Emotion Detection in the Loop from Brain Signals and Facial Images

Arman Savran[1], Koray Ciftci[1], Guillame Chanel[2], Javier Cruz Mota[3], Luong Hong Viet[4], Bülent Sankur[1], Lale Akarun[1], Alice Caplier[5] and Michele Rombaut[5]

[1]*Bogazici University,* [2]*University of Geneva,* [3]*Universitat Politecnica de Catalunya,* [4]*The Francophone Institute for Computer Science,* [5]*Institut National Polytechnique de Grenoble*

arman.savran@boun.edu.tr, rciftci@boun.edu.tr, javicm@gmail.com, lhviet@ifi.edu.vn, guillaume.chanel@cui.unige.ch, bulent.sankur@boun.edu.tr, caplier@lis.inpg.fr, akarun@boun.edu.tr

***Abstract*—**
In this project, we intended to develop techniques for multimodal emotion detection, one modality being brain signals via fNIRS, the second modality being face video and the third modality being the scalp EEG signals. EEG and fNIRS provided us with an "internal" look at the emotion generation processes, while video sequence gave us an "external" look on the "same" phenomenon.

Fusions of fNIRS with video and of EEG with fNIRS were considered. Fusion of all three modalities was not considered due to the extensive noise on the EEG signals caused by facial muscle movements, which are required for emotion detection from video sequences.

Besides the techniques mentioned above, peripheral signals, namely, respiration, cardiac rate, and galvanic skin resistance were also measured from the subjects during "fNIRS + EEG" recordings. These signals provided us with extra information about the emotional state of the subjects.

The critical point in the success of this project was to be able to build a "good" database. Good data acquisition means synchronous data and requires the definition of some specific experimental protocols for emotions elicitation. Thus, we devoted much of our time to data acquisition throughout the workshop, which resulted in a large enough database for making the first analyses. Results presented in this report should be considered as preliminary. However, they are promising enough to extend the scope of the research.

***Index Terms*—**Emotion detection, EEG, video, near-infrared spectroscopy

I. INTRODUCTION

Detection and tracking of human emotions have many potential applications ranging from involvement and attentiveness measures in multimedia products to emotion-sensitive interactive games, from enhanced multimedia interfaces with more human-like interactions to affective computing, from emotion-sensitive automatic tutoring systems to the investigation of cognitive processes, monitoring of attention and of mental fatigue.

The majority of existing emotion understanding techniques is based on a single modality such as PET, fMRI, EEG or static face image or videos. The main goal of this project was to develop a multimodal emotion-understanding scheme using hemodynamic brain signals, electrical brain signals and face images. Studies about the way to fusion the different modalities was also an important goal of the work.

Psychologists agree that human emotions can be categorized into a small number of cases. For example, Ekman et al. [1] found that six different facial expressions (fearful, angry, sad, disgust, happy, and surprise) were categorically recognized by humans from distinct cultures using a standardized stimulus set. In other words, these facial expressions were stable over races, social strata and age brackets, and were consistent even in people blind by birth.

Nevertheless, there are several difficulties in automatic human emotion identification. First, the straightforward correlation of emotions with neural signals or with facial actions may not be correct since emotions are affected by interactions with the environment. As a result, the unfolding of emotions contains substantial inter-subject and intra-subject differences, even though the individuals admit or seem to be in the claimed emotional

This report, as well as the source code for the software developed during the project, is available online from the eNTERFACE'06 web site: www.enterface.net.

situation. Moreover, to design experiments to single out a unique emotion is a very challenging task. These imply that, even small changes in the experimental setup may lead to non-negligible differences in the results.

The majority of existing emotion understanding techniques is based on a single modality such as PET, fMRI, EEG or static face image or videos. The main goal of this project was to develop a multimodal emotion-understanding scheme using functional, physiological and visible data. As an intermediate step, it was necessary to determine the feasibility of fusing different modalities for emotion recognition. These modalities are functional Near Infrared Spectroscopy (fNIRS) electroencephalogram (EEG), video and peripheral signals. Note that these modalities provide us with different aspects of the "same" phenomenon. fNIRS and EEG try to detect functional hemodynamic and electrical changes, peripheral signals give an indication of emotion-related changes in the human body and video signal captures the "visible" changes caused by emotion elicitation.

In the rapidly evolving brain-computer interface area, fNIRS (functional Near Infrared Spectroscopy) represents a low-cost, user-friendly, practical device for monitoring the cognitive and emotional states of the brain, especially from the prefrontal cortex area. fNIRS detects the light (photon count) that travels through the cortex tissues and is used to monitor the hemodynamic changes during cognitive and/or emotional activity.

The second modality to estimate cortical activity is EEG. Using the scalp electrodes, useful information about the emotional state may be obtained as long as stable EEG patterns on the scalp are produced. EEG recordings capture neural electrical activity on a millisecond scale from the entire cortical surface while fNIRS records hemodynamic reactions to neural signals on a seconds scale from the frontal lobe. In fact, electrical activity takes place in order of milliseconds, whereas hemodynamic activity may reach its peak in 6-10 seconds and may last for 30 seconds. In addition to these modalities, peripheral signals, namely, galvanic skin response (GSR), respiration and blood volume pressure (from which we can compute heart rate) were also recorded.

We have combined these four monitoring modes of emotions in two separate pairs, namely: i) fNIRS, ii) EEG, iii) peripheral signals, iv) image or video. Notice that EEG is very sensitive to electrical signals emanating from facial muscles while emotions are being expressed, hence EEG and video modalities cannot coexist. In contrast, fNIRS is the modality that can be combined with either video signals or with EEG signals.

In summary, the first short-term goal of the project has been to build a reliable database that can be used for all related future research. The second such goal was to prove the viability of a multi-modal approach to emotion recognition, both from instrumentation and signal processing points of view. The final long-term aim is to build an integrated framework for multi-modal emotion recognition for both brain research and affective-computing aspects.

II. MEASUREMENT SETUP AND EMOTION ELICITING

A. *Instrumental Setup*

To detect and estimate emotions based on brain as well as physiological signals the following sensor setup was prepared: (Figure 1):

- fNIRS sensor to record frontal brain activity,
- EEG sensor to capture activity in the rest of the brain,
- Sensors for acquiring peripheral body processes: a respiration belt, a GSR (Galvanic Skin Response) and a plethysmograph (blood volume pressure)

All these devices were synchronized using a trigger mechanism. Notice that EEG and fNIRS sensor arrangements partially overlap, so that there is no EEG recording on the front. Similarly the fNIRS device covers the eyebrows, occluding one of the image features for emotion recognition

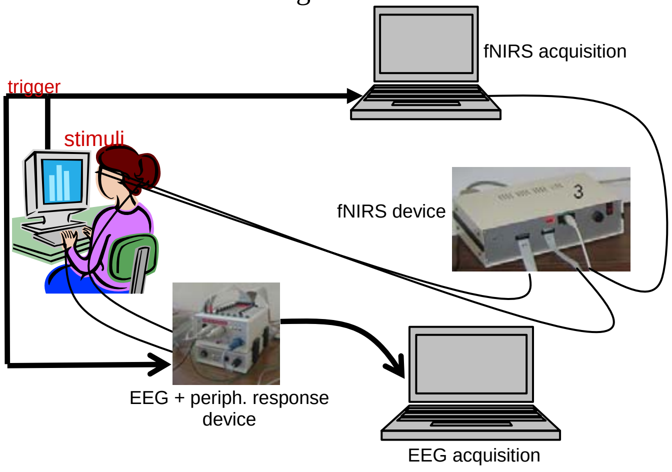

Figure 1 Schematics of EEG and fNIRS acquisition.

The Video-fNIRS acquisition scenario is composed of three computers, *Stimulus Computer*, *fNIRS Computer* and *Video Computer,* each one with the following purpose (Figure 2):

- **Stimulus Computer** shows recorded stimuli to the subjects, sends synchronization signal via the parallel port to the *fNIRS Computer* and stores stimuli start and end instants in a log file.
- **fNIRS Computer** acquires fNIRS data from the fNIRS device.
- **Video Computer** acquires video data from a Sony DFW-VL500 camera.

Synchronization becomes a critical issue when more than one modality is to be recorded, especially when they are recorded on different computers. We have used two synchronization mechanisms: In the first mechanism, the *Stimulus Computer* sends a signal to the *fNIRS Computer* each time a stimulus is shown in the screen via the parallel port. In the second mechanism, the *Stimulus Computer* writes to a log file the instants, with millisecond precision, of each stimulus. This log file is used after recording in the *Video Computer* to mark the frames corresponding to each stimulus. Before the recording process, the *Stimulus Computer* and the *Video Computer* clocks are synchronized using a free internet NTP server localized in Zagreb (ri.ntp.carnet.hr).

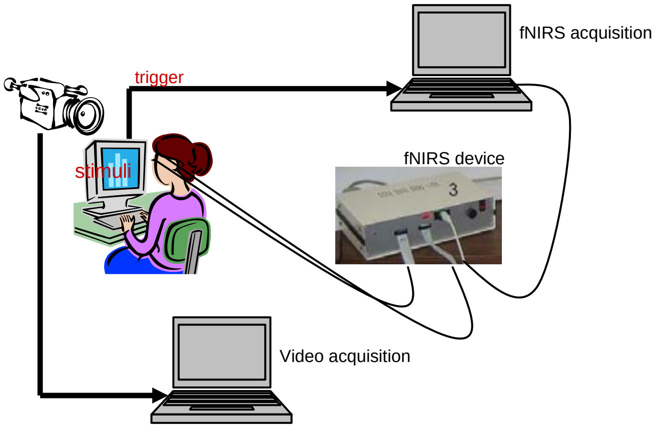

Figure 2 Schematics of Video and fNIRS acquisition.

B. Emotion Eliciting Images

The emotions were elicited in subjects using images from the IAPS (International Affective Picture System) 9. Several studies have shown the usefulness of images to elicit emotional responses that trigger discriminative patterns in both the central and peripheral nervous system (10, 11). The IAPS contains 900 emotionally evocative images evaluated by several American participants on two dimensions of nine points each (1-9): valence (ranging from positive to negative or unpleasant to pleasant) and arousal (ranging from calm to exciting). The mean and variance of participant judgments for both arousal and valence are computed from these evaluation scores.

We chose images from the IAPS that corresponded to the three emotional classes we wanted to monitor: calm, exciting positive and exciting negative. This was performed by first selecting pictures from IAPS values (1) and then eliminating particular images based on redundancy or particularity of context (for example erotic images were removed). This selection resulted in 106, 71, and 150 pictures respectively for these classes. The selection of the three images subsets, corresponding to the emotional states of interest was instrumented via empirical thresholds on valence and arousal scores:

$$\begin{aligned} &calm: && \overline{arousal} < 4; \quad 4 < \overline{valence} < 6 \\ &positive\ exciting: && \overline{valence} > 6.8; \\ & && Var(valence) < 2; \\ & && \overline{arousal} > 5 \\ &negative\ exciting: && \overline{valence} < 3; \quad \overline{arousal} > 5 \end{aligned} \tag{1}$$

C. Experimental Protocol for fNIRS, EEG and Peripheral Signals

The stimuli to elicit the three target emotions were the above selected images from the IAPS. During the experiment, the subject is seated in front of the computer screen his/her physiological responses (i.e.: fNIRS, EEG and peripheral activity) are being measured. The stimuli are brought to the screen in random order. The subject is asked to watch the images and be aware of his emotional state. In this study, we recorded data from five subjects using the Biosemi Active 2 acquisition system with 64 EEG channel and the peripheral sensors. Due to occlusion from fNIRS sensor arrangement, we had to remove the following ten frontal electrodes: F5, F8, AF7, AF8, AFz, Fp1, Fp2, Fpz, F7, F6, which left us with 54 channels. All EEG signals were recorded at 1024 Hz sampling rate except the first session of participant 1 that was recorded at 256 Hz.

The protocol is detailed in Figure 3 each stimulus consists of a block of five pictures from the same class, this to insure stability of the emotion over time. Each picture is displayed on the screen for 2.5 seconds leading to a total of 12.5 seconds per block. Blocks of different classes are displayed in random order to avoid participant habituation. A dark screen precedes each block with a cross in the middle to attract user attention and as a trigger for synchronization. The exhibition of the five block images is followed by a dark screen for 10 seconds in order for the fNIRS signals to return to their baseline level.

Emotions are known to be very dependent on past experience so that one can never be very sure whether a block elicits the expected emotion or not. To avoid this problem, we asked the participants to self-assess their emotions after the dark-screen resting period, by giving a score between 1 and 5 for respectively valence and arousal components. This reflection period is not time-limited, which in addition has the benefit of providing an interval for relaxing and/or stretching the muscles.

Self-assessment of the images is a good way to have an idea about the emotional stimulation "level" of the subject. However, since noting down this evaluation necessitates some movements in the subject and enforces an additional a prefrontal activity in the brain, some time should elapse for the brain to return to "baseline" before the next image stimulus is offered.

Because of their tight placement, EEG and fNIRS devices can cause some discomfort after a while. For this reason, the whole experiment was divided into three sessions of approximately 15 minutes each. Each session contained 30 blocks, hence 150 images; therefore an experiment consists of a total of 90 blocks or 450 images displayed. The calm and exiting positive classes, containing less than the target number of images were completed with random duplications in different sessions.

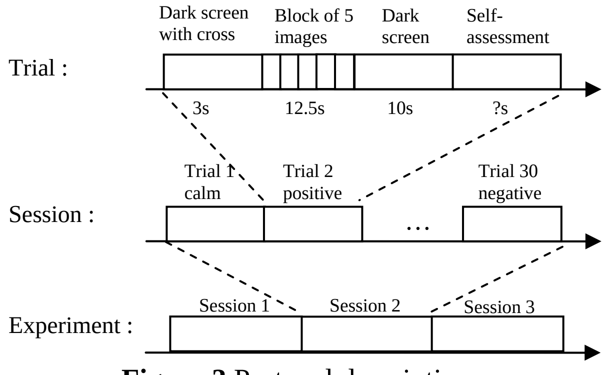

Figure 3 Protocol description

D. *Experimental Protocol for Video and fNIRS*

Three kinds of emotions, namely neutral, happiness and disgust, are stimulated using series of images and video sequences on the screen of the *Stimulus Computer*. With this purpose, two protocols have been tested during the recordings.

The first protocol, the one used in *Session 1,* consisted of 5 videos for each emotion from the DaFEx Database, separated with a 20 seconds of a fixation cross (a white cross over a black background). The second protocol, used in Session 2, was an improvement of the first protocol. It was noticed that the videos were not enough to make the subjects feel the emotions. In order to make the subject to feel the expected emotion better, a sequence of 5 images collected from the internet were added before the first video of the "happy" and "disgust" sequences.

III. Database Collection

A. *Video and fNIRS database*

In the Video-fNIRS database there are totally 16 subjects. While one experiment session is performed for 10 subjects, two sessions of experiments in different days are carried out for the other six subjects. There are six women and 10 men subjects with the average age 25 in the database.

The structure of the database is designed in order to make the video post-processing as easy as possible. Video data are recorded frame by frame into separate files. Each filename is formed by subject name, date, time and the stimuli type as follows:

SubjectName-YYYYMMDD-HHMMSS-
FFF_STIMULI.jpg

where these characters denote:

- YYYYMMDD: year in four digit format and month and day in two digit format
- HHMMSS: time expressed in hour, minutes and seconds using 24 hours format
- FFF: milliseconds
- STIMULUS: type of stimulus (happy, disgust, neutral) shown to the subject when the frame was recorded

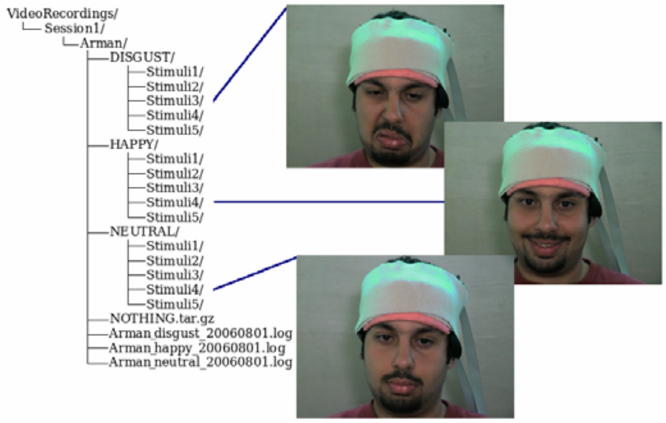

Figure 4 The structure of the database

Moreover, frames are stored in different folders depending on the type of stimulus. Under these conditions, a subject with name Arman recorded in session 1 on the 1st of August would have a folder in the video database with the structure shown in Figure 4, where the file names of the three sample frames are:

Arman-20060801-182136-599_DISGUST.jpg,
Arman-20060801-182641-586_HAPPY.jpg and
Arman-20060801-181726-131_NEUTRAL.jpg.

The frames corresponding to the cross sign, at the beginning of each recording block, are marked as NOTHING since the data is not related to any emotional state.

B. EEG + fNIRS recordings

We recorded data from five participants all male, and right handed, with age ranging from 22 to 38. For each subject data are divided in three repertories, one per session. For each session we obtained three files categories: one concerns EEG and peripheral information, another concerns fNIRS information and the last contains self-assessments of participants.

EEG and peripheral data

EGG, peripheral and the trigger signals are stored in the same BDF (Biosemi Data Format) file. This format is quite the same as the EDF (European Data Format) so that most software could use it without problems; however you can find a converter from BDF to EDF at http://www.biosemi.com/download.htm.

Remember that a trigger is sent in the beginning of each block of images as well as for the start of the protocol. For more convenience, we extracted the samples where such a trigger appears and save them as markers in a MRK file, except for the first trigger.

Finally we obtained two files: a BDF file with EEG and peripheral signals, and a MRK file containing index of samples for each block of images. These files are named as follow:

PART**A**_IAPS_SES**B**_EEG_fNIRS_**DDMMAAAA**.bdf
PART**A**_IAPS_SES**B**_EEG_fNIRS_**DDMMAAAA**.bdf.mrk

where A is the participant number (1-5), B is the session number (1-3) and DDMMAAAA represents the date of the recording.

Common data

In this section, we describe the files that are common to both modalities and concern the protocol in itself:

- IAPS_Images_EEG_fNIRS.txt, contains three columns, one per session, with the names of the IAPS pictures used in this study;
- IAPS_Eval_Valence_EEG_fNIRS.txt and IAPS_Eval_Arousal_EEG_fNIRS.txt contains in three columns the valence or arousal value for each image;
- IAPS_Classes_EEG_fNIRS.txt list in three columns the associated classes we considered for each block of pictures. Labels can be “Calm”, “Pos” or “Neg”. This can be useful if one does not want to take into account self-assessment of participants.

PartASESB.log lists self-assessment of participants. As for the EEG files, A is the number of the participant while B is the session number

C. fNIRS data

fNIRS data were stored in ASCII format with the file name,

SubjectA_SesB_EEG_fNIRS_DDMMAAAA.txt

for EEG + fNIRS recordings and

SubjectA_SesB_video_fNIRS_DDMMAAAA.txt

for video + fNIRS recordings.

where *A* is the participant number and *B* is the session number.

Note that, these files contain raw data, i.e., time-series of concentration changes for three wavelengths. A MATLAB program (loadnirs.m) is needed to convert this signal to oxygenated and deoxygenated hemoglobin values.

D. *Practical considerations and problems*

The most challenging task was making recordings simultaneously from different devices. Each device was designed to be used alone, and thus were not very suitable for multimodal recordings. For instance, EEG cap and fNIRS probe were clearly obstructing each other's functioning. Thus a special probe should be designed which may hold both EEG electrodes and fNIRS light emitting diode and detectors.

Synchronization was the most time consuming task during the workshop. It took a long time before we arrived at a reasonable and reliable solution for synchronizing the devices.

Deciding on the protocol was perhaps the most critical issue in this study. We used a well-known database for EEG and fNIRS recordings, but we tried to adapt it for our purposes. For video and fNIRS, we actually wanted from the subjects to mimic what they saw. Thus it may be argued whether "mimicking" was the same with "feeling" or not.

For video and fNIRS recordings, it is clear that facial muscle movements caused some noise for fNIRS signals.

We could not have the chance to perform the recordings in an isolated experiment room. Thus, environmental noise definitely corrupted our recordings.

During EEG and fNIRS recordings many participants reported that they had a headache at the end of each session. This is due to the different caps that become more and more uncomfortable along time. More over, they also reported that they never felt some strong positive response while they found negative images a bit too hard. Several participants claimed that the effects of the emotional stimuli decrease after viewing many images in succession, suggesting that they became accustomed to the emotional content.

IV. BRAIN SIGNAL ANALYSIS TECHNIQUES

A. *EEG Analysis Techniques*

Prior to extracting features from EEG data and performing classification, we need to pre-process signals to remove noise. Noise can originate from several sources: environment (mainly 50Hz), muscles activity and fNIRS noise. The environmental noise is the easiest to remove by applying a bandpass filter in the 4-45 Hz range. This band is selected because the frequency intervals of interest in EEG are the θ (4-8Hz), α (8-12Hz), β (12-30Hz) and γ (30-45Hz) bands. Muscle activities such as eye-blinks or jaw clenching contaminate EEG signals with strong artifacts. In this study, no special effort was done to remove these artifacts, but subjects were requested to avoid these movements during recordings. One unexpected source of contamination was the fNIRS light activations. As can be observed in Figure 5 fNIRS light activations cause spikes in the EEG recordings, especially in the frontal area. For the moment, no appropriate filtering was designed to remove this type of noise, though independent component analysis (ICA) technique is one potential tool. Finally, to obtain some better focalization on brain activity, we computed a Laplacian reference signal, which consists in subtracting for each electrode the mean signal of its neighbors.

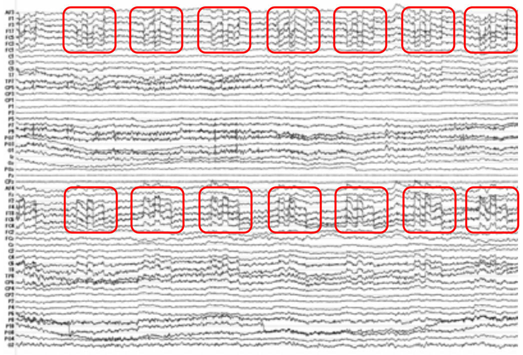

Figure 5 EEG signal sample after pre-processing. fNIRS noise can be observed approximatively every 700ms especially on the frontal electrodes (in red)

Following the preprocessing stage, there are various alternatives for feature extraction. One alternative is to collect EEG energies at various frequency bands, time intervals and locations in the brain. This approach results typically in oversized feature vectors. As a second alternative, Aftanas & al. [10] proved the correlation between arousal variation and power in selected frequency bands and electrodes. These features have also been used in 12 to assess the arousal dimension of emotions. In this project we opted for this first set of features. A third possibility is to compute the STFT (Short Term Fourier Transform) on 12.5 second segments of each trial and electrode, assuming stationarity of the signal within the chosen widow length. This allows taking into account time evolution as well as spatial distribution of energy. Each atom resulting from the STFT is then considered as a feature or relevant features can be selected by filter or wrapper methods [13].

B. *Peripheral Signals Analysis Techniques*

Several studies have shown the effectiveness of peripheral sensors in recognizing emotional states (see 12, 13, 15). While there are many variables from the autonomous nervous system that can be used to determine affective status, we will focus to three such variables: GSR, respiration and blood volume pressure. All these signals were first filtered by a mean filtering to remove noise

GSR provides a measure of the resistance of the skin. This resistance can decrease due to an increase of sudation, which usually occurs when one is feeling an emotion such as stress or surprise. Lang [11] also demonstrates correlation between mean GSR level and arousal. In this study, we recorded GSR by positioning two dedicated electrodes on the top of left index and middle fingers. In order to assess the change in resistance, we used the following features:

Value	Comment
Mean skin resistance over the whole trial	Estimate of general arousal level
Mean of derivative over the whole trial	Average GSR variation
Mean of derivative for negative values only	Average decrease rate during decay time
Proportion of negative samples in the derivative	Importance and duration of the resistance fall

The mean value of samples within a session gives us an estimate of the general arousal level of the emotion while the mean derivative reveals the variability of the signal. Computing the mean of derivative for negative values only, or the proportion of negative values for the whole session indicates the importance of the fall in resistance.

Respiration was recorded by using a respiration belt, providing the chest cavity expansion over time. Respiration is known to correlate with several emotions [13]. For example slow respiration corresponds to relaxation while irregularity or cessation of respiration can be linked to a surprising event. To characterize this activity we used features both in the time and frequency domain. In the frequency domain we computed energy by FFT (Fast Fourier Transform) in 10 frequency bands of size $\Delta f = 0.25$ ranging from 0.25Hz to 2.75Hz. Others features are listed below: (

Value	Comment
Power in the 0.25Hz-2.75Hz ($\Delta f = 0.25$Hz) bands (10 features)	-
Mean of respiration over the whole trial	Average chest expansion
Mean of derivative over the whole trial Standard deviation	Variation of respiration signal
Maximum value minus minimum value	Dynamic range or greatest breath

Finally, a plethysmograph was placed on the thumb of the participant to record his blood volume pressure. This device permits to analyze both relative vessel constriction, which is a defensive reaction [13], and heartbeats that are clearly related to emotions especially in terms of heart rate variability (HRV) (see 11, 13, 15). Heart beats were extracted from the original signal by identification of local maxima, and then the BPM (Beat Per Minute) signal was computed for each inter-beat periods *i*. This enables us to approximate HRV using standard deviation or mean derivative of the BPM signal. The following features were extracted from blood volume pressure:

Value	Comment
Mean value over the whole trial	Estimate of general pressure
Mean of heart rate over the whole trial	-
Mean of heart rate derivative Standard deviation of heart rate	Estimations of heart rate variability

Finally, all these features were concatenated in a single features vector of size 22, representing the peripheral activity.

C. *fNIRS Analysis Techniques*

fNIRS provides us with time series of oxygen-rich (HbO2) and oxygen-poor (Hb) blood concentration changes on the cortical surface. fNIRS signals should be preprocessed first to eliminate high frequency noise and low frequency drifts. Previous studies have shown that involvement of prefrontal cortex in the emotion processing is concentrated in the medial frontal cortex. Thus, it may be a good choice to concentrate on the middle 8 detectors.

Since the hemodynamic response mainly gives an idea about the area of activation, the first line of action has been to detect the presence of active regions in the brain and their variation with stimuli. On the other hand, activated regions are known to vary from subject to subject, and even within subject in the course of experiments. It follows than that detection schemes based only on single subject data may not be reliable enough. One solution to this problem is the use of multivariate methods, that is, simultaneous processing and modeling of data from a group of subjects. Some well-known examples are principal component analysis and independent component analysis. This type of methods may give us the emotion-related components.

The noise caused by facial muscle movement aroused as an important source of contamination for fNIRS signals. Since for some detectors this noise is so large with respect to the signal, it is (and will be) hard to extract cognitive and emotional component from the signals.

D. *Fusion techniques*

The main fusion strategies are data-level fusion, feature-level fusion and decision-level fusion. Due to the disparity of the nature of data in the three modalities, data fusion is not conceivable. On the other hand, fusion at the more abstract levels, feature level and decision level, are both feasible and desirable.

fNIRS-EEG fusion: Recall that the link between electrical activity and hemodynamic activity is supplied by the neurocoupling mechanisms. The EEG modality in one part and the video or fNIRS modality on the other part, have orders of magnitude difference in their relative time scales. However, feature/decision level fusion is possible if one generates fNIRS and EEG feature vectors and/or decision scores for each block of emotional stimuli (12.5 seconds long in our experiment). Alternatively, video features and fNIRS features can be fused at the feature or decision level on a block-by-block basis.

V. VIDEO BASED EMOTION DETECTION

Video signals are quite rich in facially expressed emotions, especially for the happiness and disgust cases. Facial expressions are formed by motions or deformations of mouth, eyebrows and even of eyelids. Also, facial skin may get deformed, such as wrinkles in the forehead or inflations on the cheeks. In this particular experiment, however, we do not have access to the eyebrow information due to the occlusion by the fNIRS probe on the forehead (Figure 6).

We have therefore extracted facial features from mouth and eyes, and then analyzed and classified the data as in the block diagram of Figure 6. We used comparatively two methods for facial feature segmentation: active contour-based technique [3, 4] and active appearance models (AAM) [8]. For the classification we are using Transferable Belief Model (TBM) method [2, 7].

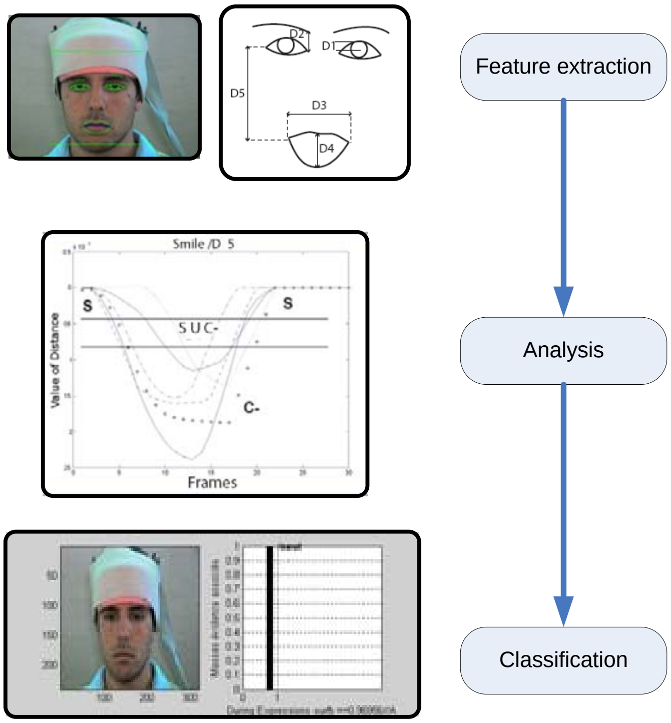

Figure 6 Illustration of the feature extraction, analysis and classification system for emotion detection.

A. *Active Contours for Facial Feature Extraction*

Active contours are widely used for segmentation purposes. However first, the face itself and the eyes must be located. We have used the detector in the Machine Perception Toolbox (MPT) [5]. We had to execute the face detection in each image and bypass its tracking ability due to stability problems. This in turn, slows down the process. Wherever MPT cannot detect a face, we recurred to the OpenCV library face detection tool. The OpenCV algorithm detects faces in general with higher precision albeit at lower speeds than the MPT.

Following fiducial point localization, lips, eyes and eyebrows are segmented by fitting curves automatically and frame-by-frame, using the algorithm described in [3, 4]. This algorithm uses a specific predefined parametric model such that all the possible deformations can be taken into account. The contours are initialized by extracting certain characteristic points, such as eye corners, mouth corners and eyebrows corners automatically. In order to fit the model to the contours, a gradient flow (of luminance or of chrominance) through the estimated contour is maximized. As remarked above, the model fitting to the eyebrows and mouth was not satisfactory, the first due to the occluding fNIRS probe and the latter whenever there were beard and moustache (Figure 7). Even in the absence of such impediments, we have observed that this algorithm works well only when the facial images are neutral (open eyes and closed mouth). Finally, the tracking mode of this algorithm was not available during the workshop. Therefore, we applied the algorithm described in the next section in order to have working results.

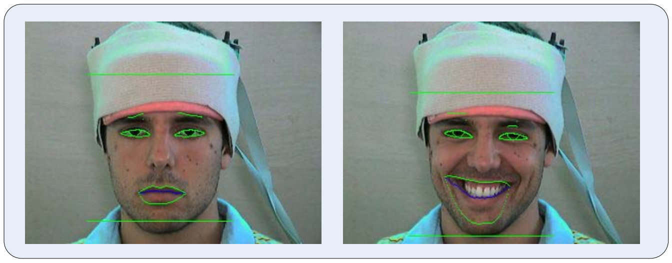

Figure 7 Examples of correct (left) and incorrect (right) localizations

B. *Active Appearance Models (AAM)*

Using Active Appearance Models (AAM) is a well-known technique for image registration [8], in which statistical models of appearances are matched to images iteratively by modifying the model parameters that control modes of shape and gray-level variation. These parameters are learned from a training dataset. The training of the AAM algorithm is initiated with manually annotated facial images, as illustrated in Figure 8. 37 landmark points are chosen from the easily identifiable locations on the face, and their 2D coordinates constitute the shape vector for the face images. In this study, they are chosen appropriately to cover the face, mouth and eye regions for eventual segmentation of face contour, lips and eyes. After creating the shape vectors from all of the training images, they are aligned in a Euclidean frame by Procrustes algorithm. Finally, principal component analysis (PCA) technique is applied to reduce the dimension of the shape vectors, resulting in 11 modes of shape variation that account for 95 percent of variance in the training set.

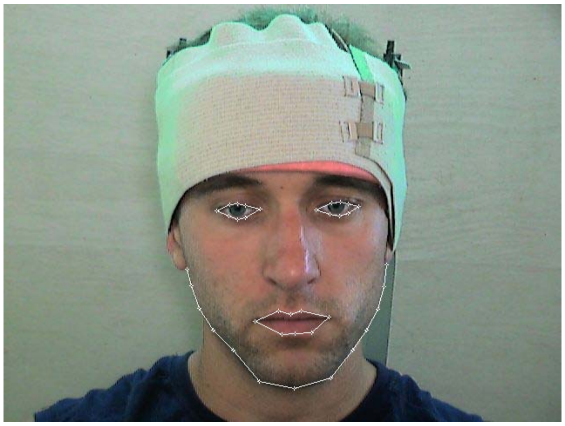

Figure 8 Annotated image with 37 landmark points

The next step is to create the texture vectors for the training images. For this purpose, Delaunay triangulation (Figure 9) is performed so that the triangular patches are transformed to the mean shape (Figure 10). Thus, shape-free patches of pixel intensities are obtained. Also, to diminish the effect of lighting differences, a further alignment in the gray level is performed. Finally, PCA is applied to the texture vectors as in the shape vectors.

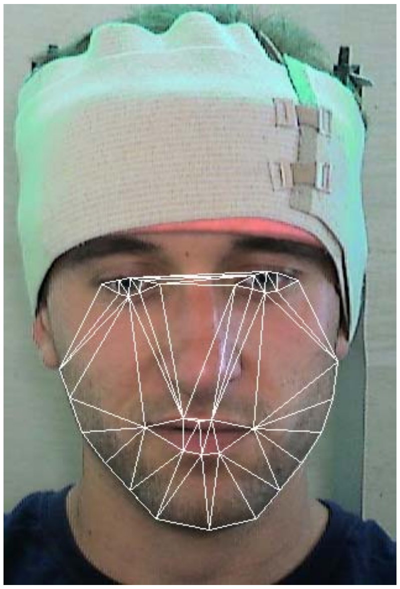

Figure 9 Delaunay triangulation

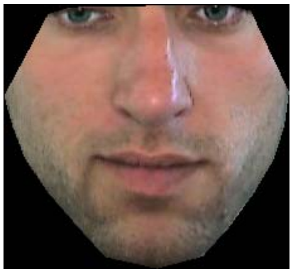

Figure 10 Texture image after transforming to mean shape

The third and final step of the algorithm is to combine shape and texture vectors. These two types of vectors are concatenated into one big feature vector, after appropriate weighting. These weighting coefficients are obtained using an estimation procedure. A final PCA is applied to this combined vector, the resulting vectors being called the appearance vector.

The goal of AAM is to fit models to the images and to synthesize various facial appearances. This is accomplished by modifying and optimizing the combination weights. In face modeling, best pose parameters, which are planar translations, rotations and scaling, are estimated. Briefly, this optimization is realized by iteratively minimizing the difference between the input image pixel intensities and the model instances.

In this study, an AAM is trained and tested for a subject. Some sample results are given in Figure 11, where the detected facial contours corresponding to happiness and disgust moods are tracked in video. For this subject total 26 training images were chosen from the training video database. These images were chosen in order to include different neutral, disgust and happiness expressions with varying head pose and eye motion to cover sufficient amount of face motion.

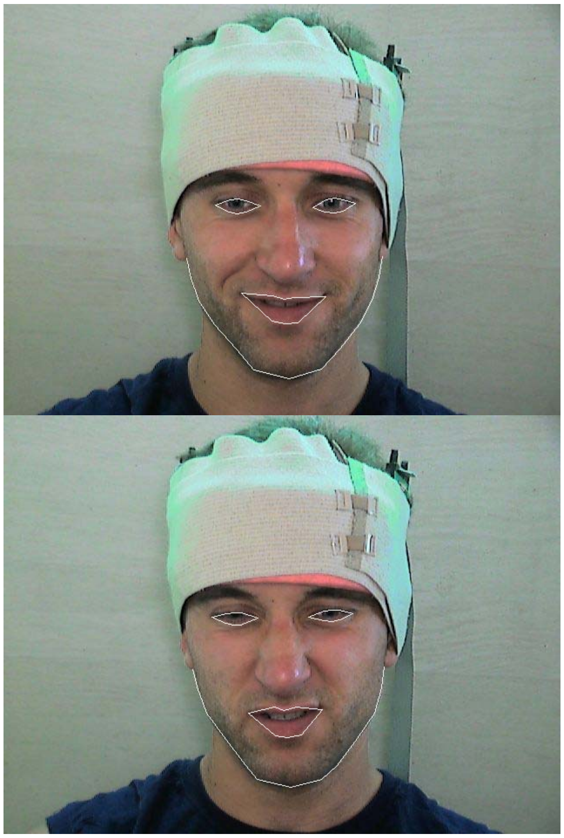

Figure 11 Detected facial contours in video corresponding to happiness (above) and disgust (bottom)

C. *Classification*

In this study, Transferable Belief Model (TBM) algorithm [7], which is based on belief theory, is applied for the classification of anger, disgust and neutral expressions. First, some facial distances, as illustrated in Figure 6, are calculated from the extracted contours. These distances are: eye opening (D_1), distance between the inner corner of the eye and the corresponding corner of the eyebrow (*D2*), mouth opening width (*D3*), mouth opening height (*D4*), distance between a mouth corner and the outer corner of the corresponding eye.

Briefly, in the TBM algorithm each facial expression is characterized by a combination of symbolic states, which are evaluated from the calculated distances. For distance *Di*, the symbolic state is found by thresholding. In Figure 12, the representation of the symbolic states {*C+, C-, S, SC+, SC*} and the thresholds (*a, b, c, d, e, f, g, h*) are shown. While states *C+, C-, S* are representing positive activation, negative activation and no activation, respectively, other two states are denoting the doubt between activation and no activation. The threshold values are found in a training phase automatically as described in [7]. In Figure 12, the y-axis shows the piece of evidence (PE) according to the belief theory. After having found the states and the PEs for each symbolic state of each distance, conjunctive combination rule, which is explained in [7], is applied to combine the information coming from each distance. With this rule, combined PE for each expression is calculated, and the decision is made by choosing the expression that gives highest value for the combined PE. Details about the fusion process can be found in [7].

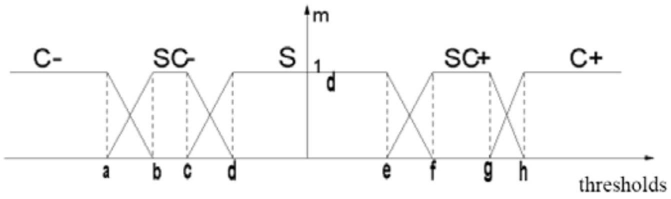

Figure 12 Characterization of a distance with thresholds

VI. Conclusions and future work

Project 7 entitled "Emotion Detection in the Loop from Brain Signals and Facial Images " at Dubrovnik eNTERFACE had three goals in mind:

i) Common database building
ii) Interest and feasibility of these modalities
iii) Assessment of emotion detection performance of individual modalities and their fusion.

Briefly, we have made significant progress in goals 1 and 2, while the 3rd goal must be revisited in a future project.

Common database building: A considerable database containing video, fNIRS and EEG signals has been built. We have already mentioned in Section 6 about the incompatibility of video and EEG. As a consequence, two separate databases were built, one encompassing EEG (including physiological signals) and fNIRS modalities, and the other encompassing video and fNIRS modalities. Second, there was an unpredicted interference effect between EEG and fNIRS setups. The elimination of the EEG & fNIRS interference is not insurmountable, though we did not have time to address the problem during the workshop. Third, the critical synchronization problem between the modality pairs has been ingeniously solved in two alternative ways. The fourth issue was the determination of the proper protocols as well as stimulation material. Although we used the standard methods and materials as in the literature, some subjects reported unsure or inadequate stimulation especially during the prolonged experiments. Subject discomfort and fatigue was another aggravating factor.

Interest and feasibility of these modalities: There is increasing interest in literature for emotion detection and estimation in humans. However, there exist separate literatures, one set of papers published in neuroimaging and neural signal processing journals, the other set of papers appearing in computer vision and man-machine interface journals. We believe the joint use of modalities was for the first time addressed in this workshop, as far as the open literature is concerned. Individual modalities do not fair very well in emotion assessment, hence we believe the multimodal approach will certainly improve the classification performance.

Assessment of emotion detection performance of individual modalities and their fusion:. This part of the project has not been completed and is left as a future work.

ACKNOWLEDGMENT

The authors gratefully acknowledge all participants that volunteered in studies.

REFERENCES

1. Ekman, P., Levenson, R.W., Friesen, W.V., 1983. Autonomic nervous system activity distinguishing among emotions. Science 221, 1208– 1210.

2. Smet, PH. Data fusion in the Transferable Belief Model. Proc ISIF, Frane(2000) 21-33

3. Eveno, N., Caplier, A., Coulon, P.Y.: Automatic and Accurate Lip Tracking. IEEE Trans. On CSVT, Vol. 14. (2004) 706–715.

4. Hammal, Z., Caplier, A. : Eye and Eyebrow Parametric Models for Automatic Segmentation. IEEE SSIAI, Lake Tahoe, Nevada (2004).

5. Machine Perception Toolbox (MPT) http://mplab.ucsd.edu/grants/project1/free-software/MPTWebSite/API/.

6. Open Computer Vision http://opencvlibrary.sourceforge.net/FaceDetection

7. Hammal: Facial Features Segmentation, Analysis and Recognition of Facial Expressions using the Transferable Belief Model 29-06-2006

8. G.J.Edwards, C.J.Taylor, T.F.Cootes, "Interpreting Face Images using Act Active Appearance Models", Int. Conf. on Face and Gesture Recognition 1998. pp. 300-305

9. P.J. Lang, M.M. Bradley, and B.N. Cuthbert, "International affective picture system (IAPS): Digitized photographs, instruction manual and affective ratings", Technical Report A-6, University of Florida, Gainesville, FL, 2005.

10. L.I. Aftanas, N.V. Reva, A.A. Varlamov, S.V. Pavlov, and V.P. Makhnev, "Analysis of Evoked EEG Synchronization and Desynchronization in Conditions of Emotional Activation in Humans: Temporal and Topographic Characteristics", Neuroscience and Behavioral Physiology, 2004, pp. 859-867.

11. J.P. Lang, M.K. Greenwald, M.M. Bradley, A.O. Hamm, "Looking at pictures: affective, facial, visceral, and behavioral reactions", Psychophysiology. 1993 May; 30(3), pp. 261-273.

12. G. Chanel, J. Kronegg, D. Granjean, T. Pun, "Emotion assessment: arousal evaluation using EEG's and peripheral physiological signals",

Proc. Int. Workshop on Multimedia Content Representation, Classification and Security, Istanbul, 2006, pp 530-537.

13. G.H. John, R. Kohavi, K. Pfleger, "Irrelevant Features and the Subset Selection Problem", Machine Learning: Proceedings of the 11th International Conference, San Francisco, 1994, pp. 121-129.

14. J. .A Healey, "Wearable and Automotive Systems for Affect Recognition from Physiology”, PhD Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2000.

15. Lisetti, F. Nasoz, "Using Non-Invasive Wearable Computers to Recognize Human Emotions from Physiological Signals", Journal on Applied Signals Processing, 2004, pp. 1672-1687.