteristics and defines, as an output, an emotion. The mapping function is a sequence of characteristic value cases. Each case returns an emotion among Ekman's six basic emotions: anger, fear, joy, sadness, disgust, surprise. Our current interpreter is rather simplistic in its recognition rules but provides an input for augmenting a ballet dance show.

The current implementation of the eMotion software has yet to be refined to produce conclusive results on emotion recognition. The emotion branch architecture, upon which eMotion is based, allows to quickly plug in new subcomponents after having developed them. Future works on the eMotion system will include adding new features to be extracted from movement and a more accurate way of interpreting them.

3 AUGMENTING A BALLET DANCE SHOW: THE SHADOZ APPLICATION

The emotion expressed by a dancer is not experienced by the dancer himself (although some sort of self-inducement may occur, as for certain schools of acting) but rather performed with the goal of inducing emotion in the audience. Through our eMotion software, we seek to interpret emotional features of gesture and positions found in psychology (such as [8] and [10]) and computer-based emotion recognition (such as [7]) literature. To conduct such studies and discriminate emotional features of movement, researchers usually rely on acted [10] or computer generated [8] material which is then evaluated. As such, those studies test the emotional perception of gesture features rather than testing whether those features are expressions of an experienced emotion. As we base ourselves on these studies, our emotion recognition system can be seen as an evaluator of such perceived features in order to interpret an emotion. Most of the gestural emotional features found in the literature are not context-related. Applying them to ballet dance will allow studying their suitability in such a context.

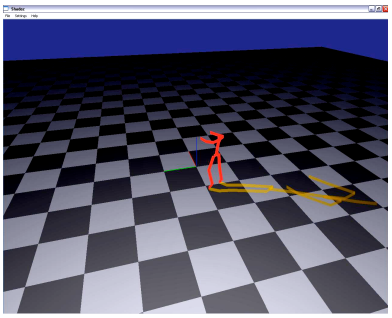

Figure 5. the Shadoz application.

We developed an alpha-version of the Shadoz application for augmenting a ballet dance show. The Shadoz application uses the movement recorded from the dancer and the emotion interpreted by the eMotion software to project a real-time computed shadow on stage as presented in figure 5. Captured motion from the dancer is sent in real-time through UDP sockets over the network. It is hence possible to use those coordinates in eMotion as well as in Shadoz. eMotion interprets the coordinates to extract an emotion that is sent to the Shadoz application over the network (also through an UDP socket). The Shadoz application listens to the network. At each frame, it

receives the captured motion information from the Moven suit, and the corresponding emotion from eMotion.

The literature about visual expression of emotions, through colors or size ratio is rich but often diverges. We based ourselves on the work by Faber Birren [3] as his studies were conducted in the frame of western culture. However, colors are verbally defined ("red", "blue"); we arbitrarily chose the exact color corresponding to those words. To add contrast between the dancer and its shadow, we introduced a size ratio for each emotion (see table 1). This ratio multiplies the normal size of the shadow to make it appear bigger or smaller.

Colour	Emotion	Ratio
Red	Anger	1/n
Black	Fear	1/n
Purple	Sadness	1/n
Yellow	Disgust	1/n
Brown	Surprise	n
Pink	Joy	n

Table 1. chosen associations between colours, emotions and ratios. N is a constant.

In our application, a joyful dancer will trigger a sad shadow. This shadow will hence be smaller than the dancer and will be colored in purple. A sad dancer will trigger a joyful, pink and bigger shadow. An example of Shadoz' output is presented figure 5.

4 DISCUSSION AND FUTURE WORKS

The current state of our research presents some limitations and points to be deepened. The architecture model needs to be refined and heavily thought on several key points. Our main focus is on refining the notion of interpreter in a less bulky way.

The eMotion software will also be subject to major modifications. In its current version, recognition results are not significant due to a simplistic feature extraction and interpretation. However, it has been built according to the model described in this paper. As such, it is an easy task to add features to be extracted and replace the interpretation system.

The Sadoz application is also an alpha-version and as such needs refinement in its mapping from emotion to visual augmentation. Better visual rendering is needed to provide the ballet choreographers with an aesthetically convincing augmentation.

These current systems work and communicate with each other. The eMotion software successfully implements the emotion branch architecture model. From a recognition point of view though, a better interpretation than the current one has to be implemented.

5 CONCLUSION

In this paper we have presented an overview of our work in progress for augmenting a ballet dance show, from the design of a flexible and multimodal architecture system for emotion recognition, through the implementation of a first prototype of gesture-based emotion-recognition to a prototype for augmenting a stage during a ballet dance show. Further versions of these prototypes will enable choreographers from the Ballet de Biarritz to explore - in an artistic sense - the differences and similarities between the emotion that is captured from features not necessarily related to dance, the emotion that is communicated through dance, and the emotion that is perceived by the audience. We give two distinct roles to the system. The first role is the one of a system that recognizes emotions. This recognized emotion is the emotion perceived by the system and by statistically most of the audience. As such, the system plays the passive role of a listener. Its second role is an active role of emotion inducer.

ACKNOWLEDGMENTS

We would like to thank Elric Delord, research engineer in ESTIA, who developed the Shadoz application. This application was developed within the frame of the CARE project, funded by the French National Research Agency.

REFERENCES

[1] Camurri, et al., Toward Real-Time Multimodal Processing: EyesWeb 4.0*, Proc. Artificial Intelligence and the Simulation of Behaviour* (AISB) 2004 Convention: Motion, Emotion and Cognition, Soc. for the Study of Artificial Intelligence and the Simulation of Behaviour (SSAISB), 2004, pp. 22-26.

[2] Benoit, A., et al., Multimodal signal Processing and Interaction for a Driving Simulation: Component-Based Architecture. In *Journal on Multimodal User Interfaces*, 2007, Vol. 1, No. 1, Springer, pp. 49-58.

[3] Birren, F.,*Color psychology and color therapy*. University Books Inc., New Hyde Park, New York, 1961

[4] Camurri A., De Poli G., Leman M., Volpe G., "A Multi-layered Conceptual Framework for Expressive Gesture Applications". In *Proc. of Intl. Workshop on Current Research Directions in Computer Music*, Barcelona, Spain, 2001 pp. 29-34.

[5] Camurri, A., Lagerlöf, I., Volpe, G., Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. In *International journal of human-computer studies*, 2003, Vol. 59, No 1-2, pp. 213-225.

[6] Camurri, A., Mazzarino, B., and Volpe, G. 2004. Expressive interfaces. In *Cogn. Technol. Work 6*, 1 (Feb. 2004), 15-22.

[7] Castellano, G., Villalba, S., & Camurri, A. (2007). Recognising human emotions from body movement and gesture dynamics. In A. Paiva, R. Prada, R. W. Picard (Ed.), *Affective Computing and Intelligent*

Interaction, Second International Conference, ACII 2007, Lisbon, Portugal, September 12-14, 2007, Proceedings: LNCS, vol. 4738 (pp. 71-82). Berlin: Springer-Verlag.

[8] Coulson, M., Attributing emotions to static body postures: recognition accuracy, confusions, and viewpoint dependence. In *Journal of Non verbal Behavior*, 2004, Vol. 28, pp. 117-139.

[9] Coutaz, J., Rey, G., Foundation for a Theory of Contextors. In *Proc. of CADUI'2002*, May, 15-17, 2002, France, Kluwer, pp. 13-34.

[10] De Meijer, M., The contribution of general features of body movement to the attribution of emotions. In *Journal of Non verbal Behavior*, 1989, Vol. 13, pp. 247-268.

[11] Fernandez, R. Picard, R.W., Signal processing for recognition of human frustration, In *Proc. of the 1998 IEEE International Conference on Acoustics*, Speech and Signal Processing, 1998, Seattle, WA, USA, May 12-15, pp 3773-3776 vol.6

[12] Krasner, G., Pope, S., A Cookbook for Using the Model-View-Controller User Interface Paradigm in Smalltalk-80. In *Journal of Object Oriented Programming*, 1988, Vol. 1, No. 3, pp. 26-49.

[13] Martin, J.-C., d'Alessandro, C., Jacquemin, C., Katz, B., Max, A., Pointal, L., and Rillard, A. 3D audiovisual rendering and real-time interactive control of expressivity in a Talking Head . *In Proceedings, 7th International Conference on Intelligent Virtual Agents*, IVA 2007, 2007, Paris, France.

[14] Nasoz, F., Ozyer, O., Lisetti, C., Finkelstein, N., Multimodal affective driver interfaces for future cars. In *Proc. Of Multimedia'02*, December, 1-6, 2002, France, ACM, pp. 319-322.

[15] Nigay, L. and Coutaz, J., Multifeature Systems: The CARE Properties and Their Impact on Software Design, In *Intelligence and Multimodality in Multimedia Interfaces: Research and Applications*, 1997.

[16] 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* (Santa Barbara, California, USA, October 27 - 27, 2006). HCM '06. ACM, New York, NY, 99-108.

[17] Pantic, M., Rothkrantz, L.J.M., Automatic Analysis of Facial Expression: the state of the art. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, Vol. 22, No. 12, pp. 1424-1445.

[18] Pelachaud, C., Bilvi, M., Computational Model of Believable Conversational Agents. In *Communication in Multiagent Systems: Background, Current Trends and Future*, 2003, M.-P. Huget, Ed., pp. 300–317, Springer, New York, NY, USA.

[19] Picard, R.W. *Affective Computing*, 1997, MIT Press.

[20] Scherer, K. and WP3 Members, Preliminary Plans for Exemplars: Theory (Version 1.0), 2004, In *Humaine Deliverable D3c*, 28/05/2004. Available online at http://emotion-research.net/deliverables/D3c.pdf (Accessed 02/03/2007).

[21] Scherer, K. R, *The neuropsychology of emotion*, Oxford University Press, Oxford/New York, 2000 p.144–151

[22] The UIMS Tool Developers Workshop, A Metamodel for the Runtime Architecture of an Interactive System. In SIGCHI Bulletin, 1992, pp. 32-37.