Physio-visual data fusion for emotion recognition

C. Maaoui *, F. Abdat, A. Pruski

Laboratoire de Conception, Optimisation et Modélisation des Systèmes, Université de Lorraine, Lorraine, France

Received 5 March 2013; received in revised form 21 February 2014; accepted 17 March 2014
Available online 3 May 2014

Abstract

Several approaches have been proposed to recognize human emotions based on facial expressions or physiological signals, relatively rare work has been done to fuse these two, and other, modalities to improve the accuracy and robustness of the emotion recognition system. In this paper, we propose two methods based on feature-level and decision-level to fuse facial and physiological modalities. At feature-level fusion, we have tested the mutual information approach for selecting the most relevant and principal component analysis to reduce their dimensionality. For decision-level fusion, we have implemented two methods; the first is based on voting process and the second is based on dynamic Bayesian Networks. The system is validated using data obtained through an emotion elicitation experiment based on the International Affective Picture System. Results show that feature-level fusion is better than decision-level fusion.
© 2014 Elsevier Masson SAS. All rights reserved.

1. Introduction

Anxiety disorders are psychiatric disorders characterized by a constant and abnormal anxiety that interferes with daily-life activities. Their high prevalence in the general population and the severe limitations they cause have drawn attention to the development of new and efficient strategies for their treatment. The evocation and detection of the user’s emotional state is becoming a crucial element in the aim for developing more effective interfaces between humans and computers, especially in applications such as games and affective computing [1,2].

Emotional state can be obtained from a broad range of behavioral cues and signals that are available via visual, auditory and physiological expressions or presentation of emotions:

- visual: the affective state is evaluated as a function of the modulations of emotions on facial expressions, gestures, postures, and generally body language. The data are captured through a camera, allowing for non-intrusive system configurations. The systems are generally very sensitive to the video quality both in terms of Signal to Noise Ratio (SNR) and in terms of illumination, pose, and size of the face on the video and is the most sensitive to false, acted facial expressions;
- auditory: the affective state can be estimated as a modulation of the vocal signal. In this case, data are captured through a microphone, once again, allowing for non-intrusive system configurations [3]. The estimation can be very accurate. The processing needs clean voice data; SNR inferior to 10 dB can severely reduce the quality of the estimation [4]. Furthermore, the processing still cannot handle the presence of more than one voice in the audio stream;
- physiology: the affective state is appraised through the modulations emotions exert to the Autonomic Nervous System (ANS). Signals such as heart beat or skin conductivity are detected through ad hoc input devices. The estimation can be very reliable [5,6] and it is less sensitive to the acting of emotions than the one extracted from the auditory and visual modalities. The main limitation is related to the intrusiveness of the sensing devices.

In this paper, we describe an intelligent solution for the monitoring of patients with anxiety disorders during therapeutic sessions based on automatic emotion recognition. Our emotion recognition system is based with consideration of facial expression and physiological signals. It recognizes an individual’s affective state based on positive and negative emotions. We
analyze external (facial expression) and internal factors (physiological signals) of human responses to determine what the inherent emotion is.

Several advantages can be expected when combining biosensors feedback with affective facial. First, the use of the bio-sensors allows to continuously gathering information on the user’s affective state while the analysis of emotions from facial expressions should only be triggered when the user face is in front of the camera. Secondly, it is much harder for the user to deliberately manipulate biofeedback than external channels of expressions or speech. According to Mehrabian [7], there are basically three elements in any face-to-face communication: words, tone of voice and body language. His study concluded that the most communication is non-verbal. Words account for 7% of our communication, tone of voice accounts for 38% and body language accounts for 55%. From another side, the user may consciously or unconsciously conceal his/her real emotions by external channels of expression. Finally, an integrated analysis of biosignals and facial expression may help to resolve ambiguities and compensate for errors.

Due to complementarity and redundancy of the data coming from the two channels physio-face human affect recognition is expected to perform more robustly than uni-modal methods. Thus, affect recognition should inherently be the issue of the multimodal analysis. In this paper, we will present a comparative study for bimodal system. The remainder of this paper is organized as follows: first, we describe related works to recognize the emotions of human user. Feature extraction is detailed in Section 3. In Section 4, we describe different approaches of different level fusion used by our bimodal system. The used protocol for emotion induction is described in Section 5. Experimental results are illustrated in Section 6. Finally, conclusion and future works are presented in the last section.

2. Previous work

Accordingly, reviewing the efforts toward the single-modal analysis of artificial affective expressions have been the focus in the previously published survey papers, among which the papers of Cowie et al. in 2001 [8] and of Pantic and Rothkrantz in 2003 [9] have been the most comprehensive and widely cited in this field.

Also, it has been shown by several experimental studies that integrating the information from audio and video leads to an improved performance of affective behavior recognition. An increased number of studies on audio-visual human affect recognition have emerged in last years [10–12]. Zhihong et al. [13] introduce and survey the recent advances in the research on human affect recognition.

Only few works have investigated the possibility to fuse together visual and physiological affective estimation [14,15].

In [14], Bailenson et al. present automated, real-time models built with machine learning algorithms which use videotapes of subjects’ faces in conjunction with physiological measurements to predict rated emotion (trained coders’ second-by-second assessments of sadness or amusement). Input consisted of videotapes of subjects watching emotionally evocative films along with measures of their cardiovascular activity, somatic activity, and electrodermal responses. They built algorithms based on extracted points from the subjects’ faces as well as their physiological responses. Strengths of their current approach are (1) they are assessing real behavior of subjects watching emotional videos instead of actors making facial poses, (2) the training data allow to predict both emotion type (amusement versus sadness) as well as the intensity level of each emotion, (3) they provide a direct comparison between person-specific, gender-specific, and general models. Results demonstrated good fits for the models overall, with better performance for emotion categories than for emotion intensity, for amusement ratings than sadness ratings, for a full model using both physiological measures and facial tracking than for either cue alone, and for person-specific models than for gender-specific or general models.

Chuang et al. propose an emotion recognition system with consideration of facial expression and physiological signals in [15]. Specifically designed mood induction experiment is performed to collect facial expressing images and physiological signals of subjects. They detected 14 feature points and extracted 12 facial features from facial expression images. Meanwhile, they measure the skin conductivity, finger temperature and heart rate from the subject. Both facial and physiological features are adopted to train the classifiers. Two learning vector quantization (LVQ) neural networks were applied to classify four emotions: love, joy, surprise and fear. Experimental results show the proposed recognition system is able to identify four emotions by facial expressions, physiological signals, and both of them. The odd sample points of physiological signals were used for training the LVQNNs, and the remaining samples were used for testing.

The contributions of this paper include not only a new means for emotion recognition, but also the finding of significant classification rates from bimodal data corresponding to two affective states measured from 10 subjects over many days of data. Next section describes feature extraction for each modality.

3. Features extraction

3.1. From facial expression

Face detection is the first step in our facial expression recognition system. This step allows an automatically labeling for facial feature points in a face image. For this, we have used a real-time face detector proposed in [16], which represents an adapted version of the original Viola-Jones face detector. Detection of facial feature points is the key step in our facial expression analysis system, we have used a simple and fast method to detect automatically facial feature points, based on our anthropometrical model combined to Shi&Thomasi method [17] for more accuracy.

The human facial expressions originate from the movements of facial muscles beneath the skin. Thus, we represent each facial muscle by a pair of key points [18], namely dynamic point and static point. As shown in Fig. 1a, the dynamic points can be moved during an expression, while Fig. 1b shows the fixed points which cannot be moved during a facial expression (face edge, nose root and outer corners of the eyes). To clarify, when
3.2. From physiological signals

Today, it is recognized that the physiological changes occupy an important place in emotional experiences. For this reason, we are interested in the physiological changes, in order to create a recognition emotion module. The James-Lange theory [20] states that emotions are the perceptions of certain bodily changes. In other words, emotions are feelings that are caused by physiological changes induced by the autonomic nervous system. Such changes include the modifications of heart rate, muscular tension, skin conductance, etc.

Recently, the low consistency of physiological configurations supported the hypothesis that the autonomic nervous system ANS activation during emotions indicates the demands of a specific action tendency and action disposition, instead of reflecting emotions [21].

In this work, we have selected five physiological signals: Blood volume pulse (BVP), Electromyography of eyebrows (EMG), Skin Conductance of the finger (SC), Skin Temperature of the finger (SKT) and Respiration (RESP) to extract 30 features for recognition. We have chosen these five signals because they are most used in the literature for emotion recognition. The used sensors are less intrusive than others. For example, the ECG requires a bare torso and the EEG requires a high accuracy placement of electrodes on the head.

Having established a set of signals which may be used for recognizing emotion, it is then necessary to define a methodology in order to enable the system to translate the signals coming from these sensors into specific emotions. The first necessary step was the extraction of useful information bearing features for pattern classification.

For emotion recognition training or testing, the features of each bio-potential data must be extracted. In this study, we have chosen a simple method that gives good performance proposed by Picard [22]. For each record, we compute the six parameters on the N values: the mean of the raw signals \( \mu_x \), the standard deviation of the raw signals \( \sigma_x \), the mean of the absolute values of the first differences of the raw signals \( \delta_x \), the mean of the absolute values of the first differences of the normalized signals \( \tilde{\delta}_x \), the mean of the absolute values of the second differences of the raw signals \( \gamma_x \) and the mean of the absolute values of the second differences of the normalized signals \( \tilde{\gamma}_x \). By using these feature values, the feature vector \( X \) is defined as follow:

\[
X = \begin{bmatrix}
\mu_{\text{bvp}} & \sigma_{\text{bvp}} & \delta_{\text{bvp}} & \gamma_{\text{bvp}} & \bar{\mu}_{\text{bvp}} & \bar{\sigma}_{\text{bvp}} & \bar{\delta}_{\text{bvp}} & \bar{\gamma}_{\text{bvp}} \\
\tilde{\delta}_{\text{bvp}} & \tilde{\gamma}_{\text{bvp}} & \mu_{\text{resp}} & \sigma_{\text{resp}} & \tilde{\delta}_{\text{resp}} & \tilde{\gamma}_{\text{resp}} & \bar{\mu}_{\text{resp}} & \bar{\sigma}_{\text{resp}} & \bar{\delta}_{\text{resp}} & \bar{\gamma}_{\text{resp}}
\end{bmatrix}
\]

4. Bimodal system

Information from multiple sources can be consolidated in several distinct levels, including the data level, feature extraction level and decision level.

In this paper we discuss fusion at the feature level and at decision level as follows:

1. feature fusion:
   (a) feature selection with mutual information;
   (b) feature transform with principal component analysis (PCA).

2. decision fusion:
4.1. Feature fusion

Feature fusion is the process of combining features extracted from the raw measurements, whether with selection methods or transform methods:

1. feature selection with mutual information;
2. feature transform with principal component analysis (PCA).

4.1.1. Feature selection

Feature selection means choosing from an initial set of features only the ones which are relevant for the classification task, having as a result a reduced set of features.

The main advantages of this technique are:

- alleviating the effect of the curse of dimensionality;
- enhancing generalization capability;
- speeding up learning process;
- improving model interpret-ability.

We present in the following a feature selection method based on mutual information.

Mutual information is considered to be a good indicator of the relevance of two random variables [23]. Formally, if \( I(Y;C) \) is the mutual information between a feature \( Y \) and the class labels \( C \), it represents the amount of information gained about the class if the feature \( Y \) is used. A high mutual information here shows that the feature is relevant for our classification task and should be part of the subset of selected features.

Denoting two discrete random variables by \( X \) and \( Y \), their mutual information can be defined in terms of their probability density functions (pdfs) \( p(x), p(y) \), and \( p(x,y) \) as:

\[
I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}
\]  

(1)

To select \( p \) features we have:

1. Minimized the redundancy:

\[
\min W_1, \quad W_1 = \frac{1}{|S|^2} \sum_{i,j \in S} I(i,j)
\]

\( S \) is the set of initial features, \( I(i,j) \) is the mutual information between \( i \) and \( j \).

2. Maximized the pertinence:

\[
\max V_1, \quad V_1 = \frac{1}{|S|^2} \sum_{i \in S} I(h,i)
\]

\( h \) is the correspondent class (emotion).

3. Combined between the minimization of the redundancy and the maximization of the pertinence by maximizing:

\[
\max (V - W)
\]

4.1.2. Feature transforms

Feature transforms, or feature extraction methods [24], produce a set of new features based on all the features in the original set. This means that all the original features are needed, and there is no reduction in the requirements for data collection. The new features are obtained as the result of applying a transform, either linear or non-linear, on the initial feature vectors. This leads to a change in the representation of the data itself, meaning for example that it could be better visualized and understood. Feature transforms can also be supervised or unsupervised. We chose a very popular transforms which is used to reduce the dimensionality of data the PCA method [25].

4.2. Decision fusion

Decision fusion is the process of combining partial decisions, given by the different modalities, since each single module uses expert knowledge to transform the information carried by the measured data into a decision [26]. Decision fusion methods can be divided into two categories (parametric or non-parametric).

4.2.1. Non-parametric methods: Vote

The principle of voting method for information fusion is easier to implement. Let \( S_j(x) = i \) that the source \( S_j \) assigns the class \( C_i \) to the observation \( x \). We assume here that the classes \( C_i \) are mutually exclusive. For each source, we associate the indicator function:

\[
M_j^i = \begin{cases} 
1 & \text{if } S_j(x) = i \\
0 & \text{else}
\end{cases}
\]

(2)

The combination of sources written by:

\[
M_k^E(x) = \sum_{j=1}^{m} M_j^i(x)
\]

(3)

The class having the highest vote is declared as the final opinion.

4.2.2. Parametric methods: Dynamic Bayesian Network (DBN)

Graphical models are a marriage between probability theory and graph theory. They provide a natural tool for dealing with two problems that occur throughout applied mathematics and engineering: uncertainty and complexity [27]. Particularly they are playing an increasingly important role in many medical applications due to their capacity to represent uncertain knowledge.

A Bayesian network (also known as Belief Network or Directed Probabilistic Independence Network) [28,29] is a graph with probabilities for representing random variables and their
dependencies. It efficiently encodes the joint probability distribution of a set of variables. Its nodes represent random variables and its arcs represent dependencies between the random variables encoded by conditional probabilities.

Dynamic Bayesian Networks (DBN) [27] encode the joint probability distribution of a time-evolving set of variables \( X [t] = \{ X_1 [t], \ldots , X_N [t] \} \). For a given time interval of size \( T \), a Dynamic Bayesian Network can be considered as a ‘static’ Bayesian network with \( T \times N \) variables. DBN are a generalization of Kalman Filter Models (KFM) and Hidden Markov Models (HMM) [30].

Next section presents our emotion recognition results with previous different methods of fusion.

5. Emotion induction protocol

A prevalent method to induce emotional processes consists of asking an actor to feel or express a particular mood. This strategy has been widely used for emotion assessment from facial expressions and to some extent from physiological signals [31]. However, even if actors are known to deeply feel the emotion they try to express, it is difficult to insure physiological responses that are consistent and reproducible by non-actors. Furthermore, emotions from actor-play databases are often far from real emotions found in everyday life.

The alternate approach for inducing emotions is to present particular stimuli to an ordinary participant. Various stimuli can be used such as images, sounds, videos [5] or video games. This approach presents the advantages that there is no need for a professional actor and that responses should be closer to the ones observed in real life. It was essential to obtain a database of physiological signals and facial expression representing specific emotional stat. To acquire a database of physiological signals and facial expression in which the influence of emotional status was faithfully reflected, we developed a set of elaborated protocols for emotion induction. We use the International Affective Picture System (IAPS) developed by Lang et al. [32], and adopted for many psychophysiological studies involving emotion induction.

The training data were taken from a study in which 10 subjects (8 males and 2 females): graduate and undergraduate students from different disciplines, looked IAPS images to elicit positive and negative emotions over 4 days. The choice of images used for induction was done by a psychologist.

Fig. 2 shows samples of facial expression images from our database.

We have implemented our approach with a standard web-cam and no specific illumination or background conditions, and we have used five physiological signals [Blood Volume Pulse (BVP), Electromyography (EMG), Electrodermal activity (SC), Skin temperature (SKT) and Respiration (Resp)].

For each subject, we presented 60 images during 300 seconds to induce emotions with positive and negative valence (5 second/image). The scroll of images is automatic, alternating between positive images and negative images with a pause is needed.

A person should be able to sit in front of his computer and have his affective state appraised in real time.

The physiological signals were acquired using the PROCOMP Infiniti system [33]. The sampling rate was fixed at 256 samples per second for all the channels. Appropriate amplification and bandpass filtering were performed. One session of experiments took approximately 5 min. The subjects were requested to be as relaxed as possible during this period. Subsequently, emotional stimulus was applied, and both physiological signals and facial expression were recorded.

Fig. 3 shows our bimodal system outline with different levels of fusion.

6. Implementation and discussion

For each modality, we have chosen Support Vector Machine SVM as classifier. Its implementation was done using libSVM [34], For every set membership, the parameter setting is the same for all the population. The SVM parameters are:

1. Kernel type: a radial basis function (RBF) kernel was used, given by:

\[
K (x, x_i) = \exp \left( -\frac{\| x - x_i \|^2}{\sigma} \right)
\]  

(4)

The RBF shows a greater performance rate for classifying non-linear problems than other types of kernels.

2. Cross-validation: in order to compute the accuracy of each SVM, we perform k-fold cross-validation.

We performed data classification by considering two factors:

1. the data to be classified:
   (a) in the dependent case, the results of emotion recognition is evaluated on test corpus for subjects who participate in the training corpus,
   (b) in the independent case, the results of emotion recognition are evaluated on test data for subjects who did not participate in the training corpus;

2. acquisition: to see the impact of the duration of acquisitions we have constructed two databases:
   (a) Database-4D: the acquisition of training corpus and test corpus is performed during 4 days,
   (b) Database-1D: the acquisition of training corpus and test corpus is performed in the same day.

We classified the bimodal data subject-dependently and subject independently since this gave us a deeper insight on what terms the multimodal systems could improve the results of unimodal emotion recognition.

6.1. Uni-modal system

Table 1 presents the emotion recognition rates for unimodal system using facial expressions and physiological signals.
Fig. 2. Samples of facial expression images: neutral, positive, negative emotions.

separately. During the experience, we noted that the best recognition rates for some subjects from database-4D are obtained using facial expressions. The best recognition rates for the other subjects are obtained using physiological signals. This is due to the change of the emotional state over 4 days.

The day acquisition dependence is probably related to variations in physiology which is due to two factors: (1) caffeine, sugar, sleep, hormones, and other non-emotional factors; (2) the mood or inability to build an intense experience of joy if the subject felt a strong sadness that day. Nonetheless, many of these sources of variation are natural and cannot be controlled in realistic long-term measuring applications [22].

However, for database-1D, the best results are obtained using physiological signals for all subjects.

Fig. 3. Bimodal system outline with different levels of fusion.
Two first bars of Fig. 4a and b show the mean and variance values of classification results of facial expressions and physiological signals separately for database-4D and database-1D. In general case, the best recognition rates for database-4D are obtained using facial expressions (58.56%) and for database-1D are obtained using physiological signals (85.62%). These results do not reveal a conclusion to generalize which modality is best for emotion recognition, for this we have chosen the use of both modalities.

It is reasonable to expect that some characteristic patterns of the emotions can be obtained by the use of physiological or facial features, for example (EMG from physiological features, eyebrows from facial features). This redundant information is very valuable to improve the performance of the emotion recognition system when the features of one of the modalities are inaccurately acquired. For example, if a person has hair on the brow or eyeglasses, the facial expressions will be extracted with high level of error. In that case, physiological features (EMG) can be used to overcome the limitation of the visual information.

6.2. Bimodal system

The fusion characteristics for emotion recognition is a critical phase which allows to select a relevant set which ensures the coherence between the two modalities and the homogeneity of the data.

6.2.1. Feature fusion

6.2.1.1. Concatenation results (Con). The concatenation of data is a simple method. It takes all the features of facial expressions and physiological signals without any processing and sorting. The classification rates for this method are presented in the third bars of Fig. 4a and b for database-4D and database-1D respectively. We note that the recognition rate for database-4D using concatenation method (56.49%) is lower than recognition rate using only facial expressions (58.56%). Also, for database-1D, the recognition rate using concatenation method (84.39%) is lower than recognition rate using only physiological signals (85.62%). In conclusion, the concatenation of characteristics of the two methods did not improve system performance.

6.2.1.2. Mutual information results (MI). Among 21 facial features and 30 physiological features, we want to identify those that contribute more in the classification. Because it is forbiddingly time-consuming to exhaustively search the best of features that give best classification, we apply Mutual Information MI to identify the subset that are more important in distinguishing affects.

The algorithm of MI sort the features by relevance. The results of the selection algorithm suggest that the six important features for database-4D are:

1. the mean of the absolute value of the first differences of the normalized signal of BVP;

![Image](image-url)

Fig. 4. Fusion results for dependent case.

### Table 1

<table>
<thead>
<tr>
<th>Base-4D</th>
<th>Exp (%)</th>
<th>Phy (%)</th>
<th>Base-1D</th>
<th>Exp (%)</th>
<th>Phy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>64.3</td>
<td>62.1</td>
<td>1</td>
<td>88.62</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>53.09</td>
<td>59.7</td>
<td>2</td>
<td>33.86</td>
<td>83.33</td>
</tr>
<tr>
<td>3</td>
<td>63.7</td>
<td>49.95</td>
<td>3</td>
<td>58.38</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>65.02</td>
<td>46.45</td>
<td>4</td>
<td>53.86</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>44.9</td>
<td>40.95</td>
<td>5</td>
<td>65.90</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>51.95</td>
<td>54.15</td>
<td>6</td>
<td>66.90</td>
<td>99.85</td>
</tr>
<tr>
<td>7</td>
<td>55.06</td>
<td>42.87</td>
<td>7</td>
<td>48.71</td>
<td>99.95</td>
</tr>
<tr>
<td>8</td>
<td>68.79</td>
<td>49.8</td>
<td>8</td>
<td>52.86</td>
<td>99.76</td>
</tr>
<tr>
<td>9</td>
<td>74.7</td>
<td>41.39</td>
<td>9</td>
<td>62.33</td>
<td>64.09</td>
</tr>
<tr>
<td>10</td>
<td>44.1</td>
<td>47.32</td>
<td>10</td>
<td>25.99</td>
<td>33.33</td>
</tr>
<tr>
<td>All</td>
<td>54.29</td>
<td>49.74</td>
<td>All</td>
<td>68.62</td>
<td>91.13</td>
</tr>
</tbody>
</table>

a: acquisition during 4 days  
b: acquisition in the same day

The numbers in bold present the best recognition rate obtained using facial expressions or physiological signals for each subject.
2. the mean of the absolute value of the second differences of the normalized signal of BVP;
3. the mean of the absolute value of the first differences of the normalized signal of EMG;
4. the mean of the absolute value of the second differences of the normalized signal of EMG;
5. the mean of the absolute value of the first differences of the normalized signal of SC;
6. the mean of the absolute value of the second differences of the normalized signal of SC.

Using database-1D, the MI method was given a different features selection for each person.

Based on the results of bimodal features selection, we chose the first six features. The fifth bars of Fig. 4a and b present the basic classification rates with MI method for the database-4D and database-1D respectively. Overall, the performance obtained with the MI is better compared to results obtained with the uni-modal systems and compared to the concatenation.

For database-4D, we obtained 62.88% of recognition rate using MI method, while the percentage of emotion recognition is 58.56%, 49.46% and 56.49% using facial expressions, physiological signals and concatenation method respectively. For database-1D, we obtained 86.80% of recognition rate using MI method, while the recognition rate is 55.74%, 85.62% and 84.39% using facial expressions, physiological signals and concatenation method respectively. This improvement is due to the relevance of features selected using MI method.

6.2.1.3. PCA results. We apply PCA method to find a well transform of features in another space. The fourth bars of Fig. 4a and b show the classification results using PCA.

As we can see in the Fig. 4, the PCA method performs much better than single modality, concatenation and MI methods. The results using PCA method of database-1D (90.30%) are much better compared to the database-4D (62.88%). Dates acquired during several days are more dispersed compared to data acquired during a single day.

6.2.2. Decision fusion
6.2.2.1. Vote. The voting process used for decision fusion is a simple method that delivers results per second without any weighting of the two methods. The results obtained with voting process for database1-D (79.93%) are much better compared to the classification results of database4-D (52.17%). This is due to the quality of each uni-modal decisions; for database1-D, we have obtained good performance using facial expressions and physiological signals separately than those for database4-D (Fig. 4a and b).

6.2.2.2. DBN. Fig. 5 illustrates the structure of our Dynamic Bayesian Network used for decision fusion for emotion recognition. The used network consists of three nodes, two observations (facial expressions decision and physiological signals decision) and one for decision fusion. The estimation of conditional probabilities is based on the learning parameters with the EM (Expectation-Maximization) [35]. The used inference algorithm is based on the junction tree detailed in [27]. The implementation of RBD is made with the library PNL (Probabilistic Network Library) [36].

The Fig. 4 shows that the results obtained with the database-4D (58.40%) and the database-1D (56.83%) are approximately similar. By comparing between decision fusion methods, we note that the RBD gave better results compared to voting for a database-4D, contrary to the database-1D where the best results are obtained with the vote. This can be explained by the fact that the voting method does not consider the right decisions for each classifier in a time interval, while RBD take into account the error of each classifier [37].

6.3. Person-independent affect recognition

In this section, we explore person-independent affect recognition in which testing data and training data are from different subjects. Thus the variation of the data is more significant and classification is more challenging than person dependent recognition.

We apply all cited methods in this paper to Person-Independent recognition. The performance of SVM from face-only, Biosignals-only, features transform with PCA, features selection with MI, voting process and RBD method are given in Fig. 6.

Among the five fusion methods mentioned above, concatenation of features gave the poorest performance (48.58% for database-4D and 50% for database-1D). The best result was obtained with PCA method.
Table 2
Confusion matrix for database-4D.

<table>
<thead>
<tr>
<th></th>
<th>Positive(^a) (%)</th>
<th>Negative(^a) (%)</th>
<th>Positive(^b) (%)</th>
<th>Negative(^b) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>78.9</td>
<td>21.1</td>
<td>11.6</td>
<td>88.4</td>
</tr>
<tr>
<td>Negative</td>
<td>21.1</td>
<td>78.9</td>
<td>4.2</td>
<td>95.8</td>
</tr>
</tbody>
</table>

\(^a\) Subject 3 with PCA (5 components).
\(^b\) Independent case with PCA (5 components).

6.4. Confusion matrix

Table 2 represents confusion matrix for subject 3 (dependent case) and 10 subjects (independent case) respectively for the database-4D.

For a single person (dependent case), the distinction between classes is very good 78.9\%, but there is poor distinction between classes for independent case, we can explain this by the dominance of a single emotion with the participant.

6.5. Self-assessment results

For our experience, we used IAPS images as an inductor. The images are classified as positive or negative valence emotion. During the measurement session, we have noticed that the pictures did not have the same effect on every subject. For this reason, self-assessment step has been applied to all subjects to get a better classifying for the IAPS images. We have recalculated the rate of recognition of the database-4D with the new classification of images.

Fig. 7 shows a comparison between the results in the dependent case with all methods mentioned above using IAPS annotation and self-assessment. Overall, the rates obtained with self-assessment are better than the rates obtained with IAPS annotation. Feature fusion based on PCA method gives the best results.

7. Conclusions

In this paper we have presented a bimodal system to recognize emotion from facial expression and physiological signals. The proposed system estimates the affective state of individuals by classifying features extracted from facial expression, BVP, EMG, SC, SKT and RESP biosignals. The system is validated using data obtained through an emotion elicitation experiment based on the International Affective Picture System.

We have developed a feature-level fusion technique and decision-level technique for emotion recognition. The used features were extracted from facial expression and physiological signals.

For feature-level, we have used PCA to transform feature and mutual information to select the best sub-features. For decision-level, we have tested two methods, voting process and Dynamic Bayesian Network.

Feature-level fusion based on PCA gives the best results versus mutual information and versus decision-level fusion.

On the basis of the experimental results, it is showed that the features level fusion gives better performance than single modality and the decision level fusion. The main reason can be
found in the fact that the used features in each modality were not
designed to avoid dependencies among the chosen parameters.
There is no features selection method applied.

The best results for our emotion recognition system are obtained when both training and test data are recorded at the
same day, where the affective state is more stable.

Among others, the contribution of this work resides in sev-
eral aspects: First, the combination of physiological signals and
facial expressions rarely used for emotion recognition. Then,
the construction of the database which was realized over 4 days
to get the real change in the emotion of each subject. The test
set was performed separately, we did not work on the same of
measure set for learning and for test. To solve the problem of
the used inductor, we have proposed a self-evaluation stage to
show the effect of the annotation of images and particularly the
effectiveness of the methods presented in this paper.

As a future work, we intend to find a more efficient inductor
and to study its influence on emotion recognition. We plan
also, to incorporate other means of emotion recognition such
as speech recognition.

The developed system is planned to be clinically tested with
real patients.

References

[1] Lee-Johnson C, Carnegie D. Mobile robot navigation modulated by artifi-
tional model for a guide robot. IEEE Trans Syst Man Cybern A Syst Humans
[3] Noble J. Spoken emotion recognition with support vector machines, Ph.D.
multi-stream fused hidden Markov model for audio-visual affect recogni-
[5] Lisetti C, Nasoz F. Using noninvasive wearable computers to recog-
nize human emotions from physiological signals. J Appl Signal Process
[6] Villon O, Lisetti C. Towards building adaptive users psychophysologi-
cal maps of emotions using bio-sensors. In: Workshop on Emotion and
[9] Pantic M, Rothkrantz L. Toward an affect-sensitive multimodal human-
Modeling naturalistic affective states via facial and vocal expression recog-
methods: Audio, visual, and spontaneous expressions. IEEE Transac-
et al. Real-time classification of evoked emotions using facial feature track-
eration of facial expression and physiological signals. In: Proceeding of
IEEE Symposium on Computational Intelligence in Bioinformatics and
[16] Viola P, Jones M. Robust real-time object detection, 2nd international
workshop on statistical and computational theories of vision - modeling,
[17] Shi J, Tomasi C. Good features to track. IEEE Conf Computer Vision and
Pattern Recognition Seattle CVPR ’94; 1994.
action, IEEE RO-MAN08. Germany: The 17th International Symposium
on Robot and Human Interactive Communication; 2008.
Intel Corporation Microprocessor Research Labs; 2000.
[21] Davidson RJ. Parsing affective space: perspectives from neuropsychology
[22] Picard RW, Vyzas E, Healey J. Toward machine emotional intelli-
gence: analysis of affective physiological state. IEEE Trans Pattern Anal
Series in Telecommunications; 2006.
[24] Liu K, Motoda H. Feature selection for knowledge discovery and data
[25] Pearson K. On lines and planes of closest fit to systems of points in space.
Philos Mag 1901:2:559–72.
[26] Verlinde P. A contribution to multi-modal identity verification using
decision fusion. Signal et images, l’École Nationale Supérieure des Télé-
communications; 1999.
[27] Murphy P. Dynamic Bayesian networks: representation, inference and
[28] Jensen F, Lauritzen S, Olsen K. Bayesian updating in recursive graphical
[29] Cowell RG, Dawid P, Lauritzen SL, Spiegelhalter DJ. Probabilistic
[30] Rabiner LR. A tutorial on hidden Markov models and selected applications
[31] Herbelin B, Benzaki P, Riquier F, Renault O, Thalmann D. Using physiologi-
cal measures for emotional assessment: a computer-aided tool for cognitive
(iaps): digitized photographs, instruction manual and affective ratings.
Technical report A-6; University of Florida.
[34] Chang C, Lin C. Libsvm: library for support vector machines.
Eyrolles; 2008.
[37] Kim E, Kim W, Lee Y. Combination of multiple classifiers for the customer’s purchase behavior prediction. Decis Support Syst
2003;34(2):167–75 [doi:http://dx.doi.org/10.1016/S0167-9236(02)00079-
9].