Neural Networks 63 (2015) 104–116

Contents lists available at ScienceDirect

Neural Networks

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

Towards an intelligent framework for multimodal affective data analysis

Soujanya Poria[a], Erik Cambria[b,*], Amir Hussain[a], Guang-Bin Huang[c]

[a] School of Natural Sciences, University of Stirling, UK

[b] School of Computer Engineering, Nanyang Technological University, Singapore

[c] School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 31 August 2013
Received in revised form 19 September 2014
Accepted 9 October 2014
Available online 6 November 2014

Keywords:
Multimodal
Multimodal sentiment analysis
Facial expressions
Speech
Text
Emotion analysis
Affective computing

ABSTRACT

An increasingly large amount of multimodal content is posted on social media websites such as YouTube and Facebook everyday. In order to cope with the growth of such so much multimodal data, there is an urgent need to develop an intelligent multi-modal analysis framework that can effectively extract information from multiple modalities. In this paper, we propose a novel multimodal information extraction agent, which infers and aggregates the semantic and affective information associated with user-generated multimodal data in contexts such as e-learning, e-health, automatic video content tagging and human–computer interaction. In particular, the developed intelligent agent adopts an ensemble feature extraction approach by exploiting the joint use of tri-modal (text, audio and video) features to enhance the multimodal information extraction process. In preliminary experiments using the eNTERFACE dataset, our proposed multi-modal system is shown to achieve an accuracy of 87.95%, outperforming the best state-of-the-art system by more than 10%, or in relative terms, a 56% reduction in error rate.

1. Introduction

Emotions play a crucial role in our daily lives. They aid decision-making, learning, communication, and situation awareness in human-centric environments (Howard & Cambria, 2013). In the past two decades, artificial intelligence (AI) researchers have been attempting to endow machines with capacities to recognize, interpret and express emotions. All such efforts can be attributed to affective computing (Picard, 1997), a new interdisciplinary research field that spans computer sciences, psychology and cognitive science.

Emotion and sentiment analysis have become a new trend in social media, helping users to understand the opinion being expressed on products. With the advancement of technology and the rapid rise of social media, along with the large amount of opinions that are expressed in textual format, there is a growing number of opinions posted in video format. Consumers tend to record their opinions on products in front of a web camera or other devices and upload them on social media like YouTube or Facebook. This is to let other people know about the products before they buy. These videos often contain comparisons of the products with products from competing brands, the pros and cons of the product, etc. All of this information is useful for people who wish to purchase the product. The main advantage of analyzing videos rather than textual analysis to detect emotions from opinions is that more cues are available in videos. Textual analysis facilities only the use of words, phrases and relations, dependencies among them which are not sufficient to understand opinions and extract associated emotion from the opinions. Video opinions provide multimodal data in terms of vocal and visual modality. The vocal modulations of the opinions and facial expressions in the visual data along with text data provide important cues to identify emotion. Thus, a combination of text and video data can help create a better emotion analysis model.

The growing amount of research conducted in this field, combined with advances in signal processing and AI, has led to the development of advanced intelligent systems that aim to detect and process affective information contained in multi-modal sources. The majority of such state-of-the-art frameworks however, rely on processing a single modality, i.e. text, audio, or video. Furthermore, all of these systems are known to exhibit limitations in terms of meeting robustness, accuracy and overall performance requirements, which in turn greatly restrict the usefulness of such systems in real-world applications.

[*] Corresponding author.
E-mail addresses: soujanya.poria@cs.stir.ac.uk (S. Poria), cambria@ntu.edu.sg (E. Cambria), ahu@cs.stir.ac.uk (A. Hussain), egbhuang@ntu.edu.sg (G.-B. Huang).

http://dx.doi.org/10.1016/j.neunet.2014.10.005
0893-6080/© 2014 Elsevier Ltd. All rights reserved.

The aim of multi-sensor data fusion is to increase the accuracy and reliability of estimates (Qi & Wang, 2001). Many applications, e.g. navigation tools, have already demonstrated the potential of data fusion. This implies the importance and feasibility of developing a multi-modal framework that could cope with all three sensing modalities – text, audio, and video – in human-centric environments. The way humans communicate and express their emotions is known to be multimodal. The textual, audio and visual modalities are concurrently and cognitively exploited to enable effective extraction of the semantic and affective information conveyed during communication. In this work, we show that the ensemble application of feature extraction from different types of data and modalities enhances the performance of our proposed multi-modal sentiment and emotion recognition system.

Specifically, we employ the supervised learning paradigm. For training, we used three datasets corresponding to the three modalities: the ISEAR dataset (Bazzanella, 2004) to build a model for emotion detection from text, the CK++ dataset (Lucey et al., 2010) to construct a model for emotion detection from facial expressions, and the eNTERFACE dataset (Martin, Kotsia, Macq, & Pitas, 2006) to build a model for emotion extraction from audio, as well to evaluate the trained models for the other two modalities.

For training the three models, we used a novel process of feature extraction from the datasets of the corresponding modalities. The information coming from the three modalities was then fused by concatenating the feature vectors of each modality. These combined feature vectors were fed into a supervised classifier to produce the final output. Several classifiers were experimented, with their performance evaluated through tenfold cross-validation. The support vector machine (SVM) classifier was found to outperform the best known state-of-the-art system by more than 10%, which in relative figures equates to a nearly 60% reduction of the error rate.

The rest of the paper is organized as follows: in Section 2 we discuss related work on multimodal fusion; in Section 3 we give detailed descriptions of the datasets used; in Sections 5–7 we explain how we processed textual, audio and visual data, respectively; Section 8 illustrates the methodology adopted for fusing different modalities; Section 9 presents the experimental results; Section 10 presents the process of developing a real-time multimodal emotion analysis system. Section 11 outlines conclusions and some future work recommendations.

2. Related work

Both feature extraction and feature fusion are crucial for a multimodal emotion analysis system. Existing works on multimodal emotion analysis can be categorized into two broad categories: those devoted to feature extraction from each individual modality, and those developing techniques for the fusion of the features coming from different modalities.

2.1. Video: recognition of facial expression

In 1970, Ekman (1970) carried out extensive studies on facial expressions. Their research showed that universal facial expressions provide sufficient clues to detect emotions. They used anger, sadness, surprise, fear, disgust and joy as six basic emotion classes. Such basic affective categories are sufficient to describe most of the emotions exhibited through facial expressions. However, this list does not include the emotion a person facially expresses when he or she shows disrespect to someone; thus a seventh basic emotion, contempt, was introduced by Matsumoto (1992).

Ekman and Friesen (1978) developed a facial expression coding system (FACS) to code facial expressions by deconstructing a facial expression into a set of action units (AU). AUs are defined via specific facial muscle movements. An AU consists of three basic parts: AU number, FACS name, and muscular basis. For example, for AU number 1, the FACS name is *inner brow raiser* and it is explicated via *frontalis, pars medialis* muscle movements. In application to emotions, Friesen and Ekman (1983) proposed the emotional facial action coding system (EFACS). EFACS defines the sets of AUs that participate in the construction of facial expressions expressing specific emotions.

The Active Appearance Model (Datcu & Rothkrantz, 2008; Lanitis, Taylor, & Cootes, 1995) and Optical Flow-based techniques (Mase, 1991) are common approaches that use FACS to understand expressed facial expressions. Exploiting AUs as features, *k*NN, Bayesian networks, hidden Markov models (HMM) and artificial neural networks (ANN) (Ueki, Morishima, Yamada, & Harashima, 1994) have been used by many researchers to infer emotions from facial expressions. The performance of several machine-learning algorithms for detecting emotions from facial expressions is presented in Table 1 (Chen, 2000). All such systems, however, use different, manually crafted corpora, which makes it impossible to perform a comparative evaluation of their performance.

2.2. Audio: emotion recognition from speech

Recent studies on speech-based emotion analysis (Chiu, Chang, & Lai, 1994; Cowie & Douglas-Cowie, 1996; Datcu & Rothkrantz, 2008; Dellaert, Polzin, & Waibel, 1996; Johnstone, 1996; Murray & Arnott, 1993; Sato & Morishima, 1996; Scherer, 1996) have focused on identifying several acoustic features such as fundamental frequency (pitch), intensity of utterance (Chen, 2000), bandwidth, and duration. The speaker-dependent approach gives much better results than the speaker-independent approach, as shown by the excellent results of Navas and Hernez (2006), where about 98% accuracy was achieved by using the Gaussian mixture model (GMM) as a classifier, with prosodic, voice quality as well as Mel frequency cepstral coefficient (MFCC) employed as speech features.

However, the speaker-dependent approach is not feasible in many applications that deal with a very large number of possible users (speakers). To our knowledge, for speaker-independent applications, the best classification accuracy achieved so far is 81% (Atassi & Esposito, 2008), obtained on the Berlin Database of Emotional Speech (BDES) (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005) using a two-step classification approach and a unique set of spectral, prosodic, and voice features, selected through the Sequential Floating Forward Selection (SFFS) algorithm (Pudil, Ferri, Novovicova, & Kittler, 1994).

Chiu et al. (1994) extracted five prosodic features from speech and used multilayered ANNs to classify emotions. As per the analysis of Scherer (1996), the human ability to recognize emotions from speech audio is about 60%. Their study shows that sadness and anger are detected more easily from speech, while the recognition of joy and fear is less reliable. Caridakis et al. (2007) obtained 93.30% and 76.67% accuracy to identify anger and sadness, respectively, from speech, using 377 features based on intensity, pitch, Mel-Scale Frequency Cepstral Coefficients (MFCC), Bark spectral bands, voiced segment characteristics, and pause length.

2.3. Text: affect recognition from textual data

Affective content recognition in text is a rapidly developing area of natural language processing, which has received growing attention from both the research community and industry in recent years. Sentiment and emotion analysis tool said companies to, for example, become informed about what customers feel in relation to their products, or help political parties to get to know how voters feel about their actions and proposals.