Toward Multimodal Fusion of Affective Cues

Marco Paleari
Affective Social Computing Group,
Eurecom Institute
2229 route des crêtes
Sophia Antipolis, France
paleari@eurecom.fr

Christine L. Lisetti
Affective Social Computing Group,
Eurecom Institute
2229 route des crêtes
Sophia Antipolis, France
lisetti@eurecom.fr

ABSTRACT

During face to face communication, it has been suggested that as much as 70% of what people communicate when talking directly with others is through paralanguage involving multiple modalities combined together (e.g. voice tone and volume, body language). In an attempt to render human-computer interaction more similar to human-human communication and enhance its naturalness, research on sensory acquisition and interpretation of single modalities of human expressions have seen ongoing progress over the last decade. These progresses are rendering current research on artificial sensor fusion of *multiple* modalities an increasingly important research domain in order to reach better accuracy of congruent messages on the one hand, and possibly to be able to detect incongruent messages across multiple modalities (incongruency being itself a message about the nature of the information being conveyed). Accurate interpretation of emotional signals - quintessentially multimodal - would hence particularly benefit from multimodal sensor fusion and interpretation algorithms. In this paper we provide a state of the art multimodal fusion and describe one way to implement a generic framework for multimodal emotion recognition. The system is developed within the *MAUI framework* [31] and Scherer's *Component Process Theory* (CPT) [49, 50, 51, 24, 52], with the goal to be modular and adaptive. We want the designed framework to be able to accept different single and multi modality recognition systems and to automatically adapt the fusion algorithm to find optimal solutions. The system also aims to be adaptive to channel (and system) reliability.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human Factor—*Human Information Process*

; D.2.11 [**Software Architectures**]: Domain Specific Architectures

; J.m [**Computer Applications**]: Miscellaneous

HCM'06, October 27, 2006, Santa Barbara, California, USA.

General Terms

Human Factors, Design

Keywords

Multimodal Fusion, Affective computing, Emotion recognition, HCI

1. GENERAL INTRODUCTION

During face to face communications it has been suggested that as much as 70% of what we communicate when talking directly with others is through paralanguage (e.g. voice tone and volume, body language) [37, 3, 35, 36].

In an attempt to render human-computer interaction more similar to human-human communication and enhance its naturalness, research on sensory acquisition and interpretation of single modalities of human expressions have seen ongoing progress over the last decade. Computers can now nearly understand speech [45, 56, 9], recognize people [32, 34] and their gestures [63, 37]. These results are rendering current research on artificial sensor fusion of *multiple* modalities an increasingly important research domain in order to reach better accuracy of congruent messages on the one hand, and possibly to be able to detect incongruent messages across multiple modalities (incongruency being itself a message about the nature of the information being conveyed).

Accurate interpretation of emotional signals - quintessentially multimodal - would hence particularly benefit from multimodal sensor fusion and interpretation algorithms.

Indeed, during face to face communications it is demonstrated that little information is linked to the actual words one says, most of the information being carried out by vocal inflection, prosodic information, facial expressions and gestures [37, 3]. Human are naturally able to fuse all those different signals and interpret them to understand the message conveyed, according to the affective information contained in the different modalities and in their congruency accross modalities. Affective information itself is carried by the body language for the 55%, by the voice tone and volume for the 38% and for the remaining 7% by the words one said [35, 36],

Albeit work still needs to be performed in the specific intra-modality fields, e.g. voice, expression, Autonomic Nervous System (ANS) signals or gesture recognition, much progress could be accomplished by the fusion of current existing works and approaches. For example, current voice recognizers cannot understand as much natural speech as

humans mostly because, unlike humans, computer algorithms cannot read lip movements and combine the information of what they have seen with the information of what they have heard [1].

In this paper we provide a state of the art multimodal fusion and describe one way to implement a generic framework for multimodal emotion recognition. The system is developed within the *MAUI framework* [31] and Scherer's *Component Process Theory* (CPT) [49, 50, 51, 24, 52], with the goal to be modular and adaptive. We want the designed framework to be able to accept different single and multi modality recognition systems and to automatically adapt the fusion algorithm to find optimal solutions. The system also aims to be adaptive to channel (and system) reliability.

2. RELATED RESEARCH

2.1 Unimodal Emotion Recognition

There are three main areas where emotion recognition systems have developed: 1) emotion recognition from still images and video; 2) emotion recognition from audio and speech; 3) emotion recognition from *Autonomous Nervous System* ANS signals.

The main effort seems to have been in developing systems to recognize the affective state of a person starting from images and video of their faces. There are several reason for that: firstly, the face provides conversational and interactive signals and humans are good at recognizing emotions from facial expressions; secondly, although the video signal is more complex than (for example) the audio signal it presents less problems relative to environmental noise.

First scientifically relevant works about facial expressions and emotions are the ones of Ekman and Friesel [10, 12, 11] who describe facial expression as combinations of independent basic facial movements or actions referred to henceforth as Action Units (AUs). Current state of the art analyzes facial expressions in term of those combinations or of features like motion energy maps or optical flows and recognizes the right emotions in the 60% to 90% of the cases depending also on the applied constraints (e.g. illuminations conditions, database size, perspective, etc.).

Basically the algorithm at the base of all those recognition systems take into account some facial fea tures and apply spatial classifier (generally Gaussian) to calculate the distances among the current observed features and the features calculated as references.

Colmenarez [7], who uses 4 facial areas and 9 facial features, Pantic [41], who takes into account 25 features as distances and angles from predefined feature points and Sebe [54], who considers 12 facial motion measures as features, are only some of the researchers who have developed systems for recognizing emotional facial expressions from still images characteristics.

Other researchers have developed systems to extract affective information from video signals (see [42, 41, 40] for surveys of the current state of the art). Picard [44] considering features from motion energy maps, Mase [33] using optical flows, and Chen [6] using features like distances and positions represent some of the main approaches that have been taken for video signals.

Some work has also been performed for developing systems for emotion recognition from speech and voice features (see [42] for a review on the state of the art). Currently developed systems reach recognition rates going from 50% to 90% depending on the system and on the applied constraints with an average of 60%, 70% and analyze vocal features like *pitch, intensity, speech rate, pitch contour, phonetic features and frequencies of the main formants* [26, 48, 16, 19].

More recently few works have analyzed the possibility of recognizing emotions from Autonomous Nervous System signals. Those systems analyze features like heart-rate, skin conductivity or blood pressure recognizing basic emotions through machine learning techniques [44, 15, 28, 21, 62].

2.2 Multimodal Fusion

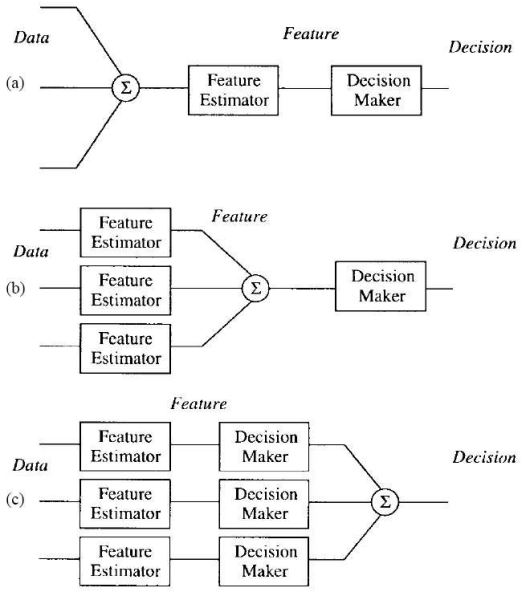

Figure 1: Three levels for multimodal fusion (from Sharma et al. 1998): a) data or signal fusion; b) feature fusion; c) decision fusion

Multimodal information fusion may be performed at different levels and usually the following three are considered (see Figure 1):

- Signal level
- Feature level
- Decision or Conceptual level

Fusing information at the *signal level* means to actually mix two or more, generally electrical, signals. This can only be done for very coupled and synchronized signals that are of the same nature (e.g. two vocal signals, two webcam signals, etc.). For multimodal fusion this is not feasible as different modalities always have different captors and different signal characteristics. General frameworks for fusion should include the possibility to fuse at this first level as different, but similar captors may be needed to get better quality on a single modality (e.g. three sensors for the three red green and blue color components of an image as in high fidelity cams, or microphone arrays).

Fusing information at the *feature level* means to mix together the features outputted by different signal processors.

Features must be pseudo-synchronized in order to provide satisfactory results. For example features can be the position of some feature points extracted from a video processor and the prosodic features of a speech signal. This approach guarantees for multimodal fusion a good amount of exploited information but it has some drawbacks. Combining at the feature level needs synchronization and it is more difficult and computationally intense than fusing at the conceptual level (see next) since the number of features is more important and features may have very different natures (e.g. distances and times). In general HMM (Hidden Markov Model) or time biased NN (Neural Network) are used to fuse at the feature level. In multimodal fusion, feature level fusion can almost always be applied. It can be used to mix information about voice and lip movements for speech recognition or voice and video features for emotion (expression) recognition.

Combining information at the *conceptual level* does not mean mixing together features or signals but directly the extracted semantic information. This implies combining representations obtained from different systems that may also be correlated just at the semantic level (e.g. positions of object, with speech indicating them). Conceptual level fusion has the advantage to avoid synchronization issues and generally to use simple algorithms to be actually computed.

A complete example showing the three possible level of fusion may be a multimodal speech recognition system. Let us assume we want to create a high fidelity gesture and speech recognition system. One of the first things we can do is to implement a microphone array instead of using one single microphone. The audio signals coming from the different microphones can therefore be fused together to get a better audio signal (*signal level fusion*). This improved signal can therefore be treated to extract audio features (e.g. phonemes). Those features may be coupled with video features obtained from a lip movement recognition system (e.g. visemes) (*feature level fusion*). These coupled features can be used to understand the speech part of a voice-gesture command. At a *decision level fusion*, we can fuse the information about what was said with the information coming from a gesture recognizer and understand to which objects some words refer to (as in the sentence "Put that there").

Several works have discussed multimodal fusion; in particular Sharma et al. [55] resume the main issues and techniques of multimodal fusion. Several works on multimodal fusion have been developed which follow the well known *"put that there"* paradigm [4] in which speech and gesture recognition are fused to interact with a 3D environment; see for example works from Bolt, Corradini, Liao and Kettebeckov [4, 8, 27, 20].

2.3 Multimodal Affective Cues Fusion

More recently some works have described how multimodal fusion mechanisms can be used for emotion/affect recognition, see as example works from Pantic, Sebe, Li, Busso and Chen [42, 53, 25, 5, 6] where vocal and facial emotion recognition are performed together to reach better results. Generally fusion is performed at the feature level and the usual clustering techniques are applied to the training data.

Busso et al. in [5] compare the feature level and the decision level fusion techniques, observing that the overall performance of the two approaches is the same, although they present different weakness and strength.

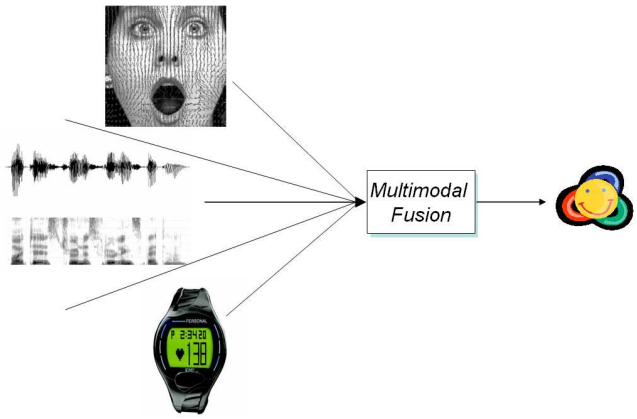

Figure 2: Fusion paradigm for emotion recognition

Busso concludes that the choice of the approach should depend on the targeted application.

For example in the case of multimodal emotion recognition, to fuse at the decision level would mean to mix together $emotion_1$ (extrapolated for example from video signal), $emotion_2$ (extrapolated from audio signal) etc. If the fusing mechanism is reliable enough then the recognition of the found emotion would have more reliability. In other words:

$$R(emotion) = f(emotion_1, .., emotion_n) > R(emotion_i);$$

$$\forall i \in [1, n];$$

Where $R(emotion_i)$ represents the reliability of the recognition of the emotion $emotion_i$. One should note that each $emotion_i$ component derives from different treatment processes and probably from different input signals.

To fuse at a feature level would mean, for example, to mix together $features_1$ which will be a set of n features obtained from video (in the case of the CPT are sets of Ekman's Action Units [10, 12, 11]) to $features_2$ (m features) obtained from audio (with Scherer's theory *pitch, energy, low frequency energy and duration*) and to search clusters in the multi-dimensional space of dimension $n + m$. In this case it is assumed that:

$$R(emotion^*) = f(features_1, .., features_n) > R(emotion_i);$$

$$\forall i \in [1, n];$$

And the results from Busso et al. [5] show that:

$$R(emotion^*) \simeq R(emotion);$$

Nothing prevents from having more than one output from one single raw input that comes from different, or even slightly different, signal processing and/or feature interpretations. At the same time it is possible to have, for example, the $emotion_1$ coming from the treatment of the video signal through AU interpretation and Scherer's theory (see next section), $emotion_2$ coming from audio features treatment,

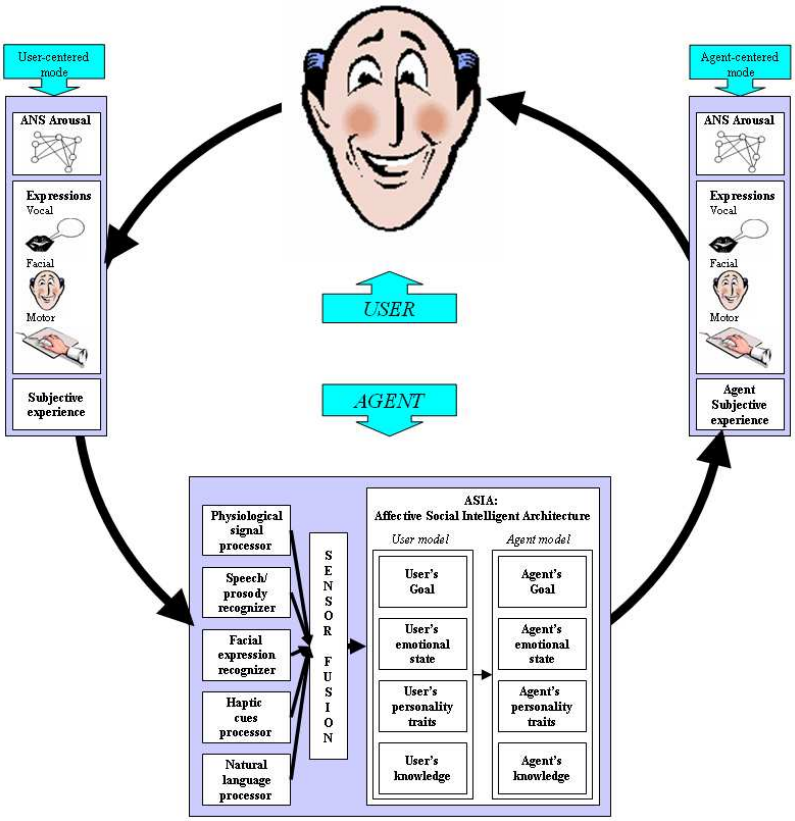

Figure 3: Human Centered Multimodal Affective User Interface. (derived from Lisetti & Nasoz, 2002)

$emotion_3$ coming from video signal treatment according to some eigen-face like recognition algorithm and $emotion_4$ coming from a feature mix of AUs (video) features and prosodic (audio) features.

One of the main limitations of these systems is that they are not upgradeable to new modalities and algorithms. In other words the described fusion algorithm are designed ad hoc for fusing information coming from two (or more) specific unimodal systems but cannot accept new modalities.

Moreover, usually the existing fusion algorithms are not adaptive to the input quality and therefore do not consider eventual changes on the reliability of the different information channels. Another limitation is that the systems cannot take in account long term affective phenomena; in other words they work on short sequences of video and sound (2 to 5 seconds) and cannot consider affective phenomena such as mood or affect of the user which need to be considered over longer times.

Our main objective is then to propose a system where the different fusion algorithms are automatically able to take into account new uni or multi modal emotion recognition systems and to dynamically adapt to the various channel conditions in order to find an optimal (not always the optimum) solution. The recognized emotion will be computed in real time and considered by our cognitive architecture on different time scales in order to get estimations of the different affective phenomena. In other words the estimated appraisals of the affective processes (i.e. 25 estimations per sec) are considered in time windows and averaged to get different estimations for the different affective phenomena (emotions, moods, affects and personalities). Therefore these affective phenomena estimations are used to describe the user affective model.

3. MAUI FRAMEWORK

Our current work builds upon the MAUI (Multimodal Affective User Interface) paradigm [31] designed to give guidelines about the development of multimodal affective user interfaces for Human-Computer interaction focusing on the affective information flow.

As shown in Figure 3, three main parts are identifiable: the User-Centered Mode referring to the various modalities used by humans to express emotions, the ASIA (Affective Social Intelligent Architecture) which processes this emotional information, and finally the Agent-Centered Mode or emotive expression generation for the agent's adaptive expressive behavior.

Looking at the interaction cycle starting from the user and following the counterclockwise arrows one can observe how the affective information can flow from the user to the agent through several different communication channels like the facial expression, the voice expression as well as the *Autonomous Nervous System* (ANS) signals. The agent interprets those signals, and translates them in term of an affective phenomena by fusing the information processed from the different sensors.

The task of the block named ASIA (Affective Social Intelligent Architecture) is therefore to use that affective information together with the knowledge about the environment

and the user (Beliefs) to take social, affective (Emotive) and intelligent decisions (Intentions) in order to reach the agent's objectives (Desires).

These decisions can be designed to have effect on the agent affective state and therefore on the agent expressed emotion, via the various agent-centered modes. All the taken decisions will influence the environment and probably influence the user (most probable in a HCI scenario) who may feel different emotions, display them and so on and so forth.

4. SCHERER'S THEORY

Among the different theory representing emotions, the ways they arise and the way they can be represented, there is one that better than the others link to the MAUI framework: this theory is the *Component Process Theory* (CPT) developed by Scherer [49, 50, 51, 24, 52]. We chose this theory to define our user model and simulate agent emotion generation. There are two main reasons for choosing this particular theory:

1. It considers emotions with their complex three levels (sensory motor, schematic and conceptual) nature.

2. It addresses emotive multimodal expressions[1] and gives guidelines for developing both emotive expression generation and recognition [51].

Scherer's CPT [51] describes emotions as arising from a process of evaluation of the surrounding events with respect to their significance for the survival and well-being of the organism. The nature of this appraisal is related to a sequential evaluation of each event with regards to some parameters called SECs or *Sequential Evaluation Checks*.

According to Scherer's theory, emotion appraisal consists of five components corresponding to the five distinctive functions that justify their existence and the human need for emotions. Those five functions are:

1. Evaluation of objects and events which is related to *information processing* functions

2. System regulation which is related to *support* functions

3. Preparation and direction of action which is related to *executive* functions

4. Communication of reaction and behavioral intention which is related to *action* functions

5. Monitoring of internal state and environment which is related to *monitor* functions

Sequential Evaluation Checks (SECs) [52] are chosen to represent the minimum set of dimensions necessary to differentiate emotions (see last column at right of Figure 6 for the complete SECs list) and they are organized in four classes or in terms of four appraisal objectives. Those objectives are the answers to the following questions:

- How relevant is the event for me? (*Relevance SECs*)

- What are the implications or consequences of this event? (*Implications SECs*)

- How well can I cope with these consequences? (*Coping Potential SECs*)

- What is the significance of this event with respect to social norms and to my self concept? (*Normative significance SECs*)

One of the primary reasons for the sequential approach is to provide a mechanism whereby focusing of attention is only employed when needed and information processing (computational loading) is theoretically reduced.

Scherer in his relatively recent paper [52] discusses the SEC approach within the context of three levels of emotion processing, as also suggested by Leventhal [23], which underlies the design of our affective-cognitive architecture [38], [30]:

- *Sensory-Motor Level*: Checking occurs through innate feature detection and reflex systems based on specific stimulus patterns. Generally it involves genetically determined reflex behaviors and the generation of primary emotions in response to basic stimulus features. For example if something big and black approaches, then the reaction of moving back and the elicited emotion of fear will both belong to this level.

- *Schematic Level*: The learned automatic non-deliberative rapid response to specific stimulus patterns largely based on social learning processes. For example the small talk sentence Good afternoon when meeting someone is a typical behavioral or schematic reaction as the emotions arising from the victory of a sport team.

- *Conceptual Level*: Checking is based on conscious reflective (deliberative) processing of evaluation criteria provided through propositional memory storage mechanisms. Planning, thinking and anticipating events and reactions are typical conceptual level actions. Emotions arise from cognitive processes, as the reproach for a non moral action or anxiety for the result of an important exam.

5. MAUI AND SCHERER

5.1 Emotion Recognition

According to Scherer's theory, it is possible to design different unimodal emotion recognition systems. For example, according to *CPT* [51] human facial expressions are dynamically displayed as consequences of the process of appraising occurring events.

Figure 4 shows the dual link between facial expressions, on the right, and SECs[2], on the left. Scherer's CPT describes predictions about the facial expression showed when appraising a specific SEC.

In that manner for emotion recognition, knowing the involved facial expression action (in terms of AUs), it is possible to retrieve the original eliciting SECs.

The approach we propose shown in Figure 5, is to recognize which facial parameters are changed by comparing the

[1] facial expressions, voice emotive expressions and ANS emotive expressions.

[2] Novelty representing if the event was novel for the agent; intrinsic pleasantness representing if the event is pleasant or unpleasant; goal conduciveness, representing if the event is goal conducive or obstructive and discrepant to the agent goals; control and power representing how much the agent thinks he/she is able to cope with the appraised event

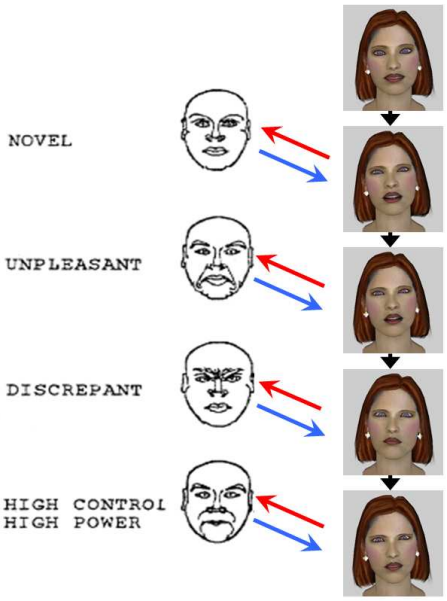

Figure 4: The facial expression process according to Scherer's CPT

current facial expression in terms of some parameters (e.g. feature points or facial motion measures), to the neutral facial state and to deduce which muscles (which Action Units[3] (AUs) [10, 12, 11]) are activated. Finally comparing the activated action units with the predictions given by Scherer's CPT, we can estimate the activated SECs, which represent the recognized emotion.

In a similar way it is possible to recognize emotions from speech (in this case parameters would be *pitch, energy, low frequency energy and duration*) or ANS signals (*heart rate, skin conductivity, salivation, pupil diameter, etc.*) starting from the descriptions of the influence that SECs have to these signals given by CPT [52].

5.2 Emotive Expression Synthesis

An interesting aspect of CPT is that the process used for recognition can also be used for the generation of believable emotive expressions [39, 14]. Looking at Figure 5 one can observe how the process is virtually almost the inverse of the process developed for emotion recognition.

Paleari and Grizard in [39, 14] showed the feasibility of facial animations based on CPT. They converted SECs chains in term of AUs [52, 10, 12, 11]) and then adapted those AUs sequences to show dynamic facial expressions on a graphical animated facial avatar [57, 39] and the Philips iCat robot [58, 14]. Adaptation was necessary since questions about how to fuse the different SECs predictions were raised and because iCat was not able to show all the desired AUs (due to limited degrees of freedom). The results were evaluated by user studies and the developed expressions showed to be recognizable and perceived as believable and not too much exaggerated.

The principal advantage of such an approach is that the cognitive architecture uses the same SECs representation for

[3]Ekman's AUs are defined as the complete set of the minimal independent facial movements that humans can perform

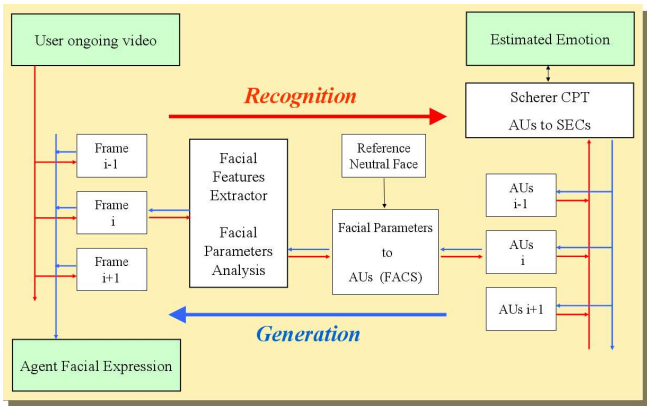

Figure 5: One algorithmic approach to emotional facial expression process based on CPT

the reasoning that are used for recognition and generation of emotive expressions, in the whole system emotions do not have to be linked to labels, e.g. happiness, sadness or fear, maintaining then the emotional nuances that otherwise might be lost.

5.3 Affective Social Intelligent Architecture

The approach we propose includes:

- emotion expression recognition based on Scherer's SECs appraisal model; it takes as inputs video, audio and ANS signals and provides as output SECs chains .
- emotion expression generation basing on the same theory; it takes as input SECs chains and output believable, and psychologically grounded, facial expressions.
- user affective state model and emotion simulation, once again, basing on CPT and allowing complex and believable behaviors.

ASIA (Affective Social Intelligent Architecture) is based on a BDI+E (Belief, Desire, Intention + Emotion) architecture (described by Lisetti and et al. [30]). BDI+E couples the well known Belief Desire and Intention paradigm, [46, 47] to Emotion capabilities at different levels. The architecture is designed on three layers: reactive layer, behavioral layer and deliberative layer and links to Scherer's CPT three levels of emotions.

The internal user emotional model is based on Scherer's CPT SECs [51] allowing the system to reason over complex and complete emotional representations.

6. MULTIMODAL FUSION FRAMEWORK

We hence currently propose a generic framework to perform multimodal fusion at the three possible fusion levels (see Figure 6).

The objective is to develop a system allowing researchers to add their input signal, if missing, and insert their emotion recognition system into a multimodal context. New multimodal systems will be allowed to use signals, or features from other inputs, and would have constraints on the output format (SECs and confidence values).

In our proposition decision or conceptual level fusion is automatically done by the architecture but is also tunable through the modification of simple text files.

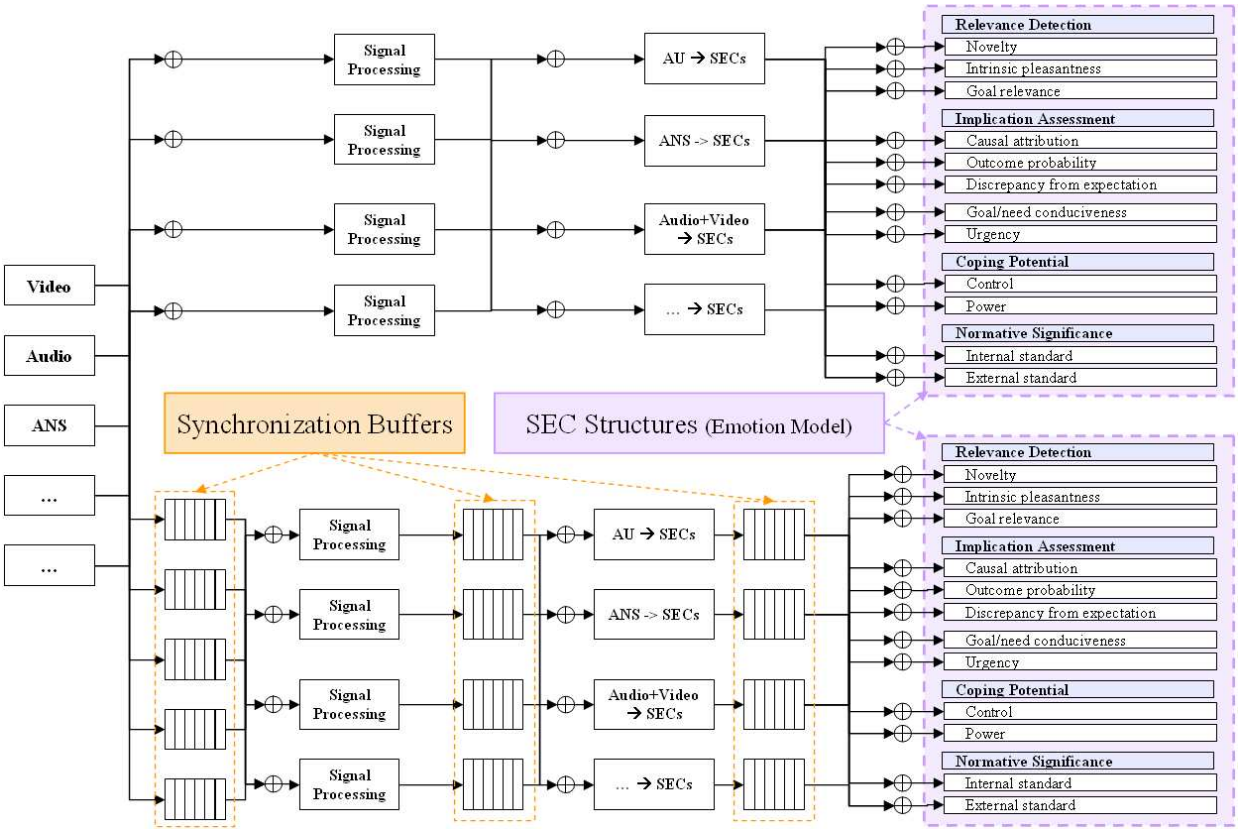

Figure 6: Double chain for signal, feature and decision realignment

In particular the algorithm analyzes different possible fusion algorithms (e.g. maximum, voting, averaging combining, and product combining) and automatically chooses the most stable one during a training phase.

In the future, the algorithm would possibly also be controlled automatically by the cognitive architecture which will be able to deliberately simulate the user's appraisal of the surrounding events and eventually manipulate it.

The framework, shown in Figure 6, has been thought for our Affective Social Intelligent Architecture, and will work on Scherer parameters. The output of each emotion recognition system is a vector representing an appraisal in terms of SECs coupled with a confidence value used by the algorithm to fuse data.

Furthermore two different fusion chains (see Figure 6) would be active in parallel. The first chain, at top in Figure 6, will treat close to real time signals and interpretations returning fast interpretations of the recognized emotion. The second chain will work on the same bufferized and re-aligned data in order to have the possibility to resynchronize data just before fusion.

There are two main reasons for this bufferized approach:

- First there are modalities, like ANS signals that are very interesting and apparently reliable but that have responses time in the order of a dozen of seconds and will not be usable for real time purposes.
- Second we are interested in having a better, more accurate appraisal of the user affective state, regardless the computational time it will take.

In other words the objective of this double chain would be to have both a fast but less reliable and a longer but more accurate evaluations of user affective states. Sensory motor (and behavioral) processes would probably use the fast appraisal while conceptual (and sometimes behavioral) processes would use the longer but more accurate version.

Finally the different fusion algorithms may be able to control the dimensions of the resynchronization buffers. In other words, if one algorithm, comparing the different estimations, observes that the ones coming from ANS signals at time t correspond to those coming from the other multimodal signals at time $(t - n)$ it may control the length of the ANS buffer in order to realign the different evaluations.

Buffers at feature and decision fusion levels are used by the algorithms for searching resynchronization patterns; the signal fusion level buffer is then the one used for the actual resynchronization given the commands coming from the algorithms working on the two higher level buffers.

Constraints would be applied to assure the stability of the system by insuring that buffers length cannot diverge. The resynchronization algorithm working on the signal level buffers will not be able to de-align $signal_a$ and $signal_b$ more than a certain time t_{a-b} or less than the time t_{b-a}.

In both cases of fast and bufferized emotional responses the resulting emotions will be evaluated averaging on different time windows (e.g. 1 sec, 3 sec and 10 minutes) to be able to take into account different affective phenomena, e.g. fast emotional responses like surprise or fear, conceptual emotions like contempt or pride but also moods and affects.

7. CONCLUSIONS

We have presented a generic framework for multimodal fusion for emotion recognition. The objective of such a project is to provide researcher in multimodal emotion recognition with a flexible platform that can accept new recognition modules based on Scherer theories. The algorithm almost automatically takes into account the new recognition system and use it to improve the emotion expression estimations.

The main reasons for basing this approach on Scherer's theory are three:

- Scherer's theory models in a very detailed and psychologically grounded way the appraisal process of emotions.
- CPT allows for a three level model of emotions.
- CPT links both emotion generation and emotion recognition to the process of appraisal and therefore to the user and agent models.

The use of such a theory implies some constraints on inputs and outputs of the recognition systems. We propose to use audio, video and ANS signals as inputs of the fusion framework (but other inputs may be added in future work). The output of each recognition system must be a SECs chain as represented in the last column of Figure 6, plus information about the reliability of the system.

The algorithm we propose takes as input the SECs chains and fuse them. In doing this operation three operations are done:

- The fusion framework automatically takes into account new recognition system.
- The algorithm computes the channels / recognition system reliabilities, comparing the different recognition system evaluations and adapting the algorithm to find optimal solutions.
- The complete system treats in parallel two chains, one near to real-time and one bufferized and more reliable one.

The fusion platform is proposed as part of the implementation of the MAUI framework and in particular will be added as a module of ASIA *Affective Social Intelligent Architecture* which is a three layer architecture designed to better link to Scherer's and Leventhal's theories.

Inside ASIA information coming from the fusion algorithm is averaged on different time scale to take into account different affective phenomena like emotion, mood and personality.

The system we propose allows agents to consider more modalities and to get computer reasoning and emotive expressions closer to the human one. This kind of mechanism is in our opinion fundamental in Human Centered Computing and will help developers and researchers that will not have to develop complete fusion algorithms.

8. REFERENCES

[1] P. Aleksic and A. Katsaggelos. Product hmms for audio-visual continuous speech recognition using facial, 2003.

[2] C. Bartneck, J. Reichenbach, and A. van Breemen. In your face, robot! the influence of a character's embodiment on how users perceive its emotional expressions. In *Proceedings of the Fourth International Conference on Design & Emotion*, Ankara, Turquey, July 2004.

[3] C. Besson, D. Graf, I. Hartung, B. Kropfhusser, and S. Voisard. The importance of non-verbal communication in professional interpretation, 2004.

[4] R. A. Bolt. Put-that-there: Voice and gesture at the graphics interface. In *SIGGRAPH '80: Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, pages 262–270, New York, NY, USA, 1980. ACM Press.

[5] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on multimodal interfaces (ICMI '04)*, pages 205–211, State College, PA, USA, 2004. ACM Press, New York, NY, USA.

[6] L. Chen, H. Tao, T. Huang, T. Miyasato, and R. Nakatsu. Emotion recognition from audiovisual information. In *Proceedings, IEEE Workshop on Multimedia Signal Processing*, pages 83–88, Los Angeles, CA, USA, 1998.

[7] A. Colmenarez, B. Frey, and T. Huang. Embedded face and facial expression recognition. In *Proceedings of ICIP 1999*, volume 1, pages 633–637, 1999.

[8] A. Corradini, M. Mehta, N. Bernsen, and J.-C. Martin. Multimodal input fusion in human-computer interaction on the example of the on-going nice project. In *Proceedings of the NATO-ASI conference on Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management*, Yerevan (Armenia), August 2003.

[9] A. Duminuco, C. Liu, D. Kryze, and L. Rigazio. Flexible feature spaces based on generalized heteroscedastic linear discriminant analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2006.

[10] P. Ekman. Universals and cultural differences in facial expressions of emotion. J. K. Cole, editor, *In Proceeding of Nebraska Symposium on Motivation*, volume 19, pages 207–283, Lincoln (NE), 1971. Lincoln: University of Nebraska Press.

[11] P. Ekman, W. V. Friesen, and J. C. Hager. *Facial Action Coding System Invistagator's Guide*. A Human Face, 2002.

[12] P. Ekman and F. W. *Facial Action Coding System*. Palo Alto (CA), 1978.

[13] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42, 2002.

[14] A. Grizard, M. Paleari, and C. Lisetti. Adaptating psychologically grounded facial emotional expressions to different platforms. In *Proceedings of KI06 26th German Annual Conference in Artificial Intelligence*, Bremen, Germany, 2006.

[15] A. Haag, S. Goronzy, P. Schaich, and J. Williams. Emotion recognition using biosensors: First steps

towards an automatic system. In *Proceedings of LNCS*, pages 36–48, 2004.

[16] R. Huber, A. Batliner, J. Buckov, E. Noth, V. Warnke, and H. Niemann. Recognition of emotions in realistic dialog scenario. In *Proceedings of ICSLP 2000*, pages 665–668, 2000.

[17] I.Poggi, C.Pelachaud, F. de Rosis, V. Carofiglio, and B. D. Carolis. *Multimodal Intelligent Information Presentation*, chapter GRETA. A Believable Embodied Conversational Agent. Kluwer, 2005.

[18] H. Ishiguro. 2006-2056 projects and vision in robotics. In *Proceedings of 50 years AI Symposium at KI06 26th German Annual Conference in Artificial Intelligence*, Bremen, Germany, 2006.

[19] T. Kang, C. Han, S. Lee, D. Youn, and C. Lee. Speaker dependent emotion recognition using speech signals. In *Proceedings of ICSLP 2000*, pages 383–386, 2000.

[20] S. Kettebekov and R. Sharma. Toward multimodal interpretation in a natural speecmgesture interface. In *Proceedings of IEEE Symposium on Image, Speech, and Natural Language Systems*, pages 328–335. IEEE, November 1999.

[21] K. Kim, S. Bang, and S. Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, 42, 2004.

[22] B. J. A. Kröse, J. M. Porta, A. J. N. van Breemen, K. Crucq, M. Nuttin, and E. Demeester. Lino, the user-interface robot. In *EUSAI*, pages 264–274, 2003.

[23] H. Leventhal. A perceptual-motor theory of emotion. *Journal of Advances in Experimental Social Psychology*, 17:117–182, 1984.

[24] H. Leventhal and K. R. Scherer. The relationship of emotion to cognition: A functional approach to a semantic controversy. *Cognition and Emotion*, 1:3–28, 1987.

[25] X. Li and Q. Ji. Active affective state detection and user assistance with dynamic bayesian networks. In *IEEE Transactions on Systems, Man , and Cybernetics. Part A: Systems and Humans*, volume 35, pages 93– 105. IEEE, January 2005.

[26] Y. Li and Y. zhao. Recognition of emotions in speech using short term and long term features. In *Proceedings of ICSLP 1998*, pages 2255–2258, 1998.

[27] H. Liao. Multimodal Fusion. Master's thesis, University of Cambridge, july 2002.

[28] C. Lisetti and F. Nasoz. Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP Journal on Applied Signal Processing*, 11:16721687, 2004.

[29] C. L. Lisetti and P. J. Gmytrasiewicz. Emotions and personality in agent design. In *Proceedings of AAMAS 2002*, 2002.

[30] C. L. Lisetti and A. Maurpang. Bdi+e framework: An affective cognitive modeling for autonomous agents based on scherers emotion theory. In *Proceedings of KI06 26th German Annual Conference in Artificial Intelligence*, Bremen, Germany, 2006.

[31] C. L. Lisetti and F. Nasoz. Maui: a multimodal affective user interface. In *Proceedings of the ACM Multimedia International Conference 2002*, Juan les Pins, December 2002.

[32] C. Mallauran, F. Dugelay, J.L.and Perronnin, and C. Garcia. Online face detection and user authentication. In *Proceedings of the ACM Multimedia Conference 2005*, Singapore, Nov 2005.

[33] K. Mase. Recognition of facial expression from optical flow. In *Proceedings of IEICE Transactions*, volume E74, pages 3474–3483, 1991.

[34] F. Matta and J. Dugelay. Towards person recognition using head dynamics. In *Proceedings of ISPA 2005, 4th International Symposium on Image and Signal Processing and Analysis*, Zagreb, Croatia, September 2005.

[35] A. Mehrabian. *Silent Messages*. Wadsworth Publishing Company, Inc, Belmont, CA, 1971.

[36] A. Mehrabian. *Nonverbal Communication*. Aldine-Atherton, Chicago, 1972.

[37] G. Merola and I. Poggi. Multimodality and gestures in the teachers communication. In *Lecture Notes in Computer Science*, volume 2915, pages 101–111, Feb 2004.

[38] R. Murphy, C. Lisetti, L. Irish, R. Tardif, and A. Gage. Emotion-based control of cooperating heterogeneous mobile robots. *IEEE Transactions on Robotics and Automation: special issue on Multi-robots Systems*, 2001.

[39] M. Paleari and C. Lisetti. Psychologically grounded avatar expressions. In *Proceedings of KI06 26th German Annual Conference in Artificial Intelligence*, Bremen, Germany, 2006.

[40] M. Pantic and L. Rothkrantz. Automatic analysis of facial expression: The state of the art. *IEEE Transaction, Issue on Pattern Analisys and Machine Intelligence*, volume 22, pages 1424–1445, 2000.

[41] M. Pantic and L. Rothkrantz. Expert systems for automatic analysis of facial expression. *Image, Vision and Computing Journal*, 18:881–905, 2000.

[42] M. Pantic and L. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. In *Proceedings of IEEE*, volume 91, pages 1370–1390. IEEE, September 2003.

[43] C. Pelachaud, V. Carofiglio, and I. Poggi. Embodied contextual agent in information delivering application. In *Proceedings of the First International Joint Conference on Autonomous Agents & Multi-Agent Systems*, Bologna, Italy, 2002.

[44] R. Picard. *Affective Computing*. MIT Press, Cambridge (MA), 1997.

[45] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, 1993.

[46] A. S. Rao and M. P. Georgeff. Modeling rational agents within a bdi-architecture. J. Allen, R. Fikes, and E. Sandewall, editors, *In Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91*, pages 473–484, 1991.

[47] A. S. Rao and M. P. Georgeff. Bdi agents: From theory to practice. In *Proceedings of 1st International Conference on Multi-Agent Systems, ICMAS'95)*, page 312319, San Francisco, CA., 1995.

[48] H. Sato, Y. Mitsukura, M. Fukumi, and N. Akamatsu. Emotional speech classification with prosodic parameters using neural networks. In *Proceedings of Australian and NewZealand Intelligner Information System Conference*, pages 395–398, 2001.

[49] K. R. Scherer. Emotion as a process: Function, origin and regulation. *Social Science Information*, 21:555–570, 1982.

[50] K. R. Scherer. Emotions can be rational. *Social Science Information*, 24(2):331–335, 1985.

[51] K. R. Scherer. Toward a dynamic theory of emotion: The component process model of affective states. Geneva Studies in Emotion and Communication, 1(1), 1-98, 1987.

[52] K. R. Scherer. *Appraisal processes in emotion: Theory, methods, research*, chapter Appraisal Considered as a Process of Multilevel Sequential Checking, pages 92–120. New York, NY, US: Oxford University Press, 2001.

[53] N. Sebe, I. Cohen, and T. Huang. *Multimodal emotion recognition*. World Scientific, 2005.

[54] N. Sebe, M. Lew, I. Cohen, A. Garg, and T. Huang. Emotion recognition using a cauchy naive bayes classifier. In *Proceedings of ICPR 2002*, volume 1, pages 17–20, 2002.

[55] R. Sharma, V. Pavlovic, and T. Huang. Toward multimodal human-computer interface. In *Proceedings of the IEEE*, 1998.

[56] V. Tyagi and C. Wellekens. Adaptive enhancement of speech signals for robust ASR. In *ASIDE 2005, COST278 Final Workshop and ISCA Tutorial and Research Workshop*, Aalborg, Denmark, Nov 2005.

[57] Haptek website: www.haptek.com, 2006.

[58] iCat website at Philips: www.research.philips.com/robotics, 2006.

[59] A. van Breemen. Animation engine for believable interactive user-interface robots. In *Proceedings of IROS 2004, 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sendai, Japan, September 2004.

[60] A. van Breemen. Bringing robots to life: Applying principles of animation to robots. In *Proceedings of Shapping Human-Robot Interaction workshop held at CHI 2004*, Vienna, Austria, 2004.

[61] A. van Breemen. icat: Experimenting with animabotics. In *Proceedings of AISB*, pages 27–32, 2005.

[62] O. Villon and C. L. Lisetti. Toward building adaptive users psycho-physiological maps of emotions using bio-sensors. In *Proceedings of KI06 26th German Annual Conference in Artificial Intelligence*, Bremen, Germany, 2006.

[63] Y. Wu and T. Huang. Vision-based gesture recognition: A review. *Lecture Notes in Computer Science*, 1739:103+, 1999.