Computer Vision and Image Understanding 147 (2016) 114–124

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Multi-modal emotion analysis from facial expressions and electroencephalogram

Xiaohua Huang*, Jukka Kortelainen, Guoying Zhao, Xiaobai Li, Antti Moilanen, Tapio Seppänen[1], Matti Pietikäinen[1]

Center for Machine Vision and Signal Analysis, Department of Computer Science and Engineering, University of Oulu, P.O Box 4500, FI-90014, Oulu, Finland

ARTICLE INFO

Article history:
Received 26 March 2015
Accepted 29 September 2015

Keywords:
Multi-modal emotion recognition
Facial expression
Electroencephalogram
Feature representation
Pattern recognition
Signal processing

ABSTRACT

Automatic analysis of human spontaneous behavior has attracted increasing attention in recent years from researchers in computer vision. This paper proposes an approach for multi-modal video-induced emotion recognition, based on facial expression and electroencephalogram (EEG) technologies. Spontaneous facial expression is utilized as an external channel. A new feature, formed by percentage of nine facial expressions, is proposed for analyzing the valence and arousal classes. Furthermore, EEG is used as an internal channel supplementing facial expressions for more reliable emotion recognition. Discriminative spectral power and spectral power difference features are exploited for EEG analysis. Finally, these two channels are fused on feature-level and decision-level for multi-modal emotion recognition. Experiments are conducted on MAHNOB-HCI database, including 522 spontaneous facial expression videos and EEG signals from 27 participants. Moreover, human perception in emotion recognition compared to the proposed approach is tested with 10 volunteers. The experimental results and comparisons with the average human performance show the effectiveness of the proposed multi-modal approach.

1. Introduction

Emotions are a central part of human communication. They are fundamental to humans, impacting on our perception and everyday activities such as communication, learning and decision-making. It is widely agreed that they are a multi-modal procedure involving facial expressions, speech, gestures and some physical characteristics, as shown in Fig. 1, and should have a key role in human–computer interactions [10,31,52]. Application scenarios include analyzing emotions while the person is watching emotional movies or advertisements, playing video games, driving a car, is under health monitoring or crime investigation, or is participating in interactive tutoring.

While computers are expected to naturally interact with humans, an emotion recognition technique should be able to process, extract and analyze a variety of cues through multi-modal procedure. Recently, multi-modal emotion recognition has gained significant scientific interests [3,32,44,45]. These works utilized various kinds of channels involving facial expressions, speech and physiological signals for emotion recognition. Among these, facial expression is an intuitive measurement for computers to understand human beings' emotions, while electroencephalogram (EEG) is an internal measure from the brain, making an interesting alternative for multi-modal emotions recognition. So far, there are few works attempting to consider facial expression and EEG together for spontaneous emotion recognition [21]. This paper proposes a new approach for multi-modal emotion recognition fusing facial expression and EEG for recognizing emotions from long continuous videos.

1.1. Related work

Facial expression is probably the most important non-verbal communication channel. Facial expressions have been directly linked to the emotional state experienced by the sender [10] and have been shown to be an important source of information regarding the emotional state of others. They can reveal how people are feeling and what their attitude and behavioral intentions are.

For recent decades, research on facial expression analysis has been developed from posed (acted) to spontaneous facial expressions [52], from isolated to continuous [19], from obvious to subtle expressions [39]. The recent studies [38,48] have extensively investigated spontaneous facial expressions, because spontaneous facial expressions are more relative to true emotion of human

* Corresponding author.
E-mail addresses: xiaohuahwang@gmail.com, xiaohua.huang@ee.oulu.fi (X. Huang), jukortel@ee.oulu.fi (J. Kortelainen), gyzhao@ee.oulu.fi (G. Zhao), lxiaobai@ee.oulu.fi (X. Li), antti.moilanen@ee.oulu.fi (A. Moilanen), tapio@ee.oulu.fi (T. Seppänen), mkp@ee.oulu.fi (M. Pietikäinen).

[1] Equal contributions.

http://dx.doi.org/10.1016/j.cviu.2015.09.015
1077-3142/

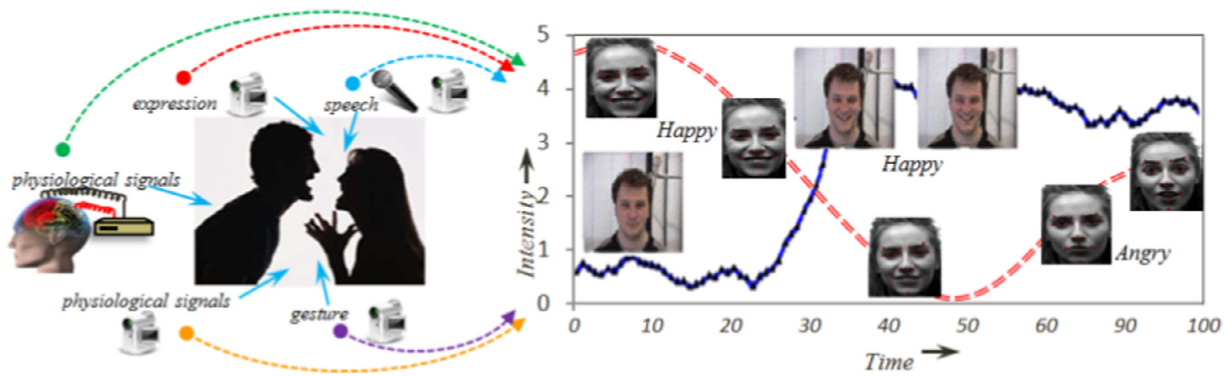

Fig. 1. Multimodal data for recognizing emotions and estimating their intensity.

beings than acted facial expressions. Technically, geometry-based and appearance-based features are two common ways to analyze spontaneous facial expressions [18,26,49]. Specifically, geometric-based feature approach represents the face geometry, such as the shapes and the locations of facial landmarks, which are obtained by an active shape model or active appearance model. On the other hand, appearance-based feature method describes the skin texture of faces, such as wrinkles and furrows [15,30,53]. However, as indicator of emotions, facial expression itself may not provide sufficiently informative characteristics of human beings' affective status [45,52]. The true expression is affected by the context of the social situation, such as different cultures [9]. As a result, one needs to use information from different modalities to increase the accuracy.

Recently, research on physiological signals has been conducted to recognize emotions [22,47], since the physiological signals, such as EEG and electromyogram (EMG), can reveal emotion through physical changes. Kolodyazhniy et al. [22] used the features from peripheral physiological signals to represent neutral, fear, and sadness responses to movie excerpts. In [47], Takahashi et al. collected EEG and peripheral physiological signals from 12 participants and classified their response to emotional videos into five classes: joy, sadness, disgust, fear, and relaxed. While conveying important affective information, EEG signals are difficult to control voluntarily [6]. Moreover, EEG, which reflects the cortical electrical activity, has been proved to provide informative characteristics in responses to the emotion states [29,34,46,54].

In recent years, several researches have made efforts to fuse facial expression and physiological signals. In [3], 93% and 89% accuracies were obtained when using facial expressions for recognizing amusement and sadness, respectively. On the other hand, the accuracy for classifying these emotions with physiological signals (including heart rate, systolic blood pressure, skin conductance level, etc.) was 82%. Combining facial expression and physiological signals improved the accuracies to 94% and 98% for amusement and sadness, respectively. In [6], Chang et al. obtained recognition rate of 90% and 88.33% for facial expression and physiological signals (including skin conductivity, finger temperature and heart rate), respectively, while combining the modalities resulted in a rate of 95%. These results indicate that physiological signals can substantially contribute to multi-modal emotion recognition. In [51], Wesley et al. used the combination of a physiological and visual information channel for user studies, where they used the thermal imaging system to obtain a physiological signal from the face. Furthermore, In [40], Pavlidis et al. applied the work of [51] at a longitudinal human performance study. According to [29,34,46,54], among physiological signals EEG holds relevant information for emotion detection, suggesting it to be a suitable supplement to facial expressions. As far as we know, there are few works involving facial expression with EEG for emotion recognition. In [21], Koelstra et al. used facial expressions together with EEG for emotion classification and implicit affective tagging, but they did not consider arousal and valence classes based on emotion keywords.

1.2. Contribution

In this paper, a new approach for multi-modal emotion recognition is proposed by fusing facial expression and EEG. These modalities are used to classify emotions while users are watching videos with emotional content. The paper's contributions include four parts: (1) emotion recognition from expressions with a new percentage feature; (2) extraction and selection of spectral power and spectral power difference features for EEG; (3) fusion of facial expressions and EEG for valence and arousal recognition on the challenging MAHNOB-HCI database; and (4) a comparison of our approach to human performance for emotion recognition and analysis.

The paper is organized as follows. In Section 2, we briefly introduce the used database, named MAHNOB-HCI database. In Section 3, we present the methods for extracting and fusing facial expression and EEG features. In Section 4, we present the experimental protocol and results of facial expression analysis, EEG classification and multi-modal emotion recognition. Section 5 concludes the paper by giving a short discussion about the results and future work.

2. Database

Different ways of defining expressions and emotions could be used, depending on the problems, *e.g.,* prototypical expressions including happiness, sadness, surprise, fear, disgust and anger, or using two main dimensions: arousal and valence. The dimension of valence ranges from highly positive to highly negative, whereas the dimension of arousal ranges from calming or soothing to exciting or agitating. This two-dimensional model of valence and arousal [43] integrates the discrete emotional labels into a looser continuous structure, which enhances the strict label-based emotion categories and is thought to be an essential representation of affective space. Describing the emotions in the two-dimensional space of valence and arousal is a widely accepted approach which is utilized in the current study.

Even though there are many emotion databases with either facial expressions or physiological signals, only few provide both data supplemented with information about valence and arousal. MAHNOB-HCI [46], used also in this study, is one of these. Even though the database utilized in [24] would also provide similar data, it uses pictures from the IAPS (International Affective Picture System) to induce participants emotions, whereas, in MAHNOB-HCI, affective videos were used for this purpose. In the current study, the latter was preferred as the evoked spontaneous emotions lasted longer and were therefore considered to be more suitable for the combined analysis of facial expression and physiological signals.