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/

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.

Table 1
Performance of various learning algorithms for detecting emotions from facial images.

Method	Processing	Classification algorithm	Accuracy
Lanitis et al. (1995)	Appearance Model	Distance-based	74%
Cohen et al. (2003)	Appearance Model	Bayesian network	83%
Mase (1991)	Optical flow	*k*NN	86%
Rosenblum et al. (1996)	Optical flow	ANN	88%
Otsuka and Ohya (1997)	2D FT of optical flow	HMM	93%
Yacoob and Davis (1996)	Optical flow	Rule-based	95%
Essa and Pentland (1997)	Optical flow	Distance-based	98%

A number of works have aimed to identify positive, negative, or neutral sentiment associated with words (Arora, Bakliwal, & Varma, 2012; Turney, 2002; Wawer, 2012; Wiebe, 2010), phrases (Wilson, Wiebe, & Hoffmann, 2005), sentences (Riloff & Wiebe, 2003; Strapparava & Mihalcea, 2007), and documents (Maas et al., 2011; Pang & Lee, 2004). The task of automatically identifying fine grained emotions, such as anger, joy, surprise, fear, disgust, and sadness, explicitly or implicitly expressed in a text, has been addressed by several researchers (Alm, Roth, & Sproat, 2005b; Mishne, 2005; Strapparava & Mihalcea, 2008; Strapparava & Valitutti, 2004). So far, approaches to text-based emotion and sentiment detection rely mainly on rule-based techniques, bag of words modeling using a large sentiment or emotion lexicon (Poria, Gelbukh, Hussain, Das, & Bandyopadhyay, 2013), or statistical approaches that assume the availability of a large dataset annotated with polarity or emotion labels (Xia, Zong, Hu, & Cambria, 2013).

Several supervised and unsupervised classifiers have been built to recognize emotional content in texts (Chaumartin, 2007; Lin, Yang, & Chen, 2007). The SNoW architecture (Alm, Roth, & Sproat, 2005a) is one of the most useful frameworks for text-based emotion detection. In the last decade, researchers have been focusing on emotion extraction from texts of different genres such as news (Lin et al., 2007), blogs (Melville, Gryc, & Lawrence, 2009), Twitter messages (Pak & Paroubek, 2010; Sidorov, Miranda-Jiménez et al., 2013), and customer reviews (Hu & Liu, 2004). Emotion extraction from social media content helps to predict the popularity of a product release or the results of an election poll, etc. To this end, several knowledge-based sentiment (Esuli & Sebastiani, 2006) and emotion (Balahur, Hermida, & Montoyo, 2012) lexicons have been developed for word- and phrase-level sentiment and emotion analysis, e.g., WordNet-Affect (WNA) (Pang & Lee, 2004), a dictionary of affective words, and SenticNet (Cambria, Olsher, & Rajagopal, 2014), a publicly available semantic resource for concept-level sentiment analysis.

2.4. Multimodal fusion

The ability to perform multimodal fusion is an important prerequisite to the successful implementation of agent–user interaction. One of the primary obstacles to multimodal fusion is the development and specification of a methodology to integrate cognitive and affective information from different sources on different time scales and measurement values. There are two main fusion strategies; feature-level fusion and decision-level fusion.

Feature-level fusion (Kapoor, Burleson, & Picard, 2007; Pun, Alecu, Chanel, Kronegg, & Voloshynovskiy, 2006; Shan, Gong, & McOwan, 2007) combines the characteristics extracted from each input channel in a "joint vector" before any classification operations are performed. Some variations of such an approach exist, e.g. Mansoorizadeh and Charkari (2010) proposed asynchronous feature-level fusion. Modality fusion at feature-level presents the problem of integrating highly disparate input features, suggesting that the problem of synchronizing multiple inputs while re-teaching the modality's classification system is a nontrivial task.

In decision-level fusion, each modality is modeled and classified independently. The unimodal results are combined at the end of the process by choosing suitable metrics such as expert rules and simple operators including majority votes, sums, products, and statistical weighting. A number of studies favor decision-level fusion as the preferred method of data fusion because errors from different classifiers tend to be uncorrelated and the methodology is feature-independent (Kuncheva, 2004). Bimodal fusion methods have been proposed in numerous instances (Datcu & Rothkrantz, 2008; Gunes & Piccardi, 2007; Zeng et al., 2007), but optimal information fusion configurations remain elusive.

Cambria, Howard, Hsu, and Hussain (2013) proposed a novel approach called Sentic Blending to fuse the modalities in order to grasp emotion associated with the multimodal content. Unlike other approaches, they fused facial expressions with natural language text. They also tracked the sentiment change over time. As datasets for the experiment, they used FGNET and MMI datasets.

Paleari and Huet (2008) carried out both decision and feature-level fusion. They experimented with the eNTERFACE dataset and showed that decision-level fusion outperformed feature-level fusion. Many multimodal methodologies have ad-hoc workarounds for the purpose of fusing information from multiple modalities, but the entire system must be retrained before new modalities can be included. Also, they are not as adaptive to quality changes in input, so do not perform long-term adjustments to better adapt to data trends.

3. Datasets employed

Our goal is to identify affective contents associated with multimodal content. In this section, we describe the various datasets used in our experiment as resources for extracting features for the three modalities.

3.1. The ISEAR dataset

As a source of various features and similarity measures between concepts, we used the International Survey of Emotion Antecedents and Reactions (ISEAR)[1] dataset (Scherer, 1996). The survey was conducted in the 1990s across 37 countries and had approximately 3000 respondents.

The respondents were instructed to describe a situation or event in which they felt a particular emotion, in the form of a *statement*—a short text of a couple of sentences (2.37 on average). Here is an example of a complete statement:

> *I had the window open and the door was shut so that the kitten would not go out. My partner came in and started talking about something and I forgot about the window and suddenly I saw the kitten hanging from the window frame. I was rigid with fright till I got hold of her.*

[1] http://www.affective-sciences.org/system/files/page/2636/ISEAR.zip, downloaded on July 14, 2012. Linked from http://www.affective-sciences.org/researchmaterial.

The choice of ISEAR as the source of corpus-based information is motivated by the fact that this corpus is particularly rich in emotion-related words, as compared to more standard corpora used in natural language processing. In the sample statement cited above, the concepts *window open*, *forget*, *suddenly*, *hang*, *rigid with fright* are all associated with the same emotion; fear. This property makes the ISEAR database particularly suitable for learning co-occurrence-based emotion similarity measures between concepts. In this work, we used ISEAR dataset as an emotion annotated corpus to build the training model for textual emotion analysis. Several features were extracted from the ISEAR corpus based on WordNet-Affect (WNA) lists (Strapparava & Valitutti, 2004) and SenticNet (Cambria et al., 2014) in order to build the model of textual data.

The dataset contains 7666 such statements, which include 18,146 sentences and 449,060 running words. Each statement is associated with the emotion felt in the situation, which takes one of the seven values: anger, disgust, fear, guilt, joy, sadness, and shame. For example, the statement cited above is labeled as fear. This set of seven emotions is different from our target set of Ekman's six basic emotions: anger, disgust, fear, joy, sadness, and surprise (see Ortony & Terence, 1990 for a comprehensive overview of different sets of basic emotions proposed in the literature). We removed this dissimilarity in the labels of these two datasets by ignoring the statements having guilt and shame as emotion labels in the ISEAR dataset. However, the ISEAR dataset does not contain any statement under surprise as an emotion category. To solve this issue and obtain the training dataset for surprise, we used a dataset produced by SemEval 2007-Task organizers. The dataset consists of newspaper headlines annotated according to Ekman's six basic emotion classes with neutral as an extra emotion. We only considered those sentences of the dataset having the surprise emotion. The dataset contains 634 sentences, which are labeled as surprise and are used in our experiment.

3.2. The CK++ dataset

To build the model for emotion recognition from facial expressions, we used CK++ (Lucey et al., 2010), a comprehensive dataset that consists of images of the facial behavior of 210 adults. The image sequences were recorded using two hardware-synchronized Panasonic AG-7500 cameras. The participants were 18–50 years old, 81% Euro-Americans, 13% Afro-Americans, and 6% from other ethnic groups; 69% were females. The experimenter asked the participants to perform a series of 23 facial displays, which included single AU or combination of AUs (Mase, 1991). The image sequences of frontal views and 30° views were digitized into 640 × 490 or 640 × 480-pixel arrays with 8-bit grayscale or 24-bit color values. The sequence of the facial images of each of the subjects was manually annotated with one of the six emotion categories, the same as in WNA and which we used in our study. CK++ dataset contains 593 facial image sequences, but only 327 of them have specific emotion labels. Detailed distribution of the data samples per emotion is shown in Table 2.

3.3. The eNTERFACE dataset

The eNTERFACE (Martin et al., 2006) database was recorded using a min-DIV digital video camera. 42 subjects of 14 nationalities were asked to listen to six successive short stories, each of them eliciting a particular emotion (Ekman's six basic emotions were used). They were instructed by the experimenter to react to each of the six situations (stories). Two human experts were judging the subjects' reactions as to whether the subjects expressed an emotion unambiguously through their reactions to the stories. Here is an example of a story which elicits anger:

Table 2
Distribution of data samples per each emotion label in CK++ dataset.

Expression	#Samples
Neutral	18
Anger	45
Joy	69
Disgust	59
Surprise	83
Fear	25
Sadness	28
Total:	327

You are in a foreign city. A city that contains only one bank, which is open today until 4 pm. You need to get 200$ from the bank, in order to buy a flight ticket to go home. You absolutely need your money today. There is no ATM cash machine and you don't know anyone else in the city. You arrive at the bank at 3 pm and see a big queue. After 45 min of queuing, when you finally arrive at the counter, the employee tells you to come back the day after because he wants to have a coffee before leaving the bank. You tell him that you need the money today and that the bank should be open for 15 more minutes, but he is just repeating that he does not care about anything else but his coffee . . . .

Different subjects' reactions after listening to the above story have been:

- *What??? No, no, no, listen! I need this money!*
- *I don't care about your coffee! Please serve me!*
- *I can have you fired you know!*
- *Is your coffee more important than my money?*
- *You're getting paid to work, not drink coffee!*

Each of the reactions expresses *anger* as emotion according to the eNTERFACE dataset.

Since all video clips in this dataset are annotated according to Ekman's emotion taxonomy, we treated this dataset as the gold standard data for all three (visual, text, and speech) modalities. We also used this dataset as a source of speech data to build the training model for speech-based emotion analysis.

3.4. Knowledge bases used/developed

In the analysis of textual data, information related to the language and the properties of individual words of concepts was used. Specifically, we used the following lexical resources.

The SenticNet dataset: As an a priori polarity lexicon of concepts, we used SenticNet 3.0 (Cambria et al., 2014), a lexical resource that contains 30,000 concepts along with their polarity scores in the range from −1.0 to +1.0. Specifically, we employed the beta version of SenticNet 3.0.[2] It contains 13,741 concepts,[3] of which 7626 are multi-word expressions, e.g., *prevent pregnancy*, *high pay job*, *feel happy*. Of the concepts in SenticNet, 6452 are found in WordNet 3.0 and 7289 are not. Of the latter, most are multi-word concepts such as *access internet* or *make mistake*, except for 82 single-word concepts, such as *against* or *telemarketer*.

The first 20 SenticNet concepts in lexicographic order along with the corresponding polarities are shown in Table 3.

ConceptNet: ConceptNet (Speer & Havasi, 2012) represents the information from the Open Mind corpus as a directed graph, in

[2] http://sentic.net/senticnet-3.0.zip, downloaded on May 14, 2014.

[3] SenticNet3.0 is currently under development; it will contain 30,000 concepts. Applying our method to this new version will automatically result in a resource of the corresponding size.

Table 3
A sample of SenticNet data.

a lot	+0.258	*abhorrent*	−0.443
a lot sex	+0.858	*able read*	+0.865
a little	+0.032	*able run*	+0.775
Abandon	−0.566	*able use*	+0.856
Abase	−0.153	*abominably*	−0.443
Abash	−0.174	*abominate*	−0.391
Abashed	−0.174	*abomination*	−0.391
Abashment	−0.186	*abortion*	−0.27
Abhor	−0.391	*abroad*	+0.255
Abhorrence	−0.391	*absolute*	+0.277

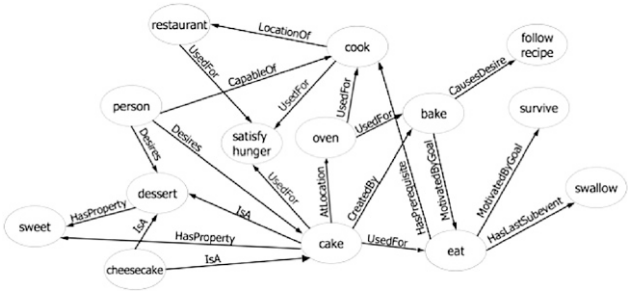

Fig. 1. A sketch of ConceptNet graph.

which the nodes are concepts and the labeled edges are commonsense assertions that interconnect them. For example, given the two concepts *person* and *cook*, an assertion between them is CapableOf, i.e., a *person* is *capable of cooking*; see Fig. 1 (Speer & Havasi, 2012).

EmoSenticNet: The EmoSenticNet dataset (Poria et al., 2013) contains about 5700 common-sense knowledge concepts, including those concepts that exist in the WNA list, along with their affective labels in the set {anger, joy, disgust, sadness, surprise, fear}.

EmoSenticSpace: In order to build a suitable knowledge base for emotive reasoning, we applied the so-called "blending" technique to ConceptNet and EmoSenticNet. Blending is a technique that performs inference over multiple sources of data simultaneously, taking advantage of the overlap between them (Havasi, Speer, & Pustejovsky, 2009). Basically, it linearly combines two sparse matrices into a single matrix, in which the information between the two initial sources is shared.

Before performing blending, we represented EmoSenticNet as a directed graph similar to ConceptNet. For example, the concept *birthday party* is assigned the emotion *joy*. We took them as two nodes and added the assertion HasProperty on the edge directed from the node *birthday party* to the node *joy*.

Then, we converted the graphs to sparse matrices in order to blend them. After blending the two matrices, we performed Truncated Singular Value Decomposition (TSVD) on the resulting matrix to discard those components representing relatively small variations in the data. We discarded all of them keeping only 100 components of the blended matrix to obtain a good approximation of the original matrix. The resulting 100-dimensional space was clustered by means of sentic medoids (Cambria et al., 2011).

4. Overview of the proposed method

We classified video clips that contained information in three modalities: visual information, sound track (speech), and captions (text). To achieve reliable affective information extraction from multimodal data, we fused the results on different modalities in order to involve all modalities in the emotion analysis process. Our algorithm proceeded as follows.

Preprocessing: Data for each modality were processed.

Table 4
Datasets and the classifier for each modality.

Modality	Training set	Test set	Best classifier
Video	CK++	eNTERFACE	ELM
Audio	ISEAR	eNTERFACE	SVM
Text	eNTERFACE	eNTERFACE	SVM

Feature extraction: Features for building training models were extracted from the datasets for each modality. For visual data, the feature extraction process includes a classification step, as explained in Section 5; this step includes its own training.

Fusion: Outputs of the classifiers for all modalities were fused using our feature-based fusion technique.

Training: Using these features, a multimodal model was built and evaluated. For comparison, a model was also built and evaluated for each modality separately.

As training data, we used the CK++ dataset for the visual modality, the ISEAR dataset for the textual modality, and the eNTERFACE dataset for the audio modality (speech). As testing data for all three modalities, we used the eNTERFACE dataset. We evaluated various supervised classifiers for each modality: for textual and speech modality, the best accuracy was achieved by using SVM (Cortes & Vapnik, 1995); and for visual modality, by means of the extreme learning machine (ELM) (Huang, Zhu, & Siew, 2006); see Table 4.

In the next four sections we describe each step in detail, and then show that our proposed technique outperforms the methods that use single modalities.

5. Use of visual data for emotion recognition

Humans are known to express emotions through the face to a great extent. Facial expressions play a significant role in the identification of emotions in a multimodal stream. A facial expression analyzer automatically identifies emotional clues associated with facial expressions and classifies facial expressions in order to define emotion categories and to discriminate between them. We used Ekman's six emotion classes along with an extra emotion category, *neutral*, as target classes for the emotion classification problem.

Our method of feature extraction for visual modality of the video clips requires previous classification of still images, as explained in Section 5.3.

5.1. Still images: data preparation

We used CK++ and eNTERFACE datasets to train and evaluate our facial expression analyzer. The CK++ dataset contains, for each subject, a sequence of n facial images expressing a particular emotion, from time T_0 to T_n. At time T_0 the subject starts to express the emotion in front of the camera, and expresses this emotion till time T_n. The first few images of the sequence correspond to a neutral expression, and the rest to the expression of a particular emotion. We manually separated the images in each sequence into two categories: those expressing a neutral emotion and those expressing a given emotion, as shown in Fig. 2.

Since our classifier worked with individual images, not with sequences, we considered the sequences as sets of individual images. These individual images, with their assigned categories – either neutral or one of the six emotions – formed our dataset. For example, the sequence in Fig. 2 contributed to the dataset with three images labeled as *neutral* and four labeled as *surprise*. In total, the resulting dataset contained 5877 facial images corresponding to the 7 emotions (including *neutral*), see Table 5.

Fig. 2. Labeling facial images in the sequence as neutral or carrying a specific emotion.

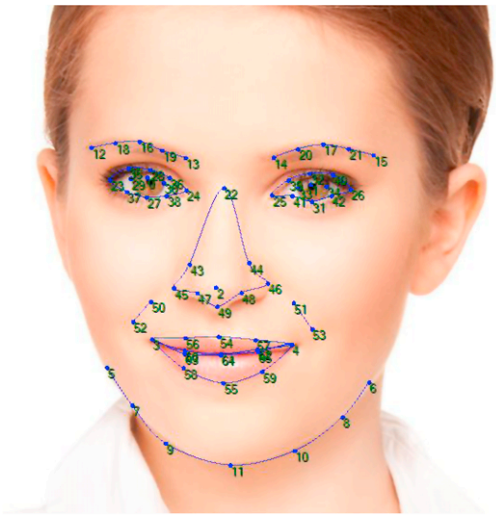

Fig. 3. Facial characteristic points of a facial image ad detected by Luxand software.

Table 5
Distribution of data samples per emotion in the final dataset.

Emotion	Number of samples
Neutral	233
Anger	1022
Joy	1331
Disgust	868
Surprise	1329
Fear	546
Sadness	548

5.2. Still images: feature extraction

To extract facial characteristic points (FCPs) from the facial images, we used the face recognition software Luxand FSDK 1.7[4]. From each image we extracted 66 FCPs as shown in Fig. 3; Table 6 lists important examples. The FCPs were used to construct facial features, which were defined as distances between FCPs; see examples in Table 7. There were, thus, a total of $\binom{66}{2} = 2145$ features per image.

5.3. Unimodal classification of still facial images

With the features just described, we trained a classifier for two-way classification of still images into those that express no emotion (neutral category) and those expressing some emotion. This classifier was used as the first step in our two-step classification procedure for emotion-based classification of images as described below, as well as for feature extraction from video clips, as described in the next section.

Table 6
Some relevant facial characteristic points (out of the 66 facial characteristic points detected by Luxand).

Facial point	Description
0	Left eye
1	Right eye
24	Left eye inner corner
23	Left eye outer corner
38	Left eye lower line
35	Left eye upper line
29	Left eye left iris corner
30	Left eye right iris corner
25	Right eye inner corner
26	Right eye outer corner
41	Right eye lower line
40	Right eye upper line
33	Right eye left iris corner
34	Right eye right iris corner
13	Left eyebrow inner corner
16	Left eyebrow middle
12	Left eyebrow outer corner
14	Right eyebrow inner corner
17	Right eyebrow middle
54	Mouth top
55	Mouth bottom

Note that complete 7-way classification of still images by emotions is not a part of our multimodal method and was performed only for comparison. To classify facial images by emotion, we designed a two-step classifier: First we used our two-way classifier to decide whether the image expressed no emotion (neutral) or some emotion. In the latter case, a 6-way classification was then carried out to identify the specific emotion category of the image.

Both classification steps used the same feature set. Of various supervised classifiers that we experimented with, ELM gave the best results. The two-stage classification process enhanced the accuracy of unimodal classification: on the CK++ dataset using the ELM classifier, one-stage 7-way classification gave 80.48% accuracy, while our two-stage procedure gave 86.47%. To estimate the accuracy, we used ten-fold cross validation.

5.4. Video clips (visual modality): feature extraction for multimodal fusion

To build a feature vector of a video clip showing the human face using its visual modality, we first burst the clip into a set of individual frames. Next, we extracted the features from these individual frames as described in Section 5.2, and subsequently classified these images into those expressing no emotion (neutral) and those expressing some emotion, as described in Section 5.3. We discarded the frames classified as showing no emotion, and used for the next step only those showing some emotion. Finally, we built the feature vector for the video clip using coordinate-wise

[4] http://www.luxand.com.

Table 7
Some important facial features used for the experiment.

Feature	Distance measure
Distance between right eye and left eye	D(0, 1)
Distance between the inner and outer corner of the left eye	D(23, 24)
Distance between the upper and lower line of the left eye	D(35, 38)
Distance between the left iris corner and right iris corner of the left eye	D(29, 30)
Distance between the inner and outer corner of the right eye	D(25, 26)
Distance between the upper and lower line of the right eye	D(40, 41)
Distance between the left iris corner and right iris corner of the right eye	D(33, 34)
Distance between the left eyebrow inner and outer corner	D(12, 13)
Distance between the right eyebrow inner and outer corner	D(14, 15)
Distance between top of the mouth and bottom of the mouth	D(54, 55)

averaging of the feature vectors of individual frames:

$$x_i = \frac{1}{N}\sum_{i=1}^{N} x_{ij},$$

where x_i is the ith coordinate of the video clip's feature vector, x_{ij} is the ith coordinate of its jth frame's vector, and N is the number of frames in the video clip; as stated earlier, only frames that were classified as having some emotion are considered.

5.5. Classification of video clips (visual modality)

Similar to the case for still images, classification of video clips is not a part of our multimodal method and was performed only for comparison.

In order to classify video clips (ignoring the sound track and captions), we burst the videos from the eNTERFACE dataset into image frames, then applied our two-stage classifier to individual frames of the sequence, and finally used majority voting on the emotion labels of all the video frames to determine the prevailing emotion of the video.

6. Use of audio (speech) for emotion recognition

For emotion recognition from speech we used eNTERFACE as both the training and testing dataset. First, the audio signal was extracted from video files in the dataset. The signal had a bit-rate of 1536 kbps and a frequency of 48 kHz. Then we extracted relevant features from the audio signal. To extract all audio features, we used the JAudio toolkit (McKay, Ichiro, & Philippe, 2005), which is a music feature extraction toolkit written in Java. There are two broad kinds of audio features: short- and long-time based features. Below we briefly describe each of these features in turn.

6.1. Short time-based features

Short time-based features are mainly used to distinguish the timbral characteristics of the signal and are usually extracted from every short-time window (or frame), during which the audio signal is assumed to be stationary—see Tzanetakis (2002) for more details on these features.

Mel-frequency cepstral coefficients (MFCC) are calculated based on short time Fourier transform (STFT). First, log-amplitude of the magnitude spectrum is taken, and the process is followed by grouping and smoothing the fast Fourier transform (FFT) bins according to the perceptually motivated Mel-frequency scaling. The JAudio tool gives the first five of 13 coefficients, which produce the best classification result.

Spectral centroid is the center of gravity of the magnitude spectrum of the STFT. Here, M_i [n] denotes the magnitude of the Fourier transform at frequency bin n and frame i. The centroid is used to measure the spectral shape. A higher value of the centroid indicates brighter textures with greater frequency. The spectral centroid is calculated as

$$C_i = \frac{\sum_{i=1}^{N} nM_i[n]}{\sum_{i=1}^{N} M_i[n]}.$$

Spectral rolloff is the feature defined by the frequency R_t such that 85% of the frequency is below this point:

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 \sum_{i=1}^{N} M_t[n].$$

Spectral flux is defined as the squared difference between the normalized magnitudes of successive windows:

$$F_t = \sum_{n=1}^{N} (N_t[n] = -N_{t-1}[n])^2,$$

where N_t [n] and N_{t-1} [n] are the normalized magnitudes of the Fourier transform at the current frame t and the previous frame $t-1$, respectively. The spectral flux represents the amount of local spectral change.

Root mean square (RMS) is calculated for each window. Suppose x_i is the energy of each sample and N is the total number of samples. Then RMS is defined as

$$\text{RMS} = \sqrt{\frac{\sum_{i=1}^{N} M_i^2}{N}}.$$

Compactness is calculated as the sum over frequency bins of an FFT. It is a measure of noisiness of the signal.

Time domain zero crossing is a timbral feature that is also used as a measure of noisiness of the signal.

6.2. Long time-based features

Long-term features can be generated by aggregating the short-term features extracted from several consecutive frames within a time window. We have used derivate, standard deviation, running mean, derivative of running mean, and standard deviation of running mean as the aggregation methods of short time-based features listed in Section 6.1.

To find the human perceptible pattern for the signal we extracted three main semantic features: beat histogram feature, beat sum, and strongest beat in the audio signal.

Beat histogram is a histogram showing the relative strength of different rhythmic periodicities in a signal. It is calculated as the auto-correlation of the RMS.

Beat sum is measured as the sum of all entries in the beat histogram. It is a very good measure of the importance of regular beats in a signal.

Strongest beat is defined as the strongest beat in a signal, in beats per minute and it is found by finding the strongest bin in the beat histogram.

7. Text-based emotion recognition

Identifying emotions in text is a challenging task, because of ambiguity of words in the text, complexity of meaning and interplay of various factors such as irony, politeness, writing style, as well as variability of language from person to person and from culture to culture. In this work, we followed the sentic computing paradigm developed by Cambria and his collaborators, which considers the text as expressing both semantics and sentics (Cambria, Hussain, Havasi, & Eckl, 2009, 2010a, 2010b; Poria, Cambria, Winterstein, & Huang, 2014). We used a novel approach for identifying the emotions in text by extracting the following key features using our new resource, EmoSenticSpace, described in Section 3.4.

Bag of concepts: For each concept in the text, we obtained a 100-dimensional feature vector from the EmoSenticSpace. Then we aggregated the individual concept vectors into one document vector through coordinate-wise summation:

$$x_i = \sum_{i=1}^{N} x_{ij},$$

where x_i is the ith coordinate of the document's feature vector, x_{ij} is the ith coordinate of its jth concept vector, and N is the number of concepts in the document. We have also experimented with averaging instead of summation:

$$x_i = \frac{1}{N} \sum_{i=1}^{N} x_{ij}.$$

But contrary to our expectation and in contrast to our past experience with Twitter data, summation gave better results than averaging.

Sentic feature: The polarity scores of each concept extracted from the text were obtained from SenticNet and summed to produce one scalar feature.

Negation: As we mentioned earlier, negations can change the meaning of a statement. We followed the approach of Lapponi, Read, and Ovrelid (2012) to identify the negation and reverse the polarity of the sentic feature corresponding to the concept that followed the negation marker.

After extracting the features, we built our text analysis by training model on the ISEAR dataset and evaluated this model on the transcriptions of the video files in the eNTERFACE dataset. Results of the evaluation are shown in Section 9.

8. Multimodal fusion for emotion analysis

Multimodal fusion is the heart of any multimodal emotion analysis engine. As discussed in Section 2.4, there are two main fusion techniques: feature-level fusion and decision-level fusion. In this work we used feature-level fusion. This fusion model (Kuncheva, 2004) aims to combine all the feature vectors of the available modalities.

We took a very simple approach to fusion: specifically, concatenating the feature vectors of all three modalities, to form a single long feature vector. This trivial method has the advantage of relative simplicity, yet is shown to produce significantly high accuracy. Yongjin, Ling, and Venetsanopoulos (2012) and Zhibing and Ling (2013) also used eNTERFACE dataset for detecting emotion from multimodal contents. They considered visual and audio clues available in the dataset and fused them to obtain the emotion associated with data. Zhibing and Ling (2013) only focused on the feature extraction and feature reduction technique in order to achieve optimality. They fused audio and visual modalities in both feature and decision level. Yongjin et al. (2012) conducted an extensive study on the eNTERFACE dataset. They first extracted the key features from audio and video data and then they analyzed the cross modal relationship between audio and visual features. After that, HMM was used as a classifier to understand emotion as well as to measure statistical dependence across the successive time segments. Table 15 shows the confusion matrix resulted from fusion experiment of all three modalities. Upon calculating the average accuracy of the fusion experiment from Table 15, we can see that the proposed approach outperforms the average accuracy obtained by both Yongjin et al. (2012) and Zhibing and Ling (2013). On average, our system obtained 87.95% accuracy when all three features were fused while Yongjin et al. (2012) and Zhibing and Ling (2013) got the accuracy between 75% and 80%. However, both of these state-of-the-art approaches did not report the extracted visual and audio features from the eNTERFACE dataset. The features extracted by our approach carry more information than the features extracted by Zhibing and Ling (2013) and they actually had lost some key information due to dimensionality reduction of the feature set. On the other hand, the kernel based fusion method by Yongjin et al. (2012) seems to be statistically significant but to give their system real time capability they had reduced the dimensionality of the large visual feature vector and that caused their system to perform more poorly than ours.

9. Experimental results and discussions

Since the videos in eNTERFACE dataset are manually annotated, we used this dataset as the gold standard for evaluation. As training data, we used the CK++ dataset for visual modality (including the two-way emotional vs. neutral classifier for video frames; see Sections 5.2–5.4) and ISEAR dataset for text modality; whereas for audio modality we used the same eNTERFACE dataset for training, with tenfold cross-validation evaluation scheme to exclude overfitting.

Table 8 shows the accuracy achieved in our experiments using the best configuration, along with the results reported by other researchers on the same dataset that we used for evaluation.

From Table 8, one can see that our approach outperforms all state-of-the art approaches tested on the eNTERFACE dataset, even on each individual modality. Since for two out of the three modalities we trained the classifier on one dataset but evaluated on another, our classifiers are not biased towards a particular dataset not over-fitted. Though some works presented in Table 1 report higher figures, they were performed on different and hand-crafted corpora and are incomparable with each other or with the works listed in Table 8. To the best of our knowledge eNTERFACE is the only corpus on which a number of state-of-the-art approaches have been reportedly evaluated so far, thus allowing for a relatively fair comparison.

We experimented with several classifiers both for multimodal classification as well as for comparative purposes, unimodal classification on each modality; see Table 9.

On the facial image sequences of the eNTERFACE dataset, the highest unimodal classification accuracy was achieved with the ELM classifier. Tables 10 and 11 show that success rates for *surprise*, *neutral*, and *joy* were very high. Main classification confusion was between *surprise* and *joy*, *surprise* and *anger*, *fear* and *anger*, and *disgust* and *anger* due to the similarity between facial expressions.

Table 8
Comparison of our fusion model with the state of the art system on eNTERFACE dataset.

Method	Algorithms and modalities used	Accuracy
Datcu and Rothkrantz (2009)	HMM, audio and video	56.27%
Paleari and Huet (2008)	SAMMI framework, audio and video	67.00%
Mansoorizadeh and Charkari (2010)	Async. feature fusion, audio and video	71.00%
Dobrišek et al. (2013)	GMM, audio and video	77.50%
Proposed uni-modal method	SVM, audio	78.57%
Proposed uni-modal method	SVM, text	78.70%
Proposed uni-modal method	ELM, video	81.21%
Proposed bi-modal method	SVM, audio and video	85.23%
Proposed multi-method	SVM, audio, video, and text	**87.95%**

Table 9
Performances of different emotion classifiers on different modalities using the eNTERFACE dataset.

Classifiers	Modalities			Fusion
	Visual	Speech	Text	
KNN	57.90%	57.25%	49.12%	59.45%
ANN	65.45%	67.28%	61.20%	68.25%
ELM	**81.21%**	72.17%	73.17%	84.45%
SVM	81.20%	**78.57%**	**78.70%**	**87.95%**

For the classification of facial images from both video files and facial image sequences, we performed two variants of classification procedure: one-stage 7-way classification (Table 10) and two-stage procedure explained in Section 5.3 (Table 11). The proposed two-stage procedure was found to significantly outperform the one-stage procedure on all labels.

Table 12 shows the result of our two-stage unimodal classification process performed on the eNTERFACE dataset. Since the *neutral* category was not used in the annotation scheme of this dataset, we do not include this category in Tables 12–15. As seen in Table 12, best emotion classification accuracies were achieved on *surprise*, and *joy* categories and worst on *disgust*. Again, discarding the neutral frames at the first stage of the two-stage procedure described in Section 5.4 was found to significantly improve the performance of the classifier, since the first frames of each clip, which expressed a neutral emotion, created noise in the classification process.

Table 13 shows the confusion matrix with tenfold cross validation on speech signals extracted from the video clips of the eNTERFACE dataset. Of various supervised classifiers that we tested on the speech dataset, SVM produced the best performance. Satisfactory accuracy was obtained for *surprise* and *joy*; similar to the results of facial image-based emotion classification, whereas the worst result was obtained for *disgust*. Schuller, Vlasenko, Eyben, Rigoll, and Wendemuth (2009) also used SVM classifier to recognize emotions from eNTERFACE dataset. If we calculate the average of the accuracies of all emotion classes obtained by our audio emotion classifier, our proposed classifier outperforms (Schuller et al., 2009). Though openEAR can extract more features from the JAudio which we used in our work, we found that except MFCC, the rest of the features extracted by openEAR are not relevant for the audios of short length. It should be noted that almost all videos in the eNTERFACE dataset have length between 2 and 3 s. Additionally, openEAR cannot extract some key features extracted by the JAudio toolkit. For example, "*area method moments of MFCC*", "*peak based spectral smoothness*" and "*compactness*" features which, had helped to improve the performance of the audio based emotion detection system. However, among all features we found MFCC as the most important audio feature.

For classifying the emotions associated with textual transcriptions of the eNTERFACE dataset, we built our training model on the ISEAR dataset using the SVM. Table 14 shows the results for the unimodal text analysis classifier.

Finally, concatenating the features of all three modalities, we formed feature vectors that fused all modalities. Table 15 shows the performance of our method with feature-level fusion. For each category, better accuracy was achieved compared with unimodal classifiers.

The main differences between the state-of-the-art approaches and our framework that may explain better performance of our approach can be summarized as follows (Table 16).

Table 10
Confusion matrix for the CK++ facial expression dataset using a one-stage emotion classifier (ELM classifier, tenfold cross-validation).

Actual classification	Predicted classification							
	Surprise	Joy	Sadness	Anger	Fear	Disgust	Neutral	Precision
Surprise	1142	57	19	43	26	11	31	85.92%
Joy	65	1121	27	45	25	19	29	84.22%
Sadness	13	23	461	19	13	15	4	84.12%
Anger	29	21	3	770	65	77	57	75.34%
Fear	11	9	3	47	396	42	38	72.52%
Disgust	20	13	24	38	45	639	89	73.61%
Neutral	3	6	9	5	7	2	201	86.26%

Table 11
Confusion matrix for the CK++ facial expression dataset using a two-stage emotion classifier (ELM classifier, tenfold cross-validation).

Actual classification	Predicted classification							
	Surprise	Joy	Sadness	Anger	Fear	Disgust	Neutral	Precision
Surprise	1170	49	25	43	15	6	21	88.03%
Joy	41	1191	21	37	17	6	18	89.48%
Sadness	7	12	492	17	9	4	5	89.78%
Anger	22	19	31	832	47	53	18	81.40%
Fear	9	7	14	32	445	27	12	81.50%
Disgust	14	10	12	34	37	732	29	84.33%
Neutral	3	7	3	0	0	0	220	94.42%

Table 12
Confusion matrix on eNTERFACE video clips using only visual modality (two-stage emotion classification procedure, using the: SVM classifier; eNTERFACE dataset does not have a neutral emotion label).

Actual classification	Predicted classification						
	Surprise	Joy	Sadness	Anger	Fear	Disgust	Precision
Surprise	187	12	10	4	6	1	85.00%
Joy	15	198	6	0	0	1	90.00%
Sadness	5	7	171	13	17	7	77.72%
Anger	7	3	2	169	19	20	76.81%
Fear	0	0	3	19	181	17	82.27%
Disgust	7	0	5	23	19	166	75.45%

Table 13
Confusion matrix for the audio modality of eNTERFACE dataset (SVM classifier).

Actual classification	Predicted classification						
	Surprise	Joy	Sadness	Anger	Fear	Disgust	Precision
Surprise	177	25	3	10	3	2	80.45%
Joy	29	181	0	5	3	2	82.27%
Sadness	12	15	173	0	7	13	78.63%
Anger	15	2	3	179	15	6	81.36%
Fear	0	3	12	27	163	15	74.09%
Disgust	0	3	8	19	25	165	75.00%

Table 14
Confusion matrix for text (transcriptions) of the eNTERFACE dataset (using the SVM-based emotion classifier).

Actual classification	Predicted classification						
	Surprise	Joy	Sadness	Anger	Fear	Disgust	Precision
Surprise	169	35	13	0	0	3	76.81%
Joy	30	187	3	0	0	0	85.00%
Sadness	0	1	173	17	19	10	78.63%
Anger	7	0	1	179	16	19	81.36%
Fear	2	3	11	30	164	10	74.54%
Disgust	5	0	7	27	14	167	75.90%

Table 15
Confusion matrix for the feature-level fusion (SVM classifier).

Actual classification	Predicted classification						
	Surprise	Joy	Sadness	Anger	Fear	Disgust	Precision
Surprise	195	10	3	2	3	7	88.63%
Joy	7	203	19	0	0	0	92.27%
Sadness	5	3	199	7	5	1	90.45%
Anger	15	2	3	196	2	2	89.09%
Fear	10	3	8	7	183	9	83.18%
Disgust	3	2	9	7	14	185	84.09%

Table 16
Performance of the emotion recognition from facial expression on different datasets (SVM classifier).

Dataset	State-of-the-art best accuracy on the dataset	Accuracy obtained by the proposed method
MMI dataset (Pantic et al., 2005)	55.60% (Valstar et al., 2011)	72.10%
FABO dataset (Gunes & Piccardi, 2006)	35.50% (Gunes & Piccardi, 2009)	61.21%

Two-stage classifier: Our facial expression analyzer is a two-stage classifier. First it identifies whether a facial expression expresses no emotion (neutral) or some emotion; in the latter case it then decides which specific emotion of Ekman's set it expresses. Thus we filter out the images that do not convey any emotion, which are essentially noise for the classifier, yet existing state-of-the-art frameworks still do consider these and try to assign them to some emotion class. Conversely, even if the neutral emotion is considered as a class, as we show below, our two-stage technique outperforms a simple seven-way classifier.

Selection of audio features: We used both prosodic and acoustic features. Almost all of them proved to be crucial for the audio emotion recognition system. In contrast, state-of-the-art approaches miss many of these important features. For example, Datcu and Rothkrantz (2009) used only fundamental frequency, bandwidth, and intensity as features for their audio emotion detection classifier. Dobrišek, Gajšek, Mihelič, Pavešić, and Štruc (2013) used acoustic features including MFCC, but not prosodic features.

Text analysis: Probably the most important difference from other research was our use of the text modality in the form of transcriptions of the eNTERFACE video clips; we fused the text-based emotion features with the audio–visual features. The last two rows of Table 8 show that the use of text-based features enhanced the accuracy of our system by 2.72% as compared with using only audio–visual fusion. None of the existing state-of-the-art approaches applied to the eNTERFACE dataset make use of text-based features.

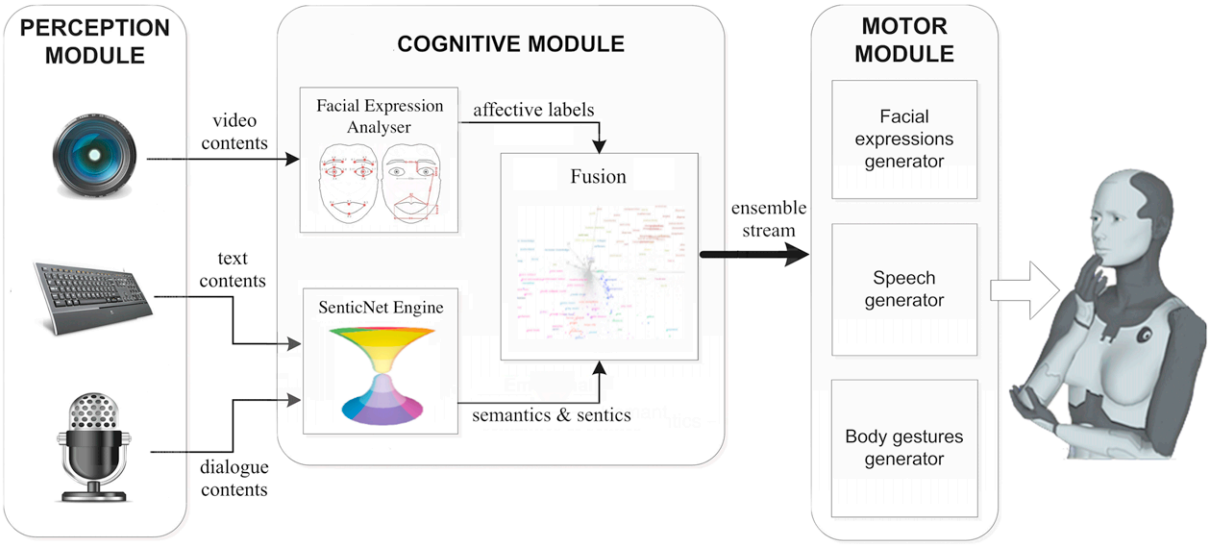

Fig. 4. Real-time emotion analysis system.

10. Developing a real-time multimodal emotion recognition system

Finally, following the steps described above, we have developed a real-time multimodal emotion recognition system. To obtain the text content of a continuous speech segment, we use a speech-to-text[5] transcription software. Fig. 4 demonstrates the system architecture. The system allows the users to upload the emotional videos and it then shows the emotion expressed by the speaker of each video. The system is available as a demo.[6]

11. Conclusions and future work

We have developed a big multimodal affective data analysis framework, which includes sets of relevant features for text, audio (speech), and visual data, as well as a simple yet effective technique for fusing the features extracted from different modalities. In particular, our textual emotion analysis module has been enriched by sentic-computing-based features, which have offered significant improvement in the performance of our textual emotion analysis system. As part of this effort, we have developed a novel lexical resource, EmoSenticSpace, which will be useful for other tasks of emotion and sentiment detection from text. Our two-stage emotion detection classifier from facial images also enhanced the system's accuracy.

Our system outperformed all state-of-the-art systems on the eNTERFACE dataset—the only publicly available dataset on which multiple systems have been analyzed, allowing for fair comparison. Moreover, our system outperformed others even when it used any one of the three single modalities, despite those systems being multimodal—this demonstrates the advantage of employing our proposed feature sets and classification techniques for video, speech, and textual data. With multimodal fusion, our system outperformed the best state-of-the-art system by more than 10%or, in relative terms, achieved a 56% reduction in error rate.

The preliminary work reported in this paper opens a number of interesting directions for future work. The most obvious ones include using Fundamental Code Unit (Howard, 2012) and sentic computing for decision-level fusion of the three modalities. Other fusion techniques can also be explored to obtain a detailed comparison of the performance of different fusion techniques. Recently introduced novel Syntactic Dependency-Based N-grams features (Jimenez Vargas & Gelbukh, 2011, 2012; Sidorov, Velasquez, Stamatatos, Gelbukh, & Chanona-Hernández, 2013a, 2013b) can also potentially improve the results for the textual modality. Finally, in order to realize our ambitious goal of developing a novel real-time system for multimodal emotion analysis, the time complexity of the methods need to be reduced to a minimum. Hence, another aspect of our future work is to effectively analyze and appropriately address the system's time complexity requirements in order to create a better, time-efficient, and reliable multimodal emotion analysis engine.

[5] http://cmusphinx.sourceforge.net/.

[6] http://sentic.net/demo/.

References

Alm, C., Roth, D., & Sproat, R. (2005a). Emotions from text: Machine learning for text-based emotion prediction. In *HLT and EMNLP*. (pp. 579–586). Vancouver, British Columbia, Canada.

Alm, C., Roth, D., & Sproat, R. (2005b). Emotions from text: Machine learning for text-based emotion prediction. in: *Proceedings of the conference on empirical methods in natural language processing* (pp 347–354) Vancouver, Canada.

Arora, P., Bakliwal, A., & Varma, V. (2012). Hindi subjective Lexicon generation using WordNet graph traversal. *International Journal of Computational Linguistics and Applications*, *3*(1), 25–39.

Atassi, H., & Esposito, A. (2008). A speaker independent approach to the classification of emotional vocal expressions (pp. 147–152).

Balahur, A., Hermida, J. M., & Montoyo, A. (2012). Building and exploiting emotinet, a knowledge base for emotion detection based on theappraisal theory model. *IEEE Transactions on Affective Computing*, *3*(1).

Bazzanella, C. (2004). Emotions, language and context. In E. Weigand (Ed.), *Emotion in dialogic interaction. Advances in the complex* (pp. 59–76). Amsterdam, Philadelphia: Benjamins.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A database of german emotional speech. In *Interspeech* (pp. 1517–1520).

Cambria, E., Howard, N., Hsu, J., & Hussain, A. (2013). Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics. In *IEEE SSCI* (pp. 108–117).

Cambria, E., Hussain, A., Havasi, C., & Eckl, C. (2009). Common sense computing: from the society of mind to digital intuition and beyond. In *LNCS*: *Vol. 5707* (pp. 252–259). Springer.

Cambria, E., Hussain, A., Havasi, C., & Eckl, C. (2010a). Sentic computing: Exploitation of common sense for the development of emotion-sensitive systems. In *LNCS*: *Vol. 5967* (pp. 148–156). Springer.

Cambria, E., Hussain, A., Havasi, C., & Eckl, C. (2010b). SenticSpace: Visualizing opinions and sentiments in a multi-dimensional vector space. In *LNAI*: *Vol. 6279* (pp. 385–393). Springer.

Cambria, E., Mazzocco, T., Hussain, A., & Eckl, C. (2011). Sentic medoids: Organizing affective common sense knowledge in a multi-dimensional vector space. In *LNCS*: *Vol. 6677* (pp. 601–610). Springer.

Cambria, E., Olsher, D., & Rajagopal, D. (2014). SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In *AAAI* (pp. 1515–1521).

Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S., et al. (2007). Multimodal emotion recognition from expressive faces, body gestures and speech. In *Artificial intelligence and innovations 2007: From theory to applications* (pp. 375–388). US: Springer.

Chaumartin, F. (2007). UPAR7: A knowledge-based system for headline sentiment tagging. In *SemEval-2007* (pp. 422–425). Prague: ACL.

Chen, L. S. H. (2000). *Joint processing of audio-visual information for the recognition of emotional expressions in human–computer interaction*. diss., University of Illinois.

Chiu, C.C., Chang, Y.L., & Lai, Y.J. (1994). The analysis and recognition of human vocal emotions. In *Proc. international computer symposium* (pp. 83–88).

Cohen, I., Sebe, N., Garg, A., Chen, Lawrence S., & Huang, Thomas S. (2003). Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, *91*(1), 160–187.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cowie, R., & Douglas-Cowie, E. (1996). Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In *Proc. international conf. on spoken language processing* (pp. 1989–1992).

Datcu, D., & Rothkrantz, L. (2008). Semantic audio-visual data fusion for automatic emotion recognition. In *Euromedia'08* (pp. 1–6).

Datcu, D., & Rothkrantz, L. (2009). Multimodal recognition of emotions in car environments. In *DCI&I 2009*.

Dellaert, F., Polzin, T., & Waibel, A. (1996). Recognizing emotion in speech. In *Proc. international conf. on spoken language processing* (pp. 1970–1973).

Dobrišek, S., Gajšek, R., Mihelič, F., Pavešić, N., & Štruc, V. (2013). Towards efficient multi-modal emotion recognition. *International Journal of Advanced Robotic Systems*, *10*(53).

Ekman, P. (1970). Universal facial expressions of emotions. In *California mental health research digest, Vol. 8* (pp. 151–158).

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: Investigator's guide*. Consulting Psychologists Press.

Essa, I. A., & Pentland, A. P. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 757–763.

Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th conference on language resources and evaluation*. LREC.

Friesen, W., & Ekman, P. (1983). EMFACS-7: Emotional facial action coding system. Unpublished manual, University of California, California.

Gunes, H., & Piccardi, M. (2006). A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *Proc. 18th int'l conf. pattern recognition, ICPR'06, Vol. 1* (pp. 1148–1153).

Gunes, H., & Piccardi, M. (2007). Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, *30*(4), 1334–1345.

Gunes, H., & Piccardi, M. (2009). Automatic temporal segment detection and affect recognition from face and body display. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, *39*(1), 64–84.

Havasi, C., Speer, R., & Pustejovsky, J. (2009). Automatically suggesting semantic structure for a generative Lexicon ontology. In *Generative Lexicon*.

Howard, N. (2012). Brain language: The fundamental code unit. *The Brain Sciences Journal*, *1*(1), 4–45. 999.

Howard, N., & Cambria, E. (2013). Intention awareness: Improving upon situation awareness in human-centric environments. *Human-Centric Computing and Information Sciences*, *3*(9).

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *SIGKDD*, Seattle.

Huang, G.-B., Zhu, Q., & Siew, C. (2006). Extreme learning machine: theory and applications. *Neurocomputing*, *70*(1), 489–501.

Jimenez Vargas, S., & Gelbukh, A. (2011). SC spectra: A linear-time soft cardinality approximation for text comparison. *Lecture Notes in Artificial Intelligence*, *7095*, 213–224.

Jimenez Vargas, S., & Gelbukh, A. (2012). Baselines for natural language processing tasks based on soft cardinality spectra. *International Journal of Applied and Computational Mathematics*, *11*(2), 180–199.

Johnstone, T. (1996). Emotional speech elicited using computer games. In *Proc. international conf. on spoken language processing* (pp. 1985–1988).

Kapoor, A., Burleson, W., & Picard, R. (2007). Automatic prediction of frustration. *International Journal of Human–Computer Studies*, *65*, 724–736.

Kuncheva, L. (2004). *Combining pattern classifiers: Methods and algorithms*. Wiley & Sons.

Lanitis, A., Taylor, C. J., & Cootes, T. F. (1995). A unified approach to coding and interpreting face images. In *Fifth international conference on computer vision, proceedings* (pp. 368–373).

Lapponi, E., Read, J., & Ovrelid, L. (2012). Representing and resolving negation for sentiment analysis. In *ICDM SENTIRE*, (pp. 687–692) Brussels.

Lin, K.H.-Y., Yang, C., & Chen, H.H. (2007). What emotions news articles trigger in their readers? In *Proceedings of SIGIR* (pp. 733–734).

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW)*. IEEE.

Maas, A., Daly, R., Pham, P., Huang, D., Ng, A., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the association for computational linguistics, ACL 2011*, Portland.

Mansoorizadeh, M., & Charkari, N. M. (2010). Multimodal information fusion application to human emotion recognition from face and speech. *Multimedia Tools and Applications*, *49*(2), 277–297.

Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The eNTERFACE'05 audio-visual emotion database. In *Proceedings of the first IEEE workshop on multimedia database management*.

Mase, K. (1991). Recognition of facial expression from optical flow. *IEICE Transactions*, *E74*(10), 3474–3483.

Matsumoto, D. (1992). More evidence for the universality of a contempt expression. *Motivation and Emotion*, *16*(4), 363–368.

McKay, Cory, Ichiro, Fujinaga, & Philippe, Depalle (2005). jAudio: A feature extraction library. In *Proceedings of the International Conference on Music Information Retrieval* (pp. 600–603).

Melville, Prem, Gryc, Wojciech, & Lawrence, Richard D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining*.

Mishne, G. (2005), Experiments with mood classification in blog posts. In *Proceedings of the 1st workshop on stylistic analysis of text forinformation access, Style 2005*, Brazil.

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: a review of the literature of human vocal emotion. *The Journal of the Acoustical Society of America*, *93*(2), 1097–1108.

Navas, E., & Hernez, L. (2006). An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS. *IEEE Transactions on Audio, Speech and Language Processing*, *14*, 1117–1127.

Ortony, A., & Terence, J. (1990). What's basic about basic emotions? *Psychological Review*, *97*(3), 315.

Otsuka, T., & Ohya, J. (1997). *A study of transformation of facial expressions based on expression recognition from temproal image sequences, Tech. rep.*. Institute of Electronic, Information, and Communications Engineers (IEICE).

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC*.

Paleari, M., & Huet, B. (2008). Toward emotion indexing of multimedia excerpts. In *CBMI*, London.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd meeting of the association for computational linguistics*, Barcelona.

Pantic, M., Valstar, M. F., Rademaker, R., & Maat, L. (2005). Web-based database for facial expression analysis. In *Proc. 13th ACM int'l conf. multimedia, Multimedia'05* (pp. 317–321).

Picard, R. (1997). *Affective computing*. Boston: The MIT Press.

Poria, S., Cambria, E., Winterstein, G., & Huang, G.-B. (2014). Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, *69*, 45–63.

Poria, S., Gelbukh, A., Hussain, A., Das, D., & Bandyopadhyay, S. (2013). Enhanced SenticNet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, *28*(2), 31–38.

Pudil, P., Ferri, F., Novovicova, J., & Kittler, J. (1994). Floating search method for feature selection with non monotonic criterion functions. *Pattern Recognition*, *2*, 279–283.

Pun, T., Alecu, T., Chanel, G., Kronegg, J., & Voloshynovskiy, S. (2006). Brain–computer interaction research at the computer vision and multimedia laboratory. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *14*(2), 210–213.

Qi, H., & Wang, X.L. (2001). Multisensor data fusion in distributed sensor. Networks using mobile agents.

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Conference on empirical methods in natural language processing, EMNLP-03* (pp. 105–112).

Rosenblum, M., Yacoob, Y., & Davis, L. (1996). Human expression recognition from motion using a radial basis function network architecture. *IEEE Transactions on Neural Networks*, *7*, 1121–1138.

Sato, J., & Morishima, S. (1996). Emotion modeling in speech production using emotion space. In *Proc. IEEE int. workshop on robot and human communication* (pp. 472–477).

Scherer, K. R. (1996). Adding the affective dimension: a new look in speech analysis and synthesis. In *Proc. international conf. on spoken language processing* (pp. 1808–1811).

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., & Wendemuth, A. (2009). Acoustic emotion recognition: A benchmark comparison of performances. In *IEEE workshop on automatic speech recognition & understanding, ASRU'09* (pp. 552–557).

Shan, C., Gong, S., & McOwan, P. (2007). Beyond facial expressions: Learning human emotion from body gestures. In *BMVC*, Warwick.

Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., et al. (2013). Empirical study of machine learning based approach for opinion mining in tweets. *Lecture Notes in Artificial Intelligence*, *7629*, 1–14.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2013a). Syntactic dependency-based N-grams as classification features. *Lecture Notes in Artificial Intelligence*, *7630*, 1–11.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2013b). Syntactic dependency-based N-grams: More evidence of usefulness in classification. *Lecture Notes in Computer Science*, *7816*, 13–24.

Speer, R., & Havasi, C. (2012). ConceptNet 5: A large semantic network for relational knowledge. In E. Hovy, M. Johnson, & G. Hirst (Eds.), *Theory and applications of natural language processing*, Springer, (Chapter 6).

Strapparava, C., & Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of the 4th international workshop on the semantic evaluations, SemEval 2007*, Prague.

Strapparava, C., & Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the ACM conference on applied computing, ACM-SAC 2008, Fortaleza*.

Strapparava, C., & Valitutti, A. (2004). WordNet-affect: an affective extension of WordNet. In *Proceedings of the 4th international conference on language resources and evaluation*, Lisbon.

Turney, P. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting of the association for computational linguistics, ACL 2002*, (pp. 417–424) Philadelphia.

Tzanetakis, G. (2002). Music genre classification of audio signal. *IEEE Transactions on Speech and Audio Processing*, *10*(5).

Ueki, N., Morishima, S., Yamada, H., & Harashima, H. (1994). Expression analysis/synthesis system based on emotion space constructed by multilayered neural network. *Systems and Computers in Japan*, *25*(13), 95–103.

Valstar, Michel François, Jiang, Bihan, Mehu, Marc, Pantic, Maja, & Scherer, Klaus (2011). The first facial expression recognition and analysis challenge. In *2011 IEEE international conference on automatic face & gesture recognition and workshops, FG 2011* (pp. 921–926).

Wawer, A. (2012). Extracting emotive patterns for languages with rich morphology. *International Journal of Computational Linguistics and Applications*, *3*(1), 11–24.

Wiebe, J. (2010). Learning subjective adjectives from corpora. In *Proceedings of the American association for artificial intelligence, AAAI 2000*, (pp. 735–740) Austin, Texas, 2000.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technologies conference/conference on empirical methods in natural language processing, HLT/EMNLP 2005*, Vancouver.

Xia, R., Zong, C. Q., Hu, X. L., & Cambria, E. (2013). Feature ensemble plus sample selection: A comprehensive approach to domain adaptation for sentiment classification. *IEEE Intelligent Systems*, *28*(3), 10–18.

Yacoob, Y., & Davis, L. (1996). Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*, 636–642.

Yongjin, W., Ling, G., & Venetsanopoulos, A. N. (2012). Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. *IEEE Transactions on Multimedia*, *14*(3), 597–607.

Zeng, Z., Tu, J., Liu, M., Huang, T., Pianfetti, B., Roth, D., et al. (2007). Audio–visual affect recognition. *IEEE Transactions on Multimedia*, *9*(2), 424–428.

Zhibing, X., & Ling, G. (2013). Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools. In *2013 IEEE international conference on multimedia and expo, ICME, Vol. 1, no. 6* (pp. 15–19).