Implicit Emotional Tagging of Multimedia Using EEG Signals and Brain Computer Interface

Ashkan Yazdani ashkan.yazdani@epfl.ch
Jong-Seok Lee jong-seok.lee@epfl.ch
Touradj Ebrahimi touradj.ebrahimi@epfl.ch

Multimedia Signal Processing Group, Institute of Electrical Engineering,
Ecole Polytechnique Fédérale de Lausanne (EPFL), EPFL/STI/IEL/GR-EB
Station 11, CH-1015 Lausanne, Switzerland

ABSTRACT

In multimedia content sharing social networks, tags assigned to content play an important role in search and retrieval. In other words, by annotating multimedia content, users can associate a word or a phrase (tag) with that resource such that it can be searched for efficiently. Implicit tagging refers to assigning tags by observing subjects behavior during consumption of multimedia content. This is an alternative to traditional explicit tagging which requires an explicit action by subjects. In this paper we propose a brain-computer interface (BCI) system based on P300 evoked potential, for implicit emotional tagging of multimedia content. We show that our system can successfully perform implicit emotional tagging and naïve subjects who have not participated in training of the system can also use it efficiently. Moreover, we introduce a subjective metric called "emotional taggability" to analyze the recognition performance of the system, given the degree of ambiguity that exists in terms of emotional values associated with a multimedia content.

Categories and Subject Descriptors

H.5.2 [**INFORMATION INTERFACES AND PRESENTATION**]: User Interfaces – *Evaluation/methodology, Input devices and strategies, Interaction styles;* H.1.2 [**MODELS AND PRINCIPLES**]: User/Machine Systems – *Human factors, Human information processing;* I.5.1 [**PATTERN RECOGNITION**]: Models– *Statistical;* I.5.2 [**PATTERN RECOGNITION**]: Design Methodology – *Classifier design and evaluation, Pattern analysis;* I.5.4 [**PATTERN RECOGNITION**]: Applications– *Signal processing, Waveform analysis;* I.5.5 [**PATTERN RECOGNITION**]: Implementation– *Interactive systems;*

General Terms

Algorithms, Measurement, Performance, Design, Experimentation, Human Factors.

WSM'09, October 23, 2009, Beijing, China.

Keywords

Brain-computer interface, Media annotation, Electroencephalogram, P300, Bayesian linear discriminant analysis, Emotional taggability.

1. INTRODUCTION

A tag for a given multimedia content is a non-hierarchical term describing that content. This metadata helps to provide users with information about it and to make search and retrieval processes for that content easier. Tagging is becoming increasingly popular in many web pages that provide multimedia content as well as social networking web sites and is an important feature of many web 2.0 based services.

In general, there are two ways to assign tags to a given content, namely, explicit and implicit tagging. The former refers to an explicit action by users such as manually typing appropriate keywords associated with the content. Most of social network-based systems such as YouTube [1] and Flickr [2] allow their users to add keywords to the content. On the other hand, implicit tagging means that the users do not necessarily input tags but automatic analysis of the users' behavior is used to generate tags. Explicit tagging is not the ultimate solution for assigning tags to the prevalent multimedia content because there exists a huge amount of multimedia data on Internet. Implicit tagging, however, does not require an effort by users to tag and thus is a promising solution to overcome the limitation of the explicit tagging.

Among various kinds of tags associated with the content, emotional tags play an important role for personalized content delivery [3]. For example, when a user feels sad, he/she may want to watch video clips containing funny stories, which will make him/her feel better. Sometimes, one may not want to watch video clips containing scenes with too much violence. In such cases, emotional tags can be used effectively in multimedia search and retrieval. Consequently, as discussed above, obtaining emotional tags implicitly by assessing the emotions of multimedia content users and assigning related keywords to that content is an important task in this context.

Implicit emotional tagging can be performed by observing users in various ways. From the facial expression analysis, tags related to the emotions for the given content can be extracted. Physiological signals such as respiration, Galvanic skin resistance, skin temperature, eye blinking rate, electromyogram (EMG) and blood flow can also be used to obtain emotional tags [4]. A user's laughter during multimedia content consumption, which is detected by acoustic and visual sensors, can also give a clue about emotional elements of that content [5].

Another modality that when compared to other alternatives has been less considered for implicit tagging is the brain activity of a user while consuming multimedia content. Electroencephalogram (EEG) is a signal that shows the electrical activity of the brain. Numerous studies have shown that this signal alters during certain mental tasks, psychiatric phenotypes, and brain disorders [6], [7]. It has also been proven that such changes in the EEG signal can be detected by means of processing and analysis and hence can be used for interaction between brain and other devices such as computers. Furthermore, several studies have shown that EEG signals can be used for the aim of emotion recognition [8], [9].

In this paper, we propose a novel EEG-based brain-computer interface (BCI) system for implicit emotional tagging of multimedia content. Our system analyzes the P300 evoked potential recorded from user's brain to recognize what kind of emotion was dominant while he/she was watching a video clip. The recognition result of the BCI system is used to assign an emotional tag to that video clip. It is shown that our system can successfully perform implicit tagging for naïve subjects. Also, we introduce a measure of easiness of tagging for the given content, namely, "Emotional Taggability (ET)", which is used to analyze the recognition performance of the system. It is shown that ET and system performance have a correlation in that, for a content with a low ET value, the recognition performance is relatively poor because of the ambiguity of the emotional value of that content, whereas for a content with high ET value, the recognition performance is satisfactory.

The rest of the paper is organized as follows. In Section 2, BCI systems are briefly introduced and explanations about the stimulus-driven BCI strategy used in this paper are given. Section 3 describes the experiments and the methods used for data preprocessing and classification. In Section 4, the results of experiments are discussed. Conclusions follow in Section 5.

2. Brain Computer Interfacing

A BCI is a system, which translates brain electrical activities into executable commands for computer and/or peripheral devices such as wheelchairs, robots, etc. Consequently, users will be able to act on their environment by using their brainwaves instead of peripheral nerves or muscles. One of the goals of BCI research is to develop systems that make it possible for disabled users to communicate with others, to control artificial limbs, or their environment. BCI systems are also increasingly used for other applications such as gaming/entertainment, biofeedback therapy, etc.

Figure 1. EEG signal acquisition setup.

To acquire EEG signals, an electrocap is placed on the scalp of the subject and electrical activities of different regions of brain are recorded. Figure 1 shows an example of a subject playing a video game by means of EEG signals.

Figure 2 illustrates a typical BCI system block diagram. In general, for a BCI application, first, the brain activity is recorded by a signal acquisition device such as EEG electrodes and an EEG amplifier. These signals are then digitized and fed to a computer. Raw EEG signals are usually corrupted with noises and artifacts. Typical sources of such artifacts are electrooculogram (EOG), electromyogram, power lines noise, and slow baseline drifts. Hence, preprocessing algorithms are used to remove such artifacts from the raw signals. After preprocessing, features that are relevant for classification of different mental activities (MA) are extracted. Finally, a classification block uses the extracted features to decide and to recognize which MA was performed by the subject. The output of the classification block can then be translated into commands and used to launch or to control devices.

The system used in this study is a stimulus-driven BCI system based on Event Related Potentials (ERPs), which was initially developed for environment control. Environment control [10] was mainly proposed for disabled subjects who are unable to interact with outside world via their neuromuscular pathways. In stimulus-driven BCI systems, mainly a block of preselected visual and/or audio stimuli is presented to the subject. Under such stimuli, the brain of the subject generates patterns called evoked potentials. These patterns can be detected by analyzing the recorded EEG signals and it can be specified which stimulus among a larger set of possible stimuli has drawn the subject's attention.

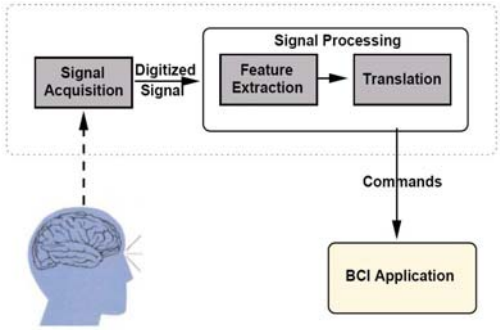

Figure 2. Basic block diagram of a BCI system [11]

An endogenous ERP that has gained much attention in the neuroscience and medical research communities is called P300 (see Figure 3). P300 is an interesting and fruitful research topic considering that it can be reliably measured, and also because the characteristics of P300 waveform, such as its amplitude and latency, can be influenced by various factors. Many studies have linked the characteristics of P300 to subject's specific factors such as gender, age, or brain disorders such as Alzheimer or schizophrenia. More details about P300, its origin and its current research challenges can be found in the following reviews [12], [13].

To evoke P300, different stimulus modalities and paradigms can be used. Regarding the stimulus modality, auditory, visual, tactile, gustatory, or olfactory stimuli types can be presented to the

subject. However, for practical reasons, often auditory or visual stimuli are preferred. Regarding the paradigms, either oddball or three-stimulus paradigms are used. In oddball paradigm, two different stimuli, namely target (or oddball) and non-target stimuli are used. The two stimuli are presented in a random sequence but the target stimulus appears rarely. Subjects are instructed to respond to each occurrence of the target stimulus and to ignore the non-target stimuli. For example subjects can be instructed to react with a button press to each 1000 Hz tone in a random sequence of 1000 Hz and 2000 Hz tones. In three-stimulus paradigm, however, a third distracter stimulus is also used. This stimulus appears in the sequence with the same frequency as target stimulus but the subject is asked to neglect it and does not perform any task when observing this stimulus.

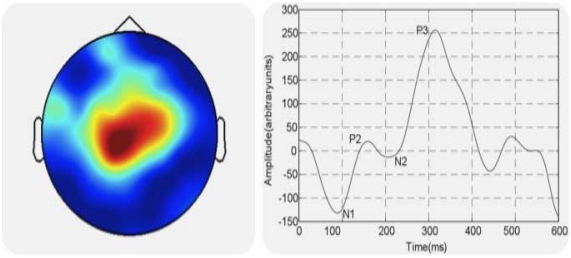

Figure 3. Spatial (left) and temporal (right) patterns of P300 component.

Different types of P300 can be observed in the two paradigms described above. In classical oddball paradigm, the target stimuli evoke the so-called P3b. The P3b has a latency of about 300-500 ms and can be observed mostly over centro-parietal brain regions. This component appears only if subjects pay attention to the stimuli. When subjects do not pay attention to the stimuli, the target stimuli in the oddball paradigm evoke a different type of P300 – the so-called P3a [14]. The P3a has a latency of about 200-400 ms and can be observed mostly over fronto-central brain regions.

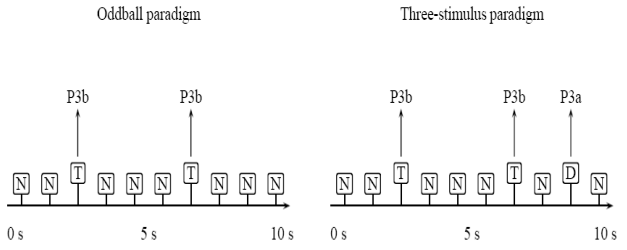

Figure 4. Paradigms for evoking the P300. Left: In oddball paradigm, a sequence of target (T) and non-target (N) stimuli is presented in a random order. Right: In three-stimulus paradigm, distracter (D) stimuli are added to the sequence of target and non-target stimuli [15].

In three-stimulus paradigm, the target stimuli also evoke P3b. The distracter stimuli, however, evoke P3a. The relation between the different paradigms and P3a and P3b components is summarized in Figure 4.

In addition to its dependence on different experimental paradigms, P300 is also influenced by many other factors. Some important factors influencing P300 are target probability, interstimulus interval, habituation, attention, and difficulty of the task. This shows that P300 is not a static, fixed phenomenon but rather an inherently variable response of the brain, occurring in situations during which, novel or improbable and task-relevant stimuli have to be processed.

3. MATERIALS AND METHODS

3.1 Experimental setup

Users faced a desktop monitor and were asked to watch some video clips. These video clips were collected from YouTube [1] for our experiments. Four clips were chosen for each of the six basic emotional categories defined by Paul Ekman (i.e. joy, sadness, surprise, disgust, fear, and anger) and thus 24 clips were used in total for the experiment. In order to ensure that the duration of the test session remains reasonable, we mostly chose relatively short video clips. The minimum, mean and maximum lengths of the clips were 15, 58, and 161 seconds, respectively.

Immediately after each video ended, six images were displayed on the screen (see Figure 5). These images were happy, sad, surprised, disgusted, afraid, and angry faces representing the six basic emotions. They were flashed in a pseudo-random order, one image at a time. During each flash, one image was intensified for 100ms followed by 300ms during which, none of the images were intensified, so that an interstimulus interval of 400ms was achieved.

Figure 5. The graphical user interface used to evoke the P300. Images were intensified, one at a time, by modification of their overall brightness.

The EEG signals were acquired at 2048 Hz sampling rate from 32 electrodes that were placed on the scalp of the subjects according to the 10-20 international electrode positioning system. A Biosemi Active Two amplifier was used for amplification and analog to digital conversion of the recorded EEG signal.

Signal processing and pattern recognition algorithms used in this study were implemented in Matlab. The stimulus display and the online access to EEG signals were implemented as dynamic link libraries (DLLs) in C. The DLLs were accessed by Matlab via a MEX interface. These algorithms and interfaces were

implemented and tested initially for P300-based environment control BCI [10].

3.2 Experimental schedule

In this study, we trained the BCI system with eight healthy subjects that were Ph.D. students recruited by our laboratory (all male, age 29±3.4). None of subjects had any known neurological deficiencies. After training a general classifier, we tested the BCI system with four other subjects (all male, age 29±1.5), who had never used the system.

In the training phase, each subject was asked to complete four training sessions for recording the EEG signals. The first two sessions were performed during one day and the remaining two sessions were performed on another day within 2 weeks after the first session. Each session consisted of six runs, one run for each image. During each run, the images of the GUI illustrated in Figure 5 were flashed in a pseudo-random order.

The subjects were asked to perform a covert task, i.e. silently count how many times a prescribed image was intensified (for example: "Now please count how often the sad face flashes"). After this massage, the six images were shown on the screen and a warning beep was played. The images then started to flash according to a random sequence staring 4ms after the preparation beep and simultaneously the EEG signals were recorded. The sequence of images to be flashed was block-randomized. In other words, after each block (6 flashes), each image was flashed one time, and after two blocks (12 flashes) each image was flashed twice and so forth. The number of the blocks inside each run was selected randomly between 20 and 25. Therefore, for instance, a sequence might include 23 blocks, which provided 23 target (P300) trials together with 23×5=115 non-target (non-P300) trials. At the end of each run, the subjects were asked to report the result of their counting. This number then was compared to the actual number of blocks to monitor the performance of the subject and also to know whether he/she was concentrated throughout the test.

A simple classifier was built based on the data recorded in the first session, and at the end of each run during the second, third, and fourth sessions, the image inferred by the classification algorithm was flashed five times so that the subjects can have a feedback of their performance. The duration of one run was approximately one minute given mean value of 22.5 blocks and six image intensification of 0.4 second long each, inside a block and the four seconds preparation time (22.5×6×0.4)+4=58 seconds. Each session took approximately 30 minutes, including the setup of electrodes and short breaks between runs and, comprised on average of 810 trials. The whole data gathered for each subject consisted on average of 3240 trials.

3.3 Data processing

The recorded EEG data has to be preprocessed and some features should be extracted from each trial in order to perform classification and to learn the discriminating functions. In this section the preprocessing and feature extraction methods are described.

During the signal acquisition, two electrodes were placed on the mastoids of the subject. The average signal of these two electrodes was used for referencing. In the next step, a sixth order forward-backward Butterworth bandpass filter with zero phase shift was used to filter the data. The cut-off frequencies of the bandpass filter were set to 1.0 Hz and 12 Hz. Computation of the filter coefficients was done using *butter* function in Matlab and to perform the forward-backward filtering *filtfilt* function was utilized. The data was then downsampled from 2048 Hz to 32 Hz. To this end, an eighth-order lowpass Chebyshev Type I filter with a cutoff frequency of 12.8 Hz was used. The input sequence was filtered in both forward and reverse directions to remove all phase distortion, effectively doubling the filter order. The decimation was performed using the *decimate* function in Matlab.

To extract the single trials from the whole EEG data gathered during each run, windows of duration 1000ms were extracted from the data. The stimulus presentation interface provided the exact system clock of the stimulus onset as well. Single trials started at stimulus onset, i.e. at the beginning of the intensification of an image, and ended 1000ms after the stimulus onset. It is worth to mention that the last 600 ms of each single trial overlaps the 600 ms of its following single trial, due to the fact that the interstimulus interval was set to 400ms.

After the data was broken down into single trials and downsampled, it needs to be purified from artifacts. Normally during EEG acquisition, eye blinks, eye movement, muscle activity or subject movement can cause large amplitude outliers in the EEG. As the source of these peaks are not directly related to brain activities, they might influence the results of the classification in that they can simply be mistaken as P300 peaks. To reduce the effects of such outliers, the data from each electrode was windsorized in the following manner. For the samples of each electrode, the 10^{th} percentile and also the 90^{th} percentile were computed. Amplitude values that fall bellow the 10^{th} percentile or above 90^{th} percentile were then replaced by the 10^{th} and 90^{th} percentiles respectively.

In the next step, the samples of each electrode were scaled to the interval [-1,1], and finally these normalized samples were concatenated to constitute the feature vectors. Considering the number of electrodes which is 32, and the number of decimated temporal samples 32 for each single trial, the dimensionality of each feature vector representing each single trial was 32×32=1024.

3.4 Bayesian linear discriminant analysis

In this study, the Bayesian Linear Discriminant Analysis (BLDA)[1] was used for classification. BLDA can be seen as an extension of Fisher's Linear Discriminant Analysis (FLDA). In contrast to FLDA, BLDA uses regularization to prevent overfitting to high dimensional and possibly noisy datasets. Through a Bayesian analysis, the degree of regularization can be estimated automatically and quickly from training data without the need for time consuming cross-validation. Algorithms that are closely related to this method are the Bayesian least-squares support vector machine and the algorithm for Bayesian non-linear discriminant analysis described in [16]. BLDA is also closely related to the so-called evidence framework for which detailed accounts are given in [17].

[1] A Matlab Implementation can be downloaded from the webpage of our group http://bci.epfl.ch

As a starting point for the description of BLDA we use the fact that FLDA is a special case of least squares regression. Least squares regression is equivalent to FLDA if regression targets are set to $\frac{N}{N_1}$ for examples from class 1, and to $\frac{N}{N_2}$ for examples from class -1 (where N is the total number of training examples, N1 the number of examples from class 1, and N2 the number of examples from class -1). A proof for the equivalence between least squares regression and FLDA can be found in [17]. Given the connection between regression and FLDA, our approach for BLDA is to perform regression in a Bayesian framework and set target values as mentioned above.

The assumption in Bayesian regression is that targets *t* and feature vectors *x* are linearly related with additive white Gaussian noise *n*.

$$t = \omega^T x + n \tag{1}$$

Given this assumption, we can write down the likelihood function for the weights ω used in regression:

$$p(D|\beta,\omega) = (\frac{\beta}{2\pi})^{\frac{N}{2}} \exp(-\frac{\beta}{2}\left\|X^T\omega - t\right\|^2) \tag{2}$$

Here X indicates the matrix that is obtained from the horizontal stacking of the training feature vectors, D denotes the pair {X, t}, and β refers to the inverse variance of the noise. It is assumed that the feature vectors contain one feature that always equals one; the bias term, which is commonly used in regression, can thus be omitted.

To perform inference in a Bayesian framework, we have to specify a prior distribution for the latent variables, i.e. for the weight vector ω. The expression for the prior distribution is:

$$p(\omega|\alpha) = (\frac{\alpha}{2\pi})^{\frac{D}{2}} (\frac{\varepsilon}{2\pi})^{\frac{1}{2}} \exp(-\frac{1}{2}\omega^T I'(\alpha)\omega) \tag{3}$$

where I′ (α) is a square, d +1 dimensional, diagonal matrix is:

$$I'(\alpha) = \begin{bmatrix} \alpha & 0 & \dots & 0 \\ 0 & \alpha & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \varepsilon \end{bmatrix} \tag{4}$$

and d is the number of features. The prior for the weights thus is an isotropic, zero-mean Gaussian distribution with inverse variance α. The effect of using a zero-mean Gaussian prior for the weights is similar to the effect of the regularization term used in ridge regression and regularized FLDA. The estimates for ω are shrunk towards the origin and the danger of overfitting is reduced. The prior for the bias (the last entry in ω) is a zero-mean univariate Gaussian with inverse variance ε. Setting ε to a very small value, the prior for the bias is practically flat. This expresses the fact that a priori we do not make any assumptions about the value of the bias parameter.

Given the likelihood and the prior, the posterior distribution can be computed using Bayes rule.

$$p(\omega|\beta,\alpha,D) = \frac{p(D|\beta,\omega)\,p(\omega|\alpha)}{\int p(D|\beta,\omega)\,p(\omega|\alpha)\,d\omega} \tag{5}$$

Since both the prior and the likelihood are Gaussian, the posterior is also Gaussian and its parameters can be derived from the likelihood and the prior by completing the square. The mean m and covariance C of the posterior satisfy the following equations.

$$m = \beta(\beta XX^T + I'(\alpha))^{-1} Xt \tag{6}$$

$$\mathrm{C} = (\beta \mathrm{XX}^{\mathrm{T}} + I'(\alpha))^{-1} \tag{7}$$

By multiplying the likelihood function (Equation 2) for a new input vector x' with the posterior distribution (Equation 5) followed by an integration over ω, we obtain the predictive distribution, i.e. the probability distribution over regression targets conditioned on an input vector:

$$p(t'|\beta,\alpha,x',\omega) = \int p(t'|\beta,x',\omega)\,p(\omega|\beta,\alpha,D)\,d\omega \tag{8}$$

The predictive distribution is again Gaussian and can be characterized by its mean μ and its variance σ^2.

$$\mu = \mathrm{m}^{\mathrm{T}} \mathrm{x}' \tag{9}$$

$$\sigma^2 = \frac{1}{\beta} + \mathrm{x}'^{\mathrm{T}}\,\mathrm{C}\mathrm{x}' \tag{10}$$

In the P300-based BCI described in the present study, only the mean value of the predictive distribution was used for taking decisions. More precisely, mean values were summed over trials and the image corresponding to the maximum of the summed mean values was then selected.

In a more general setting, class probabilities could be obtained by computing the probability of the target values used during training. Using the predictive distribution from Equation 8 and omitting the conditioning on β, α, and D we could use:

$$p(y'=1|x') = \frac{p(t'=\frac{N_1}{N}|x')}{p(t'=\frac{N_1}{N}|x') + p(t'=\frac{-N_2}{N}|x')} \tag{11}$$

Both the posterior distribution and the predictive distribution depend on the hyperparameters α and β. In the above we have assumed that the hyperparameters are known, however in real-world situations the hyperparameters are usually unknown. One way to solve this problem would be to use cross-validation to determine the hyperparameters that yield the best prediction performance. However, the Bayesian regression framework offers a more elegant and less time-consuming solution for the problem of choosing the hyperparameters. The idea is to write down the likelihood function for the hyperparameters and then maximize the likelihood with respect to the hyperparameters. The maximum likelihood solution for the hyperparameters can be found with a simple iterative algorithm, which we do not discuss in detail here, but a detailed discussion about it can be found in [17].

4. RESULTS

In this section, the results of the aforementioned processing and classification methods are presented and discussed.

We have developed a general classifier, using the training data gathered from eight subjects, and tested this classifier with another four naïve subjects, who had never been through the training phase. More precisely, in the test experiment each of the four subjects was asked to watch 24 video clips. One run of BCI

was performed immediately after each video ended. Therefore, the total duration of each test session, including the setup of the EEG signal acquisition equipment was around 90 minutes.

Before the beginning of the test experiment, the subjects were asked to select one image on the screen using their brainwaves, so that an appropriate number of blocks (B) for each subject can be defined (cf. Section 3.2). In this way, the proper value of B was chosen for each test subject separately, which varied between 6 and 10. For each run in the test session, the single trials corresponding to the B blocks were extracted using the processing techniques described in Section 3.3. In the next step, the single trials were classified using the BLDA classifier. In [10], the performances of FLDA and BLDA classifiers were compared using different electrode combinations. It has been shown that when using all the 32 electrodes, the performance of BLDA will be clearly better than FLDA.

The classifications of single trials resulted in B blocks of outputs so that each block consisted of six classifier outputs, one output for each image on the display. In order to make the final decision about the selected image, i.e. to recognize which image was selected by the subject, the classifier outputs were summed over the B blocks for each image and finally the image with the maximum summed classifier output value was selected. Table 1 shows the performance of the four subjects for annotation of 24 video clips.

Table 1. The rates of the correctly annotated video clips using the proposed BCI system for the test subjects

Subject 1	Subject 2	Subject 3	Subject 4	Average
79.17%	91.67%	70.83%	79.17%	80.19%

While the users were asked to choose only one emotional category for each video clip in our system, such task may be difficult for some video clips due to ambiguity of the messages conveyed in the clips or simultaneous elicitation of multiple emotions. This difficulty would deteriorate the accuracy of the recognition in our system in that, the EEG signal recording and the image presentation graphical user interface are already running but the subject still hesitates to choose one among the six emotional categories and consequently he/she is unable to generate P300 patterns properly.

Therefore, we conducted additional subjective experiment for further analysis of the relationship between the recognition performance of our proposed system and the difficulty of tagging for the users. More precisely, we asked another 18 subjects to rate the video clips by assigning an integer value from 0 to 10 for each of the six emotional categories. For example, if a subject does not feel sad at all for a clip, he/she enters '0' for sadness; if he/she feels that the video clip is very sad, he/she enters '10' for sadness. Thus, for each video clip and for each subject, six integer numbers corresponding to the six emotional categories (e.g. {3 6 0 8 10 2}) were obtained. To perform this subjective test, a webpage containing all the 24 video clips used in the test was created, Figure 6 shows a screenshot of this webpage and the rating scheme.

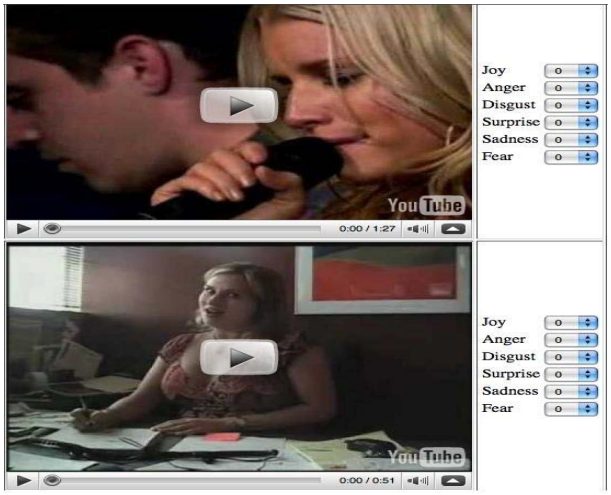

Figure 6. Screenshot of the page (2 video clips) created for the subjective experiment. Subjects were asked to rate the emotions that they associated with a given video.

By using the rating values obtained from the aforementioned subjective test, we defined a measure of "Emotional Taggability (ET)" for each video. ET indicates the feasibility and easiness of assigning an emotional tag to a video clip by using the described BCI system in this paper. It is defined in such a way that, if the six integer values are roughly equal or small, then the ET value should also be small. On the other hand, we want the ET value to be large when only one of the six values is dominantly large and the others are small. Defining a measure for this purpose has been explored in the field of audio-visual speech recognition, and it has been shown that the following definition is the best among several measures proposed in literature [18]:

$$ET = \sum_{i=1}^{M} \left(\max_{1 \le j \le M} e_j - e_i \right) \tag{12}$$

where e_i is the rating value of the i-th emotion among the M emotional categories (M=6 in our case). However, the above measure is not complete for our purpose because we also need to distinguish the two cases {10, 5, 5, 5, 5, 5} and {5, 0, 0, 0, 0, 0} which will result in the same ET value using Equation 12. For this, the above measure is weighted by the maximum value among e_i's and, thus, we finally define the ET of a video clip as follows:

$$ET = \sum_{i=1}^{M} \max_{1 \le j \le M} e_j \left(\max_{1 \le j \le M} e_j - e_i \right) \tag{13}$$

Figure 7 illustrates the average values of ratings performed by 18 subjects sorted in an ascending order with respect to their ET values. In this figure, the bright pixels and the dark pixels represent the high and low rating values, respectively. It can be easily concluded that a video with high ET value inducea only one dominant emotion. For instance, by comparing video 10, and 21 in Figure 8, it can be seen that the ratings are higher on average for video 10, but video 21 has a higher ET value, as it has only one dominant rating.

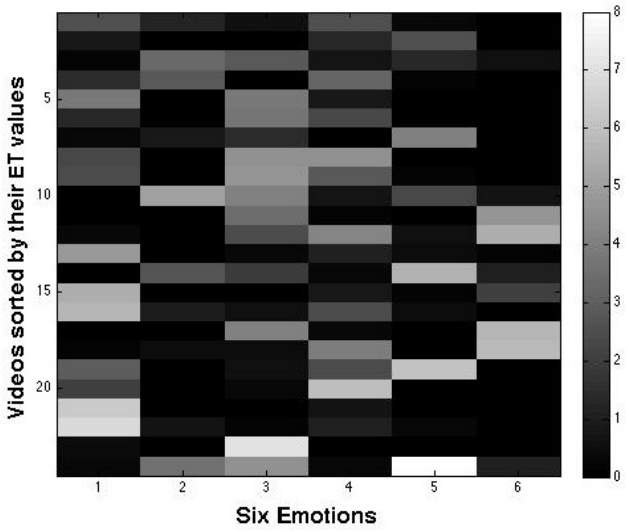

Figure 7. The ratings of the video clips sorted by the ET value of videos (i.e. video 1 and video 24 have the lowest and highest ET values, respectively).

Figure 8 shows the relationship between the ET value and the recognition (implicit tagging) performance of our BCI system. The performance value was measured among the test subjects. For instance, the value of 75% would mean that three among the four subjects could tag the video successfully. It can be observed that there exists a correlation between the ET value and the performance of the system, which supports our analysis of the recognition performance based on the ET. In other words, this figure implies that if the ET value for a given video is high enough (here above 25), then the users of the proposed BCI system can tag this video with a high accuracy.

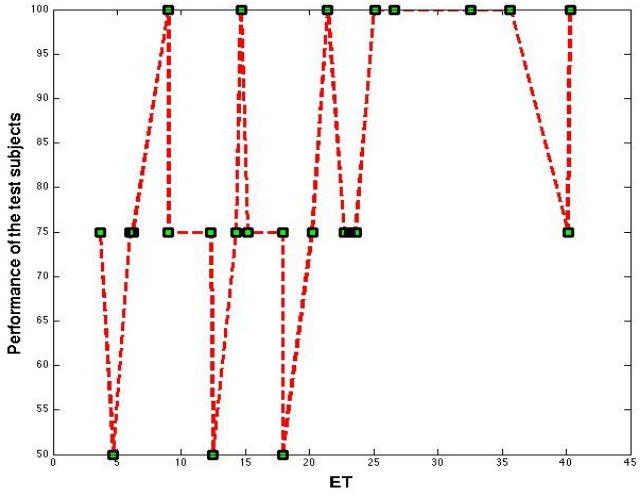

Figure 8. The performance of the proposed system for the four subjects vs. the ET value of different video clips used in this study.

As it can be seen in Figure 8, the video that has the second highest ET value resulted in a recognition accuracy of 75%. This video was the 21st video during the test session and the subject who made the error while tagging this video reported that he was too tired to fully concentrate, since the presentation of this video and the annotation occurred almost 80 minutes after the beginning of the session.

5. CONCLUSION

In this paper, a new modality for implicit emotional tagging of multimedia has been presented. More precisely, it has been shown that people can annotate the multimedia content by means of their brainwaves and a P300-based BCI system. This system can be of great interest, considering the minimization of user efforts to perform the annotation and/or enabling the disabled subjects to perform such annotation just like the normal people. To this end, a general classifier was trained after recording the EEG signals of eight subjects. This classifier has been tested with four naïve subjects and promising results were obtained.

Further improvements to the work presented in this paper might be to test the developed system with disabled (locked–in) subjects, performing tests with larger numbers of subjects in order to confirm the results obtained in the present work, and also to use other biological signals such as skin conductance, ECG, reparatory rate and etc. synchronized with EEG signals to make a multimodal analysis of biosignals. Moreover, the EEG signals can be assessed directly during consumption of multimedia for recognition of the induced emotion.

6. ACKNOWLEDGMENTS

Many thanks to all subjects who volunteered to participate in the training and/or test experiments described in this paper. We would also like to thank Ulrich Hoffmann whose work, which was initially designed for environment control for disabled subjects, was a starting point of this research. The research leading to these results has been performed in the frameworks of European Community's Seventh Framework Program (FP7/2007-2011) under grant agreement no. 216444 (PetaMedia) and Swiss national centre of competence in research (NCCR) on Interactive Multimodal Information Management (IM2).

7. REFERENCES

[1] http://www.youtube.com

[2] http://www.flickr.com

[3] Hanjalic, A. and Xu, L.-Q. 2005 Affective video content representation and modeling. IEEE Trans. Multimedia, 7,1, 143-154.

[4] Soleymani, M., Chanel, G., Kierkels J., and Pun, T. Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses. In Proceedings Int. Symposium on Multimedia, (Berkeley, California, USA, December 15-17, 2008).

[5] Petridis, S. and Pantic M. Audiovisual laughter detection based on temporal features. In Proceedings Int. Conf. on Multimodal Interfaces, (Chania, Crete, Greece, October 2008). 37-44.

[6] Anderson, C., Devulapalli, S., and Stolz, E. 1995 Determining mental state from EEG signals using parallel implementation of neural networks. Scientific programming, 4, 3, 171-183.

[7] Adeli, H., Zhou, Z., and Dadmehr, N. 2003 Analysis of EEG records in an epileptic patient using wavelet transform. J. Neurosci. Methods, 123,1, 69-87.

[8] Ekman P, Levenson R.W, and Freison W.V, 1983 Autonomic Nervous System Activity Distinguishes Among Emotions. J. Exp Soc Psychol, 195-216.

[9] Kim K.H, Band S.W, and Kim S.B, Emotion Recognition System using short-term monitoring of physiological signals. In Proceedings on Medical & Biological Engineering & Computing,. 2004, 419-427.

[10] Hoffmann, U., Vesin, J.-M., Ebrahimi, T., and Diserens, K. 2008 An efficient P300-based brain-computer interface for disabled subjects. J. Neurosci. Methods, 167,1, 115-125.

[11] Thorpe, J., van Oorschot, P. C., and Somayaji, A., Pass-thoughts: authenticating with our minds. In Proceedings of the ACSA 2005 New Security Paradigms Workshop, (Lake Arrowhead, California, USA, September 2005, 45-56.

[12] Hurby, T. and Marsalek, P. 2003 Event-related potentials-The P3 wave. Acta. Neurobiologiae. Experimentalis, 63, 1, 55-63.

[13] Nieuwenhuis, S., Aston-Jones, G., and Cohen J. 2005 Decision making, the P3, and the locus coeruleus-norepinephirine system. Psychol. Bull, 131,4, 510-532.

[14] Squires, N.K., Squires, K.C., and Hillyard, S.A. 1975 Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. Electroencephalogr. Clin Neurophysiol, 38, 4, 387-401.

[15] Hoffmann, U. 2007 Bayesian machine learning applied in a brain-computer interface for disabled users. PhD thesis. Ecole Polytechnique Fédérale de Lausanne (EPFL).

[16] Centeno, T.P. and Lawrence N.D. 2006 Optimising kernel parameters and regularisation coefficients for Non-linear discriminant Analysis. J. Mach. Learn. Res., 7, 455-491.

[17] Bishop, C.M. 2006 Pattern recognition and machine learning. Springer.

[18] Lee,J.-S. and Park, C.H. 2008 Adaptive decision fusion for audio-visual speech recognition. Speech recognition, technologies and applications, I-tech, 275-296.