Disponible en ligne sur
ScienceDirect
www.sciencedirect.com
Elsevier Masson France
EM|consulte
www.em-consulte.com
IRBM
IRBM 35 (2014) 109–118

Original article

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.

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 Autonomous 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

* Corresponding author.
E-mail address: choubeila.maaoui@univ-lorraine.fr (C. Maaoui).

http://dx.doi.org/10.1016/j.irbm.2014.03.001
1959-0318/

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], Bailonson 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