The PAMSAE system : Parallel Multimedia Stress Assessment

R.J. van Vark, L.J.M. Rothkrantz & E.J.H. Kerckhoffs
Delft University of Technology
Faculty of Technical Mathematics and Informatics
Section of Knowledge Based Systems
Julianalaan 132
2628 BL Delft
The Netherlands
E-mail: {robert, roth, eugene}@kgs.twi.tudelft.nl

Topics:
analysis and interpretation, high performance computing, image processing, image coding, multimedia, and parallel systems

Motivation:
The PAMSAE-project consists of several aspects which are interesting to the ASCI-community, besides being of interest to nowadays research society. First of all, the PAMSAE model consists of four parallel subsystems. Two of these systems incorporate vision techniques to analyse incoming data. First of these subsystems derives positional information of several body segments, like head, shoulders, hands and arms. By comparing successive positions during sequences, motion is also detected. Starting with the derived position of the head, the second subsystem derives information about facial expressions. This is done by many different techniques, including neural networks, deformable templates and snakes. At the moment, a lot of attention is paid to both topics by researchers world-wide, as can be seen from the emergence of many conferences covering these topics.
The second aspect interesting to the ASCI-community is the application of high performance computing. The final system will be a real-time system which processes four media simultaneously. The system will therefore have to be embedded in a high performance parallel system, especially as the vision prototypes are also very computation intensive.

Abstract:
A research group of Delft University of Technology, Faculty of Technical Mathematics and Informatics, has been working on automated stress assessment since 1992. Goal of the project is the development of a Parallel Multimedia Stress Assessment system (PAMSAE). The system will use multimedia techniques and is based on the analysis of human motion, facial expressions, voice signals and physiological sensor data. These media are analysed in parallel and an overall stress score is obtained based on the outcome of these parallel analyses. Analysis techniques are mostly based on artificial intelligence techniques, like knowledge based systems and artificial neural networks. Several prototypes have been developed for each medium. These prototypes have been tested on a group of 108 computer science students.
In the paper, the “term” stress will be discussed first. Secondly, the necessity of using multimedia in stress assessment is discussed. Thirdly, the PAMSAE-model will be presented, followed by some possible fields of application. Next, the developed prototypes will be discussed, followed by test results obtained by using several of these prototypes. Finally, some conclusions will be presented and future developments concerning PAMSAE will be covered.

Stress
Stress involves the internal and external stimulation of an organism which upsets its internal balance. It involves a series of physiological and behavioural changes that prepare the organism for the appropriate coping response. These changes are characterised by some degree of activation (arousal or emotional

tension in the organism) which insures the body to be ready for action. The physiological stress response of an organism is dominated by sympathetic arousal characterised by changes in heart rate, blood pressure, respiration patterns, and muscle tension. Of these, respiration and muscle tension are likely to have effect on speech production.
From both real-life stress studies and laboratory induced stress experiments, it is found that a number of speech parameters might be viewed as stress indicators: the fundamental frequency, the first and second formant, intensity, speech rate, and zero crossings. In non-verbal voice analysis for stress assessment, it is necessary to have proper software tools and hardware equipment for the extraction and analysis of relevant sound and speech parameters.
From the theory of P. Watzlawick, it is known that human communication has not only a verbal but also a non-verbal aspect. Full understanding of messages in human interaction is only partly based on the contents of the messages. The way of speaking, the corresponding facial expressions and the (involuntary) movements of head and hands greatly influence the understanding of the message. The registration and analysis of facial expressions is based on the theory of P. Ekman. He developed a classification model for emotions, based on special "action units" in the face. The interesting parts in the face are regions of the mouth, eyes and eyebrows.
Involuntary movement of body segments like head and hands become more frequent when a person is stressed. One thinks of finger tapping on a table. The mentioned body segments can be recorded by object tracking techniques. Using these techniques a measure for the amount of (involuntary) movement can be deducted.

Multimedia
Due to recent developments in multimedia, the application of computer sound and vision in the work of everyday is coming closer and closer. This increases the applicability of systems using multimedial data streams enormously as multimedia systems become cheaper and it becomes possible to obtain test results for hundreds of people without the problems which occurred a decade ago.
The main reason to use data from three different media in the PAMSAE-project, is to reduce the ambiguity in the different data streams. Using one of the data streams, it is not possible to obtain a one-to-one relationship between overt behaviour and stress. For instance, increased heart rate can be a sign of increased stress levels, but the person in question can also have just run up five stairs. Using all four data streams ambiguity can be reduced to a minimum, but it is however still possible to “fool” the system as it is possible to fool a human observer. The behaviour feedback system has not the pretension to surpass the human observer, but is mainly designed to be used in absence of a human observer.

Model
The PAMSAE-model consists of four parallel subsystems: physiology, speech, facial expressions and movement. In every data stream, the analysis of multimedial data takes place in several steps. These steps are the same for each medium and are implemented as several processing layers. The first layer is the *filtering & reduction layer*. In this layer several transformations are applied to the raw incoming measurements: noise filtering, outlier correction to eliminate recording errors and data reduction by feature abstraction. The second layer is the *preprocessor layer*. In this layer, measurements are stored for later use. Additionally, simple transformations are applied, like averaging and smoothing. The third layer, the *processor layer*, is very important as it derives hypotheses based on preprocessed features using an intermediate representation. The fourth layer is the *application layer*. This layer consists of several applications of which any subset can be processed simultaneously. Every applications integrates intermediate results from each of the parallel streams. The final layer is the *output layer*, which produces two types of feedback. Temporary feedback is used to produce feedback during the session either for

the trainer or trainee. Permanent feedback interprets complete sessions and can be used to check progress during successive sessions.

Fields of application

The system can be used in a variety of applications. Some of these possible fields of application will be discussed.

Training in stress management. Pilots and train drivers have a stressful job. Analysis of accidents and near accidents proves human factors and stress to play an important role during the onset of these accidents. Some recently developed training programs in flight simulators include stress modules to train people to manage and control the occurrence of stress. For training without supervision or as an aid to the trainer, an automated system for stress assessment and behaviour feedback is indispensable.

Teleconferencing. In traditional face-to-face meetings, all participants watch the speaker and listeners simultaneously. Sudden movements of heads, hands and feet draw their attention, so that they can immediately focus their attention on those participants. During teleconferencing, these non-verbal clues cannot be observed and an automated system can be valuable.

Identification verification. Face recognition and voice recognition is used in ID verification. In contradiction with current systems, which ignore emotions as being noise, emotions can be a characteristic personal trait. Extracting features of emotion from sensor, sound and image data, can therefore improve the results of person recognition and verification, especially when traditional recognition or verification techniques fail due to special masking techniques.

Other applications include human friendly user interfaces, realistic feedback during virtual reality sessions, lie detection, monitoring of intensive care patients and new-borns who have been hospitalised and psychotherapy.

Prototypes

The input of the physiological prototype consists of a number of physiological signals: electrocardiogram, blood pressure, respiration and galvanic skin response. In the filtering & reduction layer the following features are abstracted: heart rate, heart period, blood pressure, respiration rate, depth of respiration and galvanic skin response. The preprocessor layer averages these signals over a short period of time as temporary, cyclic perturbations can be found in physiological signals. The processor layer derives the stress hypotheses using nervous activity as the intermediate representation. Several types of output are built in the prototype. First, physiological behaviour is visualised as two-dimensional paths, in contrast to traditional one-dimensional projections. Points in this 2D-space can be labelled by different types of nervous activity and stress levels. Another type of feedback is the behavioural compass. This indicates the direction in which behaviour is changing in the near future. Based on this prediction, evasive action can be taken if necessary. Finally, a stress meter is integrated in the prototype. Several areas in the two-dimensional space correspond to certain stress levels and these are visualised as a kind of “econometer”.

The sound prototype samples the human voice signal at a rate of 44 kHz and a resolution of 16 bits. From this sampled signal a number of features are extracted: frequency spectrum, frequency spectrogram, fundamental frequency, jitter, zero crossings, energy and average amplitude. Noise filtering is very important as background noise can completely disguise the original voice signal. The first layer also detects if the observant was speaking or not speaking at the time of recording. The abstracted features are averaged over short periods of time. Hypotheses are derived using emotions as intermediate representation. Besides visualisation, only one type of feedback is implemented: a voice stressometer. Using a formula based on a linear combination of vocal features, stress levels can be deducted.

Two prototypes have been developed for facial expressions. The first prototype uses manual extracted features which can be interpreted by neural networks. The neural networks classify these features as different types of emotion. Stress corresponds to certain levels of emotions. Classification of a well known standard set of facial expressions resulted in a correct classification in over 75 percent of the cases. The second prototype concerns the automatic generation of facial expressions. By examining these expressions, more insight is gained in the correspondence between facial expressions and stress.

Test results

During the summer of 1994, some experiments were held to test the physiological and voice prototypes. These experiments were carried out in a laboratory using one computer for stimulus and two computers, one for each medium, for recording and feedback. Testing was done by 108 computer science students as part of their curriculum. Voice signals were only recorded during some parts of each session, while physiological signals were recorded during entire sessions. The tests consisted of some behavioural tasks, which were recorded after some basal recording of rest levels. These are important to compensate for personal differences. Behavioural tasks consisted of a computer game at two different playing levels, another computer game during which the observant had to perform mental arithmetic and finally the well-known Stroop Test. This task is known to be extremely stressful. Each task lasted for five minutes, followed by two minutes rest.

In general, physiological signals indicated higher stress levels during the fourth task than during the first two tasks. For the third task, two groups can be distinguished: one group of people who enjoy racing games and one group of people who dislike them or are not good at them. The second group did not react significantly, while the first group showed strong reactions. Voice analysis subscribes to the stressful nature of the Stroop Test as higher stress levels were also found using the voice stressometer.

The two-dimensional projection of the physiological prototype was found to be clearer and easier to interpret than the classical one-dimensional presentation. It was easy to classify the observed behaviour and therefore stress without observing the students. It was also possible to predict near future behaviour using the behavioural compass. The voice stressometer performed reasonably well, although the correspondence between vocal features and stress was found not to be completely linear, which presumption was used to implement the voice stressometer.

Conclusions

Although not all prototypes have been tested in detail, they can serve as a solid base for further developments towards an integrated multimedial stress assessment system. As developments in multimedia are evolving faster and faster, the diversity in possible applications of the automated stress assessment system will increase even more in nearby future.

Development of an integrated multimedial stress assessment system will still take a number of years. First, the developed prototypes have to be improved. Especially, vision prototypes have not evolved into real-time assessment tools. As can be seen from the huge list of recent publications, the research society puts a lot of effort in motion detection, feature abstraction and facial recognition. Secondly, a parallel system has to be developed in which the developed prototypes can be embedded. This parallel system has to be real-time and has to be able to interpret and integrate the outcome of each of the media intelligently.