Computer Methods and Programs in Biomedicine 140 (2017) 93–110

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Recognition of emotions using multimodal physiological signals and an ensemble deep learning model

Zhong Yin[a,*], Mengyuan Zhao[b], Yongxiong Wang[a,*], Jingdong Yang[a], Jianhua Zhang[c]

[a] Engineering Research Center of Optical Instrument and System, Ministry of Education, Shanghai Key Lab of Modern Optical System, University of Shanghai for Science and Technology, Shanghai, 200093, PR China
[b] School of Social Sciences, University of Shanghai for Science and Technology, Shanghai, 200093, PR China
[c] Department of Automation, East China University of Science and Technology, Shanghai 200237, PR China

ARTICLE INFO

Article history:
Received 20 May 2016
Revised 31 October 2016
Accepted 12 December 2016

Keywords:
Emotion recognition
Affective computing
Physiological signals
Deep learning
Ensemble learning

ABSTRACT

Background and Objective: Using deep-learning methodologies to analyze multimodal physiological signals becomes increasingly attractive for recognizing human emotions. However, the conventional deep emotion classifiers may suffer from the drawback of the lack of the expertise for determining model structure and the oversimplification of combining multimodal feature abstractions.
Methods: In this study, a multiple-fusion-layer based ensemble classifier of stacked autoencoder (MESAE) is proposed for recognizing emotions, in which the deep structure is identified based on a physiological-data-driven approach. Each SAE consists of three hidden layers to filter the unwanted noise in the physiological features and derives the stable feature representations. An additional deep model is used to achieve the SAE ensembles. The physiological features are split into several subsets according to different feature extraction approaches with each subset separately encoded by a SAE. The derived SAE abstractions are combined according to the physiological modality to create six sets of encodings, which are then fed to a three-layer, adjacent-graph-based network for feature fusion. The fused features are used to recognize binary arousal or valence states.
Results: DEAP multimodal database was employed to validate the performance of the MESAE. By comparing with the best existing emotion classifier, the mean of classification rate and F-score improves by 5.26%.
Conclusions: The superiority of the MESAE against the state-of-the-art shallow and deep emotion classifiers has been demonstrated under different sizes of the available physiological instances.

1. Introduction

1.1. Overview

Since collaborations between human and machines (or computers) exist in various working or living environments, researchers in the area of ergonomics and intelligent systems attempt to improve efficiency and flexibility of human-computer interaction (HCI) with high satisfaction levels of human agent [1]. Such intelligent HCI systems require the capability of self-adaptation of computers [2], in which the accurate comprehension of human communications is necessary for machine agent to trigger proper feedback [3]. The human intentions can be expressed in a verbal or a non-verbal manner that carries different *emotions*. A key point of approaching computer adaptability is to develop its functionality of understanding human affective behaviors [4]. This emerging research area is known as affective computing [5–7] regarding the fact that most of the contemporary HCI systems suffer from the lack of intelligence for recognizing emotional cues related to human psychophysiological states [8–10].

Emotions are known as a group of affective states of human being arising as responses to some stimuli from external environments or interpersonal events [11]. Different emotions possess critical influences on self-motivation generation and preferences of decision-making [12]. Representations of emotions include discrete scales in terms of angry, nervous, pleased, bored and so forth or using arousal-valence plane [13–15]. For the latter, 2-dimensional coordinates describe the nature of emotional experience via the core of the affections [16]. The arousal dimension is used to quantify different degrees from calm to excitement levels while the valence dimension indicates whether human feelings are positive (happy) or negative (sad) [17–20]. Fig. 1 shows a typical layout

* Corresponding author.
E-mail addresses: yinzhong@usst.edu.cn, seesawxe@126.com (Z. Yin), wyxiong@usst.edu.cn (Y. Wang).

http://dx.doi.org/10.1016/j.cmpb.2016.12.005
0169-2607/

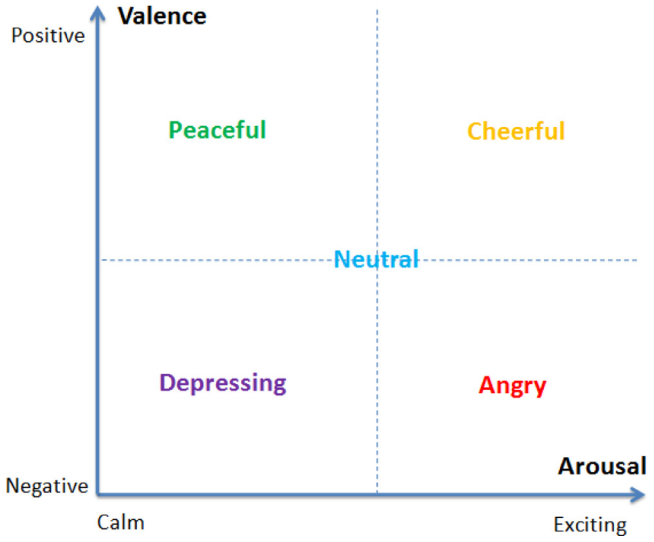

Fig. 1. Arousal-valence plane.

of the arousal-valence plane, where multiple discrete emotional states, e.g., neutral, cheerful, peaceful, depressing and angry, can be defined with different combinations of arousal and valence levels.

1.2. Emotion recognition using physiological signals and pattern classifiers

The function of an intelligent emotion estimator or classifier is to detect emotional clues from human reactions, integrate emotional responses and finally give the prediction of the transient emotional state. The corresponding approaches are mainly classified as two categories, i.e., facial/vocal expressions [21] and physiological signals [22]. Due to specific users who are conditioned to be expressionless, the generalization capability of the behavior data may be limited [23,24]. On the other hand, physiological measures that record electrophysiological information in real time from central nervous system (CNS) or peripheral nervous system (PNS) become attractive because of their the repeatability and objectivity to infer human cognitive or affective state as well as ease of use with a portable implementation via wireless data transmission devices [25].

In well-documented works, the accessibility of various physiological signals for evaluating emotions has been investigated [26–34]. More specifically, emotion variation could be identified via electroencephalogram (EEG) from several frontal and parietal cortical areas. Verma and Tiwary indicated the EEG power spectral density (PSD) in alpha (8–13 Hz) band significantly varies with different valence levels [26]. Frantzidis et al. showed that an EEG feature subset of delta (1–4 Hz) and theta (4–7 Hz) PSD extracted from three central channels (Cz, Fz, and Pz) are quite useful for indicating both arousal and valence levels [28]. The phase synchronization and coherence between EEG channel pairs in the brain areas with functional connectivity were found as effective emotion indicators [29]. The usability of event-correlated potential (ERP) was also examined. Konstantinidis et al. extracted the ERP components of N100 and N200 to classify emotions in arousal-valence plane [30]. Frantzidis et al. calculated P100 and P300 for emotion recognition [29]. In addition, the multimodal PNS physiological signals, e.g., galvanic skin response (GSR) [31], electrooculargram (EOG) [32], electromyogram (EMG) [33], and electrocardiogram (ECG) [34], were extensively explored.

Considering high spatial and temporal resolutions of sophisticated CNS and PNS signal acquisition devices, machine learning approaches facilitate analyzing the massive volume of neurophysiological data [35–39]. In particular, pattern classifiers could fuse physiological features of different modality. Recently, Iacoviello et al. have combined discrete wavelet transformation, principal component analysis (PCA) and support vector machine (SVM) to build a hybrid classification framework [38]. Khezri et al. employed three-channel forehead EEG combined with GSR to recognize six basic emotions via *K*-nearest neighbors (KNN) classifiers [31]. Verma et al. [26] developed an ensemble classification approach fusing EEG, EMG, ECG, GSR, and EOG. Mehmood and Lee used independent component analysis to extract emotional indicators from EEG, EMG, GSR, ECG, and ERP [39].

Due to the superiority of abstracting high-dimensional physiological features, a number of deep learning approaches were investigated for emotion classification and elicit promising results. The popular deep learning primitives include deep belief networks (DBN), stacked autoencoders (SAE) and convolutional neural nets (CNN). In particular, Wand and Shang adopted the standard DBN to extract features from raw physiological data based on unsupervised pre-training and build three deep classifiers to estimate the levels of arousal, valance, and liking [40]. The classification accuracies of DEAP database are 60.9%, 51.2%, and 68.4%, respectively. Similarly, Li et al. adopted a two-layer DBN ensemble to fuse multichannel EEG data in DEAP and the binary emotion classification accuracies of arousal and valence scales are 0.5840, and 0.6420, respectively [41]. Li et al. employed the supervised restrict Boltzmann machine (RBM) to modify the standard DBN and proposed the supervised DBN based affective state recognition (SDA) model [42]. By using the EEG data of DEAP as the deep model inputs, the average AUC (i.e., the area under the receiver operating characteristic curve) score is 0.75. Jia et al. proposed the semi-supervised deep learning model (semi-DLM) based on DBNs for binary emotion classification [43]. The essential of the semi-DLM classifier is to utilize the label information for EEG channel selections instead of the pre-training procedure of the DBNs. The average AUC score of the liking scale in DEAP is 0.7890. Jirayucharoensak et al. combined the dimensionality reduction technique, i.e., PCA, with the standard SAE network to build emotion classifiers [44]. The designed SAE network possessed two hidden layers with 100 hidden neurons in each layer. Based on three levels of arousal and valence scales targeted in DEAP, the average classification accuracies are 0.4952 and 0.4603, respectively. Besides the physiological signals, Acar et al. combined the CNN and SVM to identify four affective categories from the audio and visual modality of videos [45].

1.3. Motivation of the present study

The brief literature review suggests the machine-learning-based methodologies are promising to reveal the latent patterns of certain emotional states hidden in the physiological signals. In particular, the deep classifiers are able to abstract the intermediate representations of physiological features in multiple modalities via hierarchical architectures. However, the deep network structure of emotion classifiers is usually selected based on the prior knowledge from other domains. Considering the nature of the high dimensionality and limited training instances of the physiological data, transferring the empirical expertise from massive data problems may not be always reliable. More specifically, too deep network with too many hidden neurons in each layer may lead to the severe model overfitting. The oversimplification and insufficient abstraction of physiological features could arise when employing too simple model structure. Hence, it is necessary to develop a physiological-data-driven approach to identify the optimal topology of the deep emotion classifier.

On the other hand, the classifier ensemble has the capability to tackle the multimodality in physiological signals since it improves