AFFECT-SENSITIVE MULTI-MODAL MONITORING IN UBIQUITOUS COMPUTING: ADVANCES AND CHALLENGES

Maja Pantic, Leon J.M. Rothkrantz
Delft University of Technology, MediaMatica Department, PO Box 356, 2600 AJ Delft, The Netherlands
Email: M.Pantic@cs.tudelft.nl, L.J.M.Rothkrantz@cs.tudelft.nl

Key words: Human-centred computing, ubiquitous computing, affective computing, facial affect analysis, vocal affect analysis, information fusion, affect-sensitive multi-modal monitoring

Abstract: The topic of automatic interpretation of human communicative behaviour, that is, giving machines the ability to detect, identify, and understand human interactive cues, has become a central topic in machine vision research, natural language processing research and in AI research in general. The catalyst behind this recent 'human-centred computing hoopla' is the fact that automating monitoring and interpretation of human communicative behaviour is essential for the design of future smart environments, next generation perceptual user interfaces, and ubiquitous computing in general. The key technical goals concern determining of the context in which the user acts, that is, disclosing in an automatic way where is the user, what is he doing, and how is he feeling, so that the computer can act appropriately. This paper is pertained with the last of these issues, that is, with providing machines with the ability to detect and interpret user's affective states. It surveys the past work done in tackling this problem, provides taxonomy of the problem domain, and discusses the research challenges and opportunities.

1. INTRODUCTION

One of the key challenges in making human-computer interfaces (HCI) more satisfactory usable and universally accessible is to establish human-computer interaction that captures attributes of human-human communication and approaches its naturalness. Interpersonal interaction is a complex interplay of thoughts, language, and non-verbal communicative signals. If that is the intended model for future smart virtual environments (Thalmann et all 1998), natural HCI (Sharma et all 1998, Marsic et all 2000), and ubiquitous computing in general (Pentland 2000), then this next generation of HCI systems requires translation and emulation of human behavioural cues. Hence, the problems related to facilitating context sensing and understanding (who is the user, where is he, what is he doing, how is he feeling), constructing theories of mind (what does user want, when to interact, and how to adapt the interaction), and facilitating automatic intelligent responding (in which way to interact: which words, intonation, and facial expression to synthesise), have become critical issues in the design and development of the next generation HCI systems. These problems of *human-centred computing* are still far from being settled but, at a minimum, they are among the most exciting and economically important research topics in information technology (Pentland 2000).

Due to numerous areas where benefits could accrue from automating affect-sensitive monitoring of human communicative displays, this aspect of context sensing and understanding attracted interest of many AI researchers. In addition to facilitating more satisfactory usable and universally accessible HCI systems by giving them the ability to sense and respond appropriately to user affective feedback (Picard 1997), automatic affect-sensitive monitoring tools will facile the research in areas as diverse as behavioural science (e.g. in topics discussed in (Bassili 1979), (Ekman et all 1969)), medicine (e.g. in topics like those in (Steimer-Krause et all 1990)) and political sciences (e.g. in topics of (McHugo et all 1985)). Automatic assessment of attitudinal states like boredom, inattention, and stress, will be of high value in preventing critical situations in hazardous working environments like aircraft cockpits, nuclear power plan surveillance rooms, air traffic control towers, or simply in the ground traffic vehicles like trucks, trains, and personal cars. An advantage of affect-sensitive monitoring done by a computer is that human observers need not to be present to perform privacy-intruding monitoring; an automated tool could provide prompts for better performance based on the sensed user's affective state. Besides, automated monitoring will be more accurate since computers possess sensory modalities that humans lack (e.g. the EEG).

This paper examines the past work done in the field of automating affect-sensitive monitoring of human communicative signals. It summarises the relevant issues debated in the psychological research literature (§2), explains the affect-recognition ability of human sensory system (§3), and based upon these findings provides a taxonomy of the problem domain (§4). Then, the paper examines the state-of-the-art (§5), discusses some of the challenges and opportunities facing researchers in this area (§6), and provides the concluding remarks (§7).

2. PSYCHOLOGICAL ISSUES

Since an automated analyser of human affective states would be extremely beneficial, the question of how to best characterize the human perception of affective states has become an important concern for many researchers in *affective computing* (Picard 1997). Ironically, the growing interest in affective computing is coming at a time when the established wisdom on human affect states is being strongly challenged in the basic research literature.

On one hand, the classic psychological research claims the existence of universally displayed and recognized six basic expressions of emotions: anger, happiness, sadness, surprise, disgust, fear (Bezooijen 1984, Ekman 1994). This implies that, except of the verbal communicative signals (spoken words) which are person-dependant (Furnas et all 1987), the non-verbal communicative signals (e.g. facial expression, vocal intonations, body gestures, clamminess, etc.) involved in these basic emotions are displayed and recognized cross-culturally. On the other hand, there is now a growing psychological research that strongly challenges the classical theory on emotion. The psychologist James Russell argues that emotion in general can be best characterized in terms of a multi-dimensional affect space, rather than discrete emotion categories (Russell 1994). Furthermore, it has been shown that the comprehension of a given emotion label and the ways of expressing the related affective state are culture dependent (Matsumoto 1990, Cacioppo et all 2000). In turn, it is not certain that each of us will express a particular affective state by modulating the same communicative signals in the same way, nor it is certain that a particular modulation of communicative signals will be interpreted always in the same way independently of who the observer is.

Consequently, there is no psychological scrutiny on universal expressions of affective states that can be safely assumed and employed in studies on affective computing. One source of help for this problem is *machine learning*: rather than having a priori generic rules for affective state recognition, we can potentially learn the rules by interacting with the user about his/her interpretations of the observed affective displays. Thus, a promising strategy is to build a personalised, affect-sensitive analyser of human communicative signals capable of adapting the employed communicative-signals classification-mechanism according to the user’s wishes.

3. HUMAN PERFORMANCE

Affective arousal modulates all, the verbal- and the non-verbal communicative signals. As shown by Furnas et all (1987), anticipating a person’s word choice and the associated intent is very difficult: even in highly constrained situations different people choose different words to mean exactly the same thing. On the other hand, in usual face-to-face interaction, people detect and interpret non-verbal communicative signals in terms of affective states expressed by their communicator with little or no effort (Ekman et all 1969). Although a correct recognition of someone’s affective state depends on many factors (the attention given to the speaker and the familiarity with the speaker’s personality, face, usual vocal intonation, etc.), humans perform affect recognition with an apparent ease.

A main characteristic of human sensory system for affect recognition is the multi-modal analysis of multiple communication channels. A channel is a communication medium (e.g. the visual channel that carries facial expressions) while a modality is a sense used to perceive signals from the outside world (e.g. the senses of sight and hearing). In usual interpersonal face-to-face interaction, many channels are employed simultaneously and various modalities are activated in combination. As a result, the process of analysing the interaction that takes place becomes highly flexible and robust. Failure of one channel is recovered by another channel and a message in one channel can be explained by another channel (e.g. in noisy environments we can “hear” what has been said by the means of lip reading).

The abilities of the human sensory system define, in some way, the expectations for an automated affect-sensitive monitoring tool. Though it may not be possible to incorporate all features of the human sensory system into an automated alike system, the capabilities of the human sensory system can certainly serve as the ultimate goal and a guide for determining recommendations for the design of an automatic affect-sensitive monitoring tool.

4. PROBLEM TAXONOMY

We can build a taxonomy of the affect-sensitive-monitoring problem domain by considering the observation channels and their time scale. The domain can be analysed by analysing different channels of information that correspond to different human communication channels carrying non-verbal communicative signals displayed by the observed subject. People employ the communication channels in a complementary and redundant manner. Affect-sensitive monitoring tools should perform similarly: different observation channels must be considered together. Furthermore, each observation channel, in general, carries information at a wide range of time scales. At the longest scale are *static and semi-permanent signals* like bony structure, fat deposits, metabolism, and phonetic peculiarities like accent. At shorter time scales are *rapid behavioural signals* which represent temporal changes in neuromuscular and physiological activity that can last from a few milliseconds (e.g. blink) to minutes (e.g. respiration rate) or hours (e.g. sitting). In consequence, an ideal, automated, user-profiled, affect-sensitive monitoring tool will perform (Fig. 1):

generic, time-instance/time-scale analyses of all non-verbal communicative signals, and

user-defined affect-discriminative interpretation of these data previously combined by applying a multi-sensory information fusion.

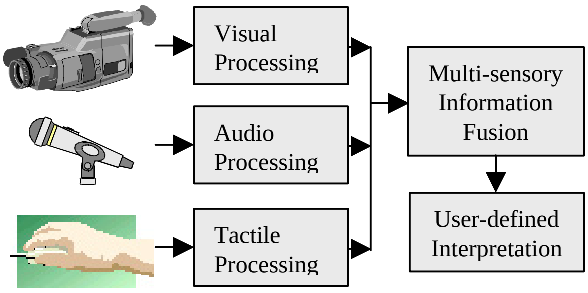

Fig. 1: Architecture of an ideal automated affect-sensitive tool

Since the potential applications of an automated affect-sensitive monitoring tool involve continuous observation of a subject in a time interval, sensing of non-verbal communicative signals should proceed in a fully automatic way. An efficient and effective tool should start with generic analyses of the sensed signals (independently of the subject’s sex, age, ethnicity and personal characteristics). Then, in order to perform a user-defined interpretation of the affective state displayed by the monitored subject, it should adapt to the current user (which might be but does not have to be the motored subject at the same time). Finally, it should perform robustly despite (inevitable) auditory noise, changes in viewing and lightning conditions, and occlusions such as glasses and facial hair.

It is interesting to note that facial- and vocal expressions of attitudinal states are widely thought to be the most important in human communication and human recognition of affect. As indicated by Mehrabian (1968), spoken words contribute for only 7%, vocal utterances for 38%, and facial expressions contribute for even 55% to whether a listener feels liked or disliked. This implies that an automated affect-sensitive monitoring tool should combine, at least, automated modalities for perceiving facial and vocal expressions of attitudinal states.

5. THE STATE OF THE ART

This section will survey current state-of-the-art in the affect-sensitive-monitoring problem domain. Rather than an exhaustive survey, the focus will be on the efforts recently proposed in the literature that had the greatest impact on the community (as measured by, e.g., coverage of the problem domain, citations and received testing).

Relatively few of the existing works combine different modalities into a single system for human affective state analysis. Examples are the works of Chen et all (1998) and de Silva & Ng (2000) who studied the effects of a combined detection of facial- and vocal expressions of affective states. Other existing studies treat various human communicative signals separately.

In the last decade, tremendous progress has been made in the field of automating sensing, detection, tracking and interpretation of human hand and body gestures (from simple pointing through manipulative gestures to more complex symbolic gestures such as those in sign languages). Several exhaustive surveys on this topic have been published recently: hand gestures visual recognition/interpretation (Pavlovic et all 1997), human body modelling techniques (Cerezo et all 1999), and human body tracking (Pentland 2000). However, after a careful literature research, we did not find any report on a system that performs human affective state recognition based on an automatic analysis of the sensed body gestures.

Also, we found merely a single work aimed at automatic analysis of affective physiological signals, namely, the work presented in (Healey & Picard 1998) and in (Vyzas & Picard 1999). In this work automatic recognition of 8 user-defined affective states has been reported. Five physiological signals have been recorded: EMG from jaw (coding the muscular tension of the jaw), blood volume pressure

(BVP), hart rate calculated from the BVP, skin conductivity, and respiration. For emotional classification, an algorithm has been used that combines the Sequential Floating Forward Search and the Fisher Projection achieving an average correct recognition rate of 81.25%.

For these reasons, this survey is divided into merely two parts. The first is dedicated to the work done in automating facial affect analysis in digitised images or image sequences. The second explores and compares automatic systems for affective state recognition from audio input.

5.1 Automatic Facial Affect Analysis

Facial expressions are our primary means of communicating emotion. In addition, human face-to-face interaction is inherently natural and substantial evidence suggests this may also be true for human-computer interactions (Marsic et all 2000, Schiano et all 2000). These findings, together with advances in image analysis and pattern recognition, produced a surge of interest in automatic recognition of facial affect. For exhaustive surveys, readers are referred to: (Samal et all 1992) for a review of early works, (Donato et all 1999) for an overview of techniques for detecting micro facial actions (AUs), (Pantic & Rothkrantz 2000a) for a survey of current efforts.

The problem of affect-sensitive monitoring of facial expressions includes three sub-problem areas:

finding faces,

detecting facial features, and

classifying these data into some affect classes.

The problem of finding faces can be viewed as a segmentation problem (in machine vision) or as a detection problem (in pattern recognition). Possible strategies for face detection vary a lot, depending on the type of input images. The existing systems for facial expression analysis process either *facial* image sequences or static *facial* images. In other words, current studies assume, in general, that the presence of a face in the scene is ensured. Posed portraits of faces (uniform background and good illumination) constitute input data processed by the majority of the current systems. Yet, in many instances, the systems do not utilize a camera mounted on the subject's head as proposed in (Otsuka & Ohya 1998, Pantic & Rothkrantz 2000b/c) what will ascertain correctness of that assumption. Except of (Essa & Pentland 1997), (Hong et all 1998), and (Colmenarez et all 1999), presently existing systems do not perform automatic face detection in an arbitrary scene.

Facial feature extraction from input images may be divided into at least four dimensions:

are the features extracted in an automatic way,

is temporal information (image sequence) used,

are the features holistic (spanning the whole face) or analytic (spanning subparts of the face),

are the features view-based (2D) or volume-based (3D).

Given this glossary, most of the recently proposed approaches to facial affect analysis in facial images are directed towards automatic, static, analytic, 2D facial feature extraction. Still, many of the proposed systems do not perform facial information extraction in an automatic way (e.g. Chen et all 1998). Though the techniques for facial affect classification employed by these systems are relevant to the present goals, the systems themselves are of limited use for affect-sensitive monitoring where analyses of human communicative signals should be fully automatic and preferably achieved in real time. The approaches to automatic facial data extraction, utilised by the existing systems, include analyses of:

facial motion (e.g. Essa & Pentland 1997, Otsuka & Ohya 1998, de Silva & Ng 2000),

holistic spatial pattern (e.g. Hong et all 1998),

analytic spatial pattern (e.g. Colmenarez et all 1999, Pantic & Rothkrantz 2000b/c).

In many instances strong assumptions are made to make the problem of facial feature detection more tractable (e.g. images contain portraits of faces with no facial hair or glasses, the illumination is constant, the subjects are young and of the same ethnicity). Few of the existing systems deal with rigid head motions (e.g. Hong et all 1998, Colmenarez et all 1999) and only the method proposed by Essa & Pentland (1997) deals with the images of faces with facial hair and glasses.

Eventually, an automated facial affect analyser should classify the extracted facial features and provide a description of the displayed facial affect. However, exactly which affective/attitudinal states the system should recognise will depend on its application domain. If the intended application is, e.g., monitoring of a nuclear-power-plant operator, then the facial affect analyser to be deployed will be probably aimed at discerning stress and inattention. Except for the system of Pantic & Rothkrantz (2000c) that performs facial expression classification into user-defined interpretation classes, the existing facial affect analysers perform classification into a number of the six basic emotion categories as defined by Ekman (1994). Overall, the classification techniques used by the existing systems include:

template-based classification in static images (e.g. Hong et all 1998),

template-based classification in image sequences (e.g. Essa & Pentland 1997, Otsuka & Ohya 1998, Colmenarez et all 1999),

ANN-based classification in static images (e.g. Zhang et all 1998), rule-based classification in static images (e.g. Chen et all 1998, Pantic & Rothkrantz 2000b/c),

rule-based classification in image sequences (e.g. de Silva & Ng 2000).

Given that humans detect six basic emotional expressions with an accuracy ranging from 70% to 98% (Bassili 1979), it is rather significant that the automated systems achieve accuracy of 74% to 98% when detecting 3-7 emotions deliberately expressed by 8-40 subjects (Pantic & Rothkrantz 2000a).

5.2 Automatic Vocal Affect Analysis

In contrast to spoken language processing, which witnessed significant advances in the last decade (Juang & Furnai 2000), processing of “emotional” speech has not been widely explored by the auditory research community. However, recent data show that automated speech recognition, which works at about 80-90% accuracy on neutrally spoken speech, tends to drop to 50-66% accuracy on emotional speech (Steeneken & Hansen 1999). Although such findings triggered some efforts at automating vocal affect analysis, the focus of most researchers in this field has emphasized synthesis of emotional speech (Murray & Arnott 1996).

The problem of vocal affect analysis includes two sub-problem areas:

specifying auditory features to be estimated from the input audio signal, and

classifying those data into some affect classes.

The research in psychology/psycholinguistics provides an immense body of results on acoustic and prosodic features which encode the affective state of a speaker (e.g. Frick 1985, Schrer & Banse 1996). These studies point to the pitch as the main vocal cue for affective state recognition in speech. Most of the works on automating affect-sensitive analysis of vocal expressions, presented in the literature up to date, use this finding and estimate the pitch of the input audio signal. Other acoustic and prosodic features used in the existing works are:

intensity (i.e. vocal energy, power) (e.g. Tosa et all 1996, Chen et all 1998, Petrushin 1999),

slope (e.g. Li & Zhao 1998, Polzin 2000),

temporal features like speaking rate (e.g. Tosa et all 1996, Amir & Ron 1998, Petrushin 1999),

derivate features such as the smoothed pitch contour and its derivatives (Dellaert et all 1996),

phonetic features like the signal’s LPC-linear predictive coding parameters (Tosa et all 1996),

supra-segmental features such as the intensity and pitch over the duration of a syllable, word or sentence (Li & Zhao 1998, Polzin 2000).

Virtually all of the existing work on automating vocal affect analysis performs singular classification of input audio signals into few of the basic emotion categories. Utilised classification techniques include:

K-nearest neighbours (e.g. Dellaert et all 1996)

HMM (de Silva & Ng 2000, Polzin 2000)

Gaussian mix density models (Li & Zhao 1998)

Rule-based approach (Chen et all 1998)

Fuzzy membership indexing (Amir &Ron 1998)

ANN (e.g. Tosa et all 1996, Petrushin 1999)

In general, people can recognize emotion in a neutral-content speech with an accuracy of 60-70% when choosing from among six basic affective states (Bezooijen 1984). Automated vocal affect analysers match this accuracy when recognizing 4-8 emotions deliberately expressed by 2-100 subjects recorded while pronouncing sentences of 1-12 words length.

In many instances strong assumptions are made to make the problem of automating vocal expression analysis more tractable (e.g. the recordings are noise free; the recorded sentences are short, delimited by pauses, and carefully pronounced to express the required affective state; subjects are non-smoking professional or non-professional actors). Only one of the existing automated vocal affect analysers, i.e. (Petrushin 1999), has been tested on ‘almost’ real world data composed of short telephone massages spoken by 18 non-professional actors expressing mainly neutral and angry vocal affects (recognition rates reported are 73-77%). Overall, the testing data sets are small (5-50 sentences spoken by few subjects) containing exaggerated vocal expressions of affective states. Hence, the state of the art in automatic affective state recognition from speech is similar to that of speech recognition several decades ago when computers could classify the carefully articulated digits spoken with pauses in between, but could not accurately detect these digits if they were spoken in a way not previously encountered and forming a part of a longer continuous conversation.

6. KEY CHALLENGES

The limitations of the existing affect-sensitive monitoring tools are probably the best place to start a discussion of the challenges and opportunities that face researches of affective computing. The issue that strikes and surprises us most is that, though the recent advances in video and audio processing make automatic *multi-modal* affect-sensitive monitoring a remarkably tractable problem and though all agreed that solving this problem would be extremely beneficial, merely two efforts (i.e. Chen et all 1998 de Silva & Ng 2000) aimed at actual implementation of such a multi-modal tool have been presented in the literature up to date. Also, there is no record of a research endeavour towards inclusion of all non-verbal modalities into a single system for affect-sensitive monitoring of human behaviour. Next to the problem of achieving a deeper integration of the presently detached visual and auditory research communities, there are a number of related issues.

6.1 Visual Input

As already remarked in section 4, acquisition of video input for an affect-sensitive monitoring system concerns, at least, detection of monitored subject's face in the observed scene (if not of the upper part of body as well). The problematic issue here, typical of all visual processing, is that of occlusion, scale, and pose. Namely, in most real-life situations it cannot be assumed that the subject will remain immovable; rigid head motions can be expected causing changes in the viewing angle and in the visibility and illumination of the tracked facial features. Although highly time-consuming, the scale problem can be solved by forming a multi-resolution representation of the input image/frame and performing the same detection procedure at different resolutions. Pose and occlusion are more difficult problems, initially thought to be intractable or at least the hardest to solve. However, interesting progress is being made in machine vision research. The focus of active vision on *foveal purposeful vision* is the design and development of special sensors, which serve a specified purpose and are based on the principal of human-eye fovea in the sense that they can pan and zoom on relatively small regions of the scene that contain critical information. Further, statistical methods have been developed that essentially try to predict/guess the pose of monitored objects from whatever image information is available. Finally, methods for the monitored object's representations at several orientations, employing data acquired by multiple cameras, are currently thought to provide the most promising solution to the problems of pose and occlusion. For an extensive review of the methods for video-surveillance, the reader is referred to (Collins et all 2000).

Next to these standard problems of all visual processing, another issue typical for facial image processing concerns 'universality' of the employed technique for detection of the face and its features. Namely, the employed detection method must not be prone to the physiognomic variability and the current outlook of monitored subjects. As explained in section 4, an ideal automated affect-sensitive monitoring tool should perform generic analyses of the sensed facial information independently of possibly present static facial signals such as wrinkles and artificial facial signals like glasses and make-up. Essa & Pentland (1997) proposed such a method.

6.2 Audio Input

As already remarked in section 5.2, virtually all of the work done on automating vocal affect analysis assumes a fixed listening position, a closely placed microphone, non-smoking subjects, and noise-free recordings of short sentences that are delimited by pauses and carefully pronounced to express the required affective state. Hoping for such a clean audio input is not realistic, especially in the case of unconstrained environments characteristic for most applications in ubiquitous computing. One possible way of enhancing the state-of-the-art in vocal affect analysis is to explore existing methods for human language and speech processing and employ the most prominent pattern-recognition methods that minimize classification error rate. Excellent reviews of the existing methods for spoken language processing could be found in (Juang & Furnai 2000).

Another intriguing issue is the kind of features that should be adopted in order to achieve robust vocal affect recognition from speech. One standpoint is that the features should be solely prosodic and different from the phonetic features used for speech recognition. The other standpoint is that prosodic and phonetic features are tightly combined when uttering speech; it is impossible for us to express and recognize vocal affects by concerning prosodic features only. The later is experimentally proved – the observers who didn't speak Sinhala language performed correct recognition of six different emotions in Sinhala spoken speech merely with an average of 32.3% (de Silva et all 1998). Another interesting observation is that the information encoded in the speech signal becomes far more meaningful if peach and intensity could be observed over the duration of a syllable, word, or phrase (Polzin 2000). For researchers of automatic vocal affect analysis this suggests investigating towards robust, speaker-independent, temporal analysis of phonetic and prosodic characteristics of speech.

6.3 Multi-modal Input

As far as an automatic multi-modal monitoring of human affective states is concerned, the goal is to achieve generic, time-instance/time-scale analyses of audio, visual, and tactile human communicative signals. An ideal human affect analyser (Fig. 1, §4) should generate a reliable result based on multiple input signals acquired by different sensors. Let us explain this issue in more detail.

Considering the state-of-the-art in audio, visual, and tactile processing, inaccurate, noisy and missing data should be expected. An (ideal) affect-sensitive monitoring tool should be able to deal with these imperfect data and to generate its conclusion so that the certainty, associated with it, varies in accordance to the certainty of the input data. A way of achieving this is to consider time-instance vs. time-scale dimension of human paralanguage. Namely, there is a certain grammar of neuromuscular actions and physiological reactions: only a certain subclass of these actions/reactions with respect to the currently encountered action/reaction (time-instance) and the previously observed actions/reactions (time-scale) is plausible. If the current input data affirm these statistically predicted actions/reactions, the certainty associated with that data should be 'high' and the certainty of the drawn conclusion is to be computed accordingly. Nevertheless, such a *temporal analysis* involves untangling the grammar of human behaviour, which is a rather unexplored topic even in the psychological and sociological research areas. The issues, which make this problem even more difficult to solve in a general case, concern the dependency of human behaviour upon the monitored person's personality, cultural and social vicinity, current mood, and the context (situation) in which the observed behavioural cues occur. One source of help for these problems is machine learning – rather than having a priori rules of human behaviour, we can potentially learn application-, user-, and context-dependent rules by watching the user's behaviour in the sensed context. Though context-sensing and the time needed to learn appropriate rules are significant problems (Pentland 2000), usefulness and universal accessibility of such an adaptive affect-sensitive HCI tool could dwarf previous generations of HCI systems.

Another issue that is typical of all multi-modal processing is that of processing multi-sensory data separately, combining them only at the end (Sharma et al 1998). The system proposed by de Silva & Ng (2000) is an example. Yet, this is almost certainly incorrect; people display audio, visual, and tactile communicative signals in a complementary and redundant manner. Chen et al. (1998) have proved this experimentally for the case of audio and visual input. In order to accomplish a multi-modal analysis of multiple signals acquired by different sensors, which will resemble human recognition of affective states, the input signals cannot be considered mutually independent and cannot be combined at the end of the intended analysis. In turn, the input data should be processed in the joint feature space. In practice, yet, there are two major difficulties:

a huge joint feature space resulting in a heavy computational burden, and

different feature formats and timing.

A way of dealing with these problems and achieving tightly-coupled multi-sensory data fusion is to apply a Bayesian inference method as presented in (Pan et all 1999). However, due to the complexity of the phenomena and a general luck of researchers having expertise in all domains (audio, visual, and tactile processing), untangling the problem of joint audio-visual-tactile human affect analysis is still a significant challenge facing the researchers of multi-modal human affect analysis.

6.4 Interpreting Multi-modal Input

Currently existing methods aimed at automating human affect analysis are not context-sensitive. Yet, interpreting human communicative signals is strongly situation-dependent (Russell 1994). Initially thought to be the research topic that would be hardest to solve, context-sensing in terms of who is the user, where is he, and what is he doing, has been proven remarkable tractable. For a discussion on advances and challenges in this research topic, readers are referred to (Pentland 2000). Yet, due to the complexity of this wide-ranging problem and a general luck of researchers having the full extent of necessary expertise, the problem of context-sensitive human affect analysis poses, perhaps, the most significant research challenge.

Another issue concerns the actual interpretation of human communicative signals in terms of affect/ attitudinal states. Almost all of the existing work employs singular classification of input data into one of the six basic emotion categories (section 5). This approach has many limitations. As explained in section 2, the theory on existence of six universal emotion categories is nowadays strongly challenged in the psychological research area. Further, as noted by the inventor of this theory himself, pure expressions of basic emotions are seldom elicited. Most of the time people show blends of emotional displays. Hence, classifying human communicative signals into a single basic-emotion category isn't realistic. An affect-sensitive analyser of sensed human communicative signals must at least realise a quantified classification into multiple emotion categories, e.g., as proposed in (Pantic & Rothkrantz 2000b) and (Zhang et all 1998) for the case of automatic facial affect analysis. Yet, not all human communicative displays can be classified as a combination of the six basic emotion categories. Think for instance about contempt, stress, boredom, or 'I don't know' attitudinal states. Besides, it has been shown that the comprehension of a given emotion label and the ways of expressing the related affective state differ from culture to culture (section2). Hence, defining interpretation categories into which any set of human communicative signals can be classified is one of the key challenges in design of a realistic affect-sensitive monitoring tool.

The lack of psychological scrutiny on the topic makes this problem even harder. One source of help for this problem is machine learning: instead of building rigid generic rules into the intended tool, the system can potentially learn its own expertise by allowing the user to define his own interpretation categories, e.g., as proposed in Pantic & Rothkrantz (2000c) for an automated facial affect analyser. As already remarked in section 6.3, an adaptive (user-, application-, and context-profiled) affect-sensitive monitoring tool would represent an ideal automated tool for understanding of human behaviour that could greatly enhance the state-of the-art in HCI.

7. CONCLUSION

Automating user-profiled, multi-modal, context- and affect-sensitive monitoring and interpretation of human behavioural cues is likely to be the single most widespread research topic of the AI research community in general (Pentland 2000). The catalyst behind is that untangling the problems related to this research topic is prerequisite for the design of next generation perceptual interfaces and ubiquitous computing in general.

However, currently existing methods aimed at automating human affect analysis are:

uni-modal, except of the systems proposed by Chen et all (1998) and de Silva & Ng (2000) that perform a joint audio-visual affect analysis,

context-insensitive, and

user-inadaptable, except of the automated facial analyser of Pantic & Rothkrantz (2000c), which performs facial data interpretation in terms of affect-descriptive labels learned from the user.

In summary, though the fields of machine vision, audio processing, and affective computing generally, witnessed rather significant advances in the past few years, realisation of robust, fully automated, multi-modal, adaptive, affect-sensitive analyser of human communicative cues is still in a rather distant future.

Another problematic issue, which jeopardises a future wide deployment of adaptive affect-sensitive monitoring tools proposed in this paper, concerns the efficiency of such HCI tools. Namely, since embedded computing devices are generally thought to be everywhere in the future, having the user train each of those devices will be inefficient. The computers of our future must know enough about the people and the environment in which they act to be capable of acting appropriately with a minimum of explicit instruction (Pentland 2000). A long-term way of achieving this is:

to develop multi-modal affect-sensitive tools, as proposed in this paper, which will be capable of monitoring human behaviour and adapting to the current user (who is he, what is the grammar of his behavioural actions/reactions), his context (where is he, what is he doing at the point), and the application domain (e.g. observing stress by a nuclear power plant operator while he reads his e-mail is not the reason for an alarm), then

to make those self-adaptive tools commercially available to the users that will profile them in the context in which the tools are to be used, and finally

to withdraw the trained systems after some time and combine the stored knowledge in order to derive generic statistical rules/models of human behaviour in the given context/environment.

Though willingness of people to participate in such a privacy-intruding large-scale project is a significant problem in its own right, this approach could resolve many intriguing questions. The most important is that this could resolve the social impact of interaction in electronic media, i.e., the effects of information technology on: interpersonal interaction, overall related human behaviour, and our cultural and social vicinity.

While all agreed that giving the machines the ability to interpret human behaviour without explicit instruction would be enormously beneficial, would represent the coming of universally usable and accessible HCI systems, and would probably define the impact information technology has on our social behaviour, we also should recognise the likelihood that such a goal is still in the relatively distant future.

REFERENCES

Thalmann, N.M., et all, 1998, ‘Face to virtual face’, *Proc. IEEE* 86(5): 870-883.

Sharma, R., et all, 1998, ‘Toward multimodal human computer interface”, *Proc. IEEE* 86(5): 853-869.

Marsic, I., et all, 2000, ‘Natural communication with information systems’, *Proc. IEEE* 88(8): 1354-1366.

Pentland, A., 2000, ‘Looking at people: Sensing for ubiquitous and wearable computing’, *IEEE Trans. PAMI* 22(1): 107-119.

Picard, R.W., 1997. *Affective computing*, Cambridge: MIT Press.

Bassili, J.N., 1979, ‘Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face’, *Journal of Personality and Social Psychology* 37: 2049-2058.

Ekman, P., et all, 1969, ‘The repertoire of nonverbal behavioral categories – origins, usage, and coding’, *Semiotica* 1: 49-98.

Steimer-Krause, E., et all, 1990, ‘Interaction regulations used by schizophrenics and psychosomatic patients’, *Psychiatry* 53: 209-228.

McHugo, G.J., et all, 1985, ‘Emotional reactions to a political leader’s expressive displays’, *Journal of Person. and Soc. Psychology* 49: 1513-1529.

Bezooijen, R.V., 1984. *Characteristics and recognizability of vocal expression of emotions*, Dordrecht, NL: Floris.

Ekman, P., 1994, ‘Strong evidence for universals in facial expressions: A reply to Russell’s mistaken critique’, *Psychological Bulletin* 115(2): 268-287.

Furnas, G., et all, 1987, ‘The vocabulary problem in human-system communication’, *Commun. ACM* 30(11): 964-972.

Russell, J.A., 1994, ‘Is there universal recognition of emotion from facial expression?’, *Psychological Bulletin* 115(1): 102-141.

Matsumoto, D., 1990, ‘Cultural similarities/differences in display rules’, *Motivation & Emotion*, 14: 195-214.

Cacioppo, J.T., et all, 2000, ‘The psychophysiology of emotion’. *Handbook of Emotions*, pp. 173-191, New York: Guilford Press.

Mehrabian, A., 1968 ‘Communication without words’, *Psychology Today* 2(4): 53-56.

Chen, L.S., et all, 1998, ‘Multimodal human emotion/ expression recognition’, *Proc. FG’98*, pp. 396-401.

de Silva. L.C., Ng, P.C., 2000, ‘Bimodal Emotion Recognition’, *Proc. FG 2000*, pp. 332-335.

Pavlovic, V.I., et all, 1997, ‘Visual interpretation of hand gestures for human-computer interaction: A review’, *IEEE Trans. PAMI* 19(7): 677-695.

Cerezo, E., et al, 1999, ‘Motion and behavior modeling: State of art’, *The Visual Computer*, 15: 124-146.

Healey, J., Picard, R., 1998, ‘Digital processing of affective signals’, *Proc. ICASSP’98*, pp. 3749-3752.

Vyzas, E., Picard, R., 1999, ‘Offline and online recognition of emotion expression from physiological data’, *Proc. Autonomous Agents ’99 – Emotion-Based Agent Architectures Workshop*, pp. 135-142.

Schiano, D.J., et all, 2000, ‘Face to interface: Facial affect in human & machine’, *Proc. CHI 2000*, pp. 193-200.

Samal, A., et all, 1992, ‘Automatic recognition and analysis of human faces and facial expressions: survey’, *Pattern Recognition* 25(1): 65-77.

Donato, G., et all, 1999, ‘Classifying Facial Actions’, *IEEE Trans. PAMI* 21(10): 974-989.

Pantic, M., Rothkrantz, L.J.M., 2000a, ‘Automatic analysis of facial expression: The state of the art’, *IEEE Trans. PAMI* 22(12): 1424-1445.

Otsuka, T., Ohya, J., 1998, ‘Spotting segments displaying facial expression from image sequences using HMM’, *Proc. FG’98*, pp. 442-447.

Pantic, M., Rothkrantz, L.J.M., 2000b, ‘Expert system for automatic analysis of facial expression’, *Image and Vision Computing* 18(11): 881-905.

Pantic, M., Rothkrantz, L.J.M., 2000c, ‘Self-adaptive expert system for facial expression analysis’, *Proc. IEEE SMC 2000*, pp. 73-79.

Essa, I., Pentland, A., 1997, ‘Coding analysis interpretation recognition of facial expressions’, *IEEE Trans. PAMI* 19(7): 757-763.

Hong, H., et all, 1998, ‘Online fac. exp. recognition based on personalised galleries’, *Proc. FG ‘98*, pp. 354-359.

Colmenarez, A., et all, 1999, ‘Probabilistic framework for embedded face and facial expression recognition’, *Proc. CVPR'99*, pp. 592-597.

Zhang, Z., et all, 1998, ‘Comparison between geometry-based and Gabor wavelets-based facial expression recognition…’, *Proc. FG’98*, pp. 454-459.

Juang, B.H., Furnai, S., Eds., 2000, *Issue on Spoken Language Processing*, *Proc. IEEE* 88(8): 1139-1366.

Steeneken, H.J.M, Hansen, J.H.L., 1999, ‘Speech under stress conditions: Overview of the effect on speech…’, *Proc. ICASSP’99-4*, pp. 2079-2082.

Murray, I.R., Arnott, J.L., 1996, ‘Synthesizing emotion in speech: Is it time to get excited?’, *Proc. ICLSP’96*, pp. 1816-1819.

Frick, R., 1985, ‘Communicating emotion. The role of prosodic features’, *Psych. Bulletin* 97(3): 412-429.

Scherer, K.R., Banse, R., 1996, ‘Acoustic profiles in vocal emotion expression’, *Personality & Social Psychology* 70: 614-636.

Tosa, N., et all, 1996, ‘Life-like communication agent – emotion sensing character MIC and feeling session character MUSE’, *Proc. MULTIMEDIA’96*, pp. 12-19.

Petrushin, V.A., 1999, ‘Emotion in speech: Recognition and application to call centers’, *Proc. ANNIE’99*.

Li, Y., Zhao, Y., 1998, ‘Recognizing emotions in speech using short-term and long-term features’, *Proc. ICSLP’98-6*, pp. 2255-2258.

Polzin, T.S., 2000, *Detecting verbal and non-verbal cues in the communications of emotions*. PhD thesis, Carnegie Mellon University.

Amir, N., Ron, S., 1998, ‘Towards automatic classification emotions in speech”, *Proc. ICSLP’98-3*, pp. 555-558.

Dellaert, F., et all, 1996, ‘Recognizing emotion in speech’, *Proc. ICLSP’96*, pp. 1970-1973.

Collins, R.T., et all, Eds., 2000, *Special Section on Video Surveillance, IEEE Trans. PAMI* 22(8): 745-887.

De Silva, L.C., et all, 1998, ‘Use of multimodal info in facial emotion recognition”, *IEICE Trans. Inf. & Syst.* E81-D(1): 105-114.

Pan, H., et all, 1999, ‘Exploiting the dependencies in information fusion’, *Proc. CVPR’99*-2, pp. 407-412.