Information Fusion 37 (2017) 98–125

Contents lists available at ScienceDirect

Information Fusion

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

A review of affective computing: From unimodal analysis to multimodal fusion

Soujanya Poria[a], Erik Cambria[c,*], Rajiv Bajpai[b], Amir Hussain[a]

[a] School of Natural Sciences, University of Stirling, UK
[b] Temasek Laboratories, Nanyang Technological University, Singapore
[c] School of Computer Science and Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 7 September 2016
Revised 31 January 2017
Accepted 1 February 2017
Available online 3 February 2017

Keywords:
Affective computing
Sentiment analysis
Multimodal affect analysis
Multimodal fusion
Audio, visual and text information fusion

ABSTRACT

Affective computing is an emerging interdisciplinary research field bringing together researchers and practitioners from various fields, ranging from artificial intelligence, natural language processing, to cognitive and social sciences. With the proliferation of videos posted online (e.g., on YouTube, Facebook, Twitter) for product reviews, movie reviews, political views, and more, affective computing research has increasingly evolved from conventional unimodal analysis to more complex forms of multimodal analysis. This is the primary motivation behind our first of its kind, comprehensive literature review of the diverse field of affective computing. Furthermore, existing literature surveys lack a detailed discussion of state of the art in multimodal affect analysis frameworks, which this review aims to address. Multimodality is defined by the presence of more than one modality or channel, e.g., visual, audio, text, gestures, and eye gage. In this paper, we focus mainly on the use of audio, visual and text information for multimodal affect analysis, since around 90% of the relevant literature appears to cover these three modalities. Following an overview of different techniques for unimodal affect analysis, we outline existing methods for fusing information from different modalities. As part of this review, we carry out an extensive study of different categories of state-of-the-art fusion techniques, followed by a critical analysis of potential performance improvements with multimodal analysis compared to unimodal analysis. A comprehensive overview of these two complementary fields aims to form the building blocks for readers, to better understand this challenging and exciting research field.

1. Introduction

Affective computing is an emerging field of research that aims to enable intelligent systems to recognize, feel, infer and interpret human emotions. It is an interdisciplinary field which spans from computer science to psychology, and from social science to cognitive science. Though sentiment analysis and emotion recognition are two distinct research topics, they are conjoined under the field of *Affective Computing* research.

Emotions and sentiments play a crucial role in our daily lives. They aid decision-making, learning, communication, and situation awareness in human-centric environments. Over the past two decades or so, AI researchers have been attempting to endow machines with cognitive capabilities to recognize, interpret and express emotions and sentiments. All such efforts can be attributed to affective computing. Emotion and sentiment analysis have also become a new trend in social media, avidly helping users to understand the opinion being expressed on different platforms [1,2].

With the advancement of technology, abundance of smartphones and the rapid rise of social media, huge amount of Big Data is being uploaded as videos, rather than text alone [3]. Consumers for instance, tend to record their reviews and opinions on products using a web camera and upload them on social media platforms like YouTube or Facebook to inform subscribers about their views. These videos often contain comparisons of products from competing brands, the pros and cons of product specifications, etc., which can aid prospective buyers in making an informed decision.

The primary advantage of analyzing videos over textual analysis, for detecting emotions and sentiments from opinions, is the surplus of behavioral cues. Whilst textual analysis facilities only make use of words, phrases and relations, as well as dependencies among them, these are known to be insufficient for extracting associated affective content from textual opinions [4]. Video opinions, on the other hand, provide multimodal data in terms of vocal and visual modality. The vocal modulations of opinions and facial

* Corresponding author.
E-mail address: cambria@ntu.edu.sg (E. Cambria).

http://dx.doi.org/10.1016/j.inffus.2017.02.003
1566-2535/

expressions in the visual data, along with textual data, can provide important cues to better identify true affective states of the opinion holder. Thus, a combination of text and video data can help create a better emotion and sentiment analysis model.

To date, most of the research work in this field has focused on multimodal emotion recognition using visual and aural information. On the other hand, there is currently rather scarce literature on multimodal sentiment analysis. Furthermore, most of the work in sentiment analysis has thus far been carried out in the field of natural language processing (NLP), hence, the primary datasets and resources available are restricted to text-based opinion mining. However, with the advent of social media, people are extensively using multimodal social media platforms to express their opinions, making use of videos (e.g., YouTube, Vimeo, VideoLectures), images (e.g., Flickr, Picasa, Facebook) and audios (e.g., podcasts). Thus, it is highly crucial to effectively mine opinions and identify affective information from these diverse Big Data modalities.

Research in the affective computing field is continuing to attract the attention of academia and industry alike. This, 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 multimodal 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 practical real-world applications.

The aim of multi-sensor data fusion is to increase the accuracy and reliability of estimates [5]. Many applications, e.g., navigation tools, have already demonstrated the potential of data fusion. This depicts the importance and feasibility of developing a multimodal framework that could cope with all three sensing modalities: text, audio, and video, in human-centric environments. The way humans naturally communicate and express their emotions and sentiments is usually 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, thereby emphasizing the importance of such seamless fusion.

Aural data in a video expresses the tone of the speaker while the visual data conveys facial expressions, which in turn can further aid in understanding of the users' affective state. The data obtained from a video can be a useful source of information for emotion and sentiment analysis, but there are major challenges which need to be addressed. For example, the way we convey and express opinions varies from person to person [6]. A person may express his/her opinions more vocally while others do so more visually. When a person expresses his/her opinions with more vocal modulation, the audio data may contain major cues for opinion mining. On the other hand, when a person is making use of more facial expressions, most of the cues needed for opinion mining can be assumed to reside in their facial expressions. Hence, a generic context-sensitive model needs to be developed which can adapt itself for any user and can give a consistent result in any real-world environment.

As human beings, we also rely on multimodal information more than unimodal [7]. It is apparent that we get a better understanding of a speaker's intention when we see his/her facial expressions while he/she is speaking. Together aural and visual mediums provide more information than they provide alone. This is often the case when the brain relies on several sources of sensory inputs in validating events. Using them, it compensates for any incomplete information which can hinder decision processes. For example, during a car accident, a person may not see any flames (visual) but the smell of burning rubber and heat proliferating through the dash would signal to the brain that a fire is kindling, thus demanding an immediate exit from the vehicle. In this example, the information driving the brain's reaction is greater than the aggregated dissimilar sensory inputs.

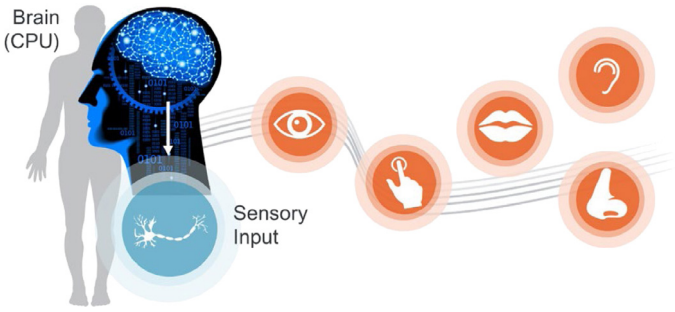


Fig. 1. The human brain considers multisensor information together for decision making.

The ability of a multimodal system to outperform a unimodal system is well established in the literature [8]. However, there is a lack of a comprehensive literature survey, focusing on recent successful methods employed in this research area. Unimodal systems are building blocks for a multimodal system, hence, we require them to be performing well in order to build an intelligent multimodal system. In this paper, we not only discuss the multimodal fusion of unimodal information, but also discuss the state of the art in unimodal affect recognition methods, as these are key preprocessing steps for the multimodal fusion.

Tables 3 and 4 give us an insight of the inclusion of more than two modalities for multimodal affect analysis. Despite considerable advances reported in this field to date, there remain significant outstanding research challenges, before real-time multimodal affect recognition enabled smart devices make their way into our every day lives. Some of the major challenges that need to be addressed, are listed below:

- Continuous data from real noisy sensors may generate incorrect data.
- Identifying whether the extracted audio and utterance refer to the same content.
- Multimodal affect analysis models should be trained on Big Data from diverse contexts, in order build generalized models.
- Effective modeling of temporal information in the Big Data.
- For real-time analysis of multi-modal Big Data, an appropriately scalable Big Data architecture and platform needs to be designed, to effectively cope with the heterogeneous Big Data challenges of growing space and time complexity.

1.1. The scope of this survey

As multimodal sentiment analysis and emotion recognition research continues to gain popularity among the AI and NLP research communities, there is need for a timely, thorough literature review to define future directions, which can, in particular, further the progress of early stage researchers interested in this multidisciplinary field (Fig. 1).

The only other recent survey of multimodal affect analysis [8], focuses mainly on the state of the art in collecting sample data, and reports performance comparison of selected multimodal and unimodal systems, as opposed to comprehensively reviewing key individual systems and approaches, from the growing literature in the field. In this study, for example, it is reported that multimodal "systems were consistently (85% of systems) more accurate than their best unimodal counterparts, with an average improvement of 9.83% (median of 6.60%)".

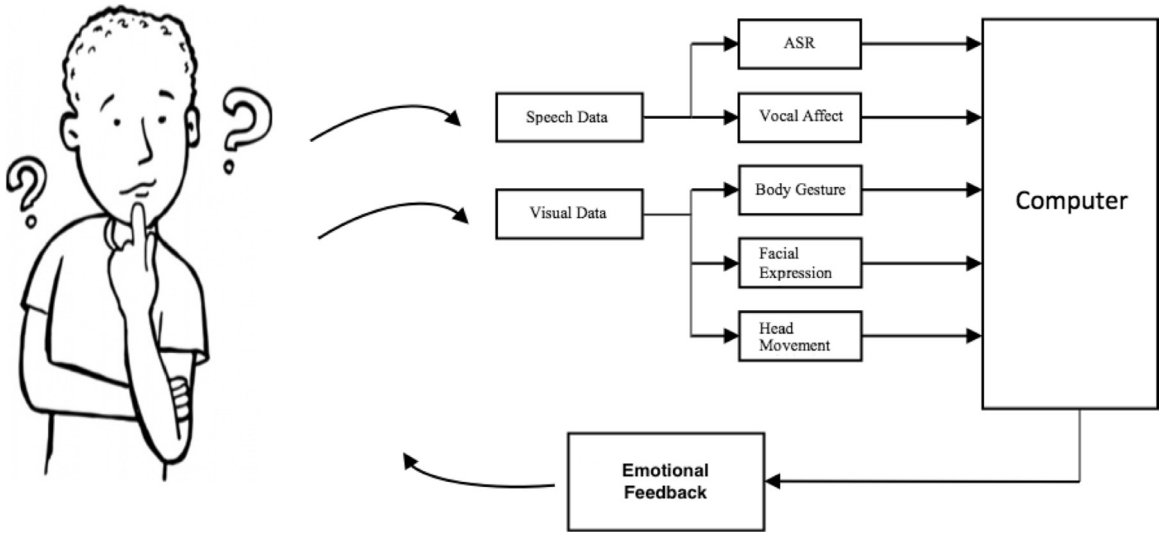


Fig. 2. A typical multimodal affect analysis framework.

On the other hand, our paper provides a first of its kind, comprehensive literature review, which also serves as an educational tutorial-type medium for novice researchers to enable them to understand the complex literature and carry out comparative experimental evaluations of benchmark methods and resources. According to this recent survey [8], numerous studies have confirmed the potential for multimodal systems to significantly outperform unimodal systems. This finding alone is a sufficient motivation for beginners to explore this field. Apart from the survey by D'Mello et al. [8], there is another extensive literature survey done by Zeng et al. [9] on multimodal emotion recognition, which mainly focuses on identifying challenges involved in collecting and processing audio, visual and audio-visual (i.e., multimodal) data.

In this article, we aim to give readers an overview of key state-of-the-art methodologies in order to inform them about major past and recent works, and trends in this field. We start by discussing the available datasets and key works in the visual, audio and textual modalities, by briefly describing the methodologies used for each. The methodologies are clustered into definite categories for readers to easily identify their points of interest. This is followed by a detailed discussion of the different types of fusion methodologies prevalent in the literature. In the end, we also provide a list of available application program interfaces (APIs) for multimodal affect analysis.

D'Mello et al. [8] thoroughly discuss unimodal and multimodal accuracy comparison using statistical measures. They have proposed statistical methods in order to compare accuracy of different algorithms on different datasets. So, as there exists a recent literature survey which discusses primarily the accuracy comparison of multimodal methods across datasets, in this paper we do not focus on that aspect. Instead, we focus on categorization of different methods, comparing them based on their individual approach. In particular, we avoid comparing the performance of multimodal methods on different datasets as it is highly challenging and controversial to find a generic method for that comparison. Thus, we compare research on the same datasets (see Table 5) based on their accuracy. We also separately discuss multimodal emotion and sentiment classification methods.

Fig. 2 shows the overall framework of a typical multimodal affect detector. The framework consists of two fundamental steps: processing unimodal data separately and fusing them all together. Both steps are equally important: poor analysis of a modality can worsen the multimodal system's performance, while an inefficient fusion can ruin the multimodal system's stability. Thus, in this article, we decided to review both unimodal affect analysis literature, as well as the research works on fusion. Finally, this paper is presented in a way that researchers can identify key visual, audio and text analysis methods from the literature, and fuse them using state-of-the-art methods.

The rest of the paper is organized as follows: Section 2 defines affective computing; Section 3 lists available datasets for multimodal emotion and sentiment analysis; Section 4 discusses feature extraction from visual, audio and text modalities; Section 5 illustrates the various fusion methods for data collected from multimodal sources and presents an overview of notable multimodal emotion recognition and sentiment analysis research carried out recently; Section 6 provides a list of available APIs for multimodal affect analysis; Section 7 presents our observations on the literature, as well as future work directions; finally, Section 8 concludes the research article.

2. Affective computing

Before discussing the literature on unimodal and multimodal approaches to affect recognition, we introduce the notion of 'affect' and 'affect taxonomy'. Affective computing is the set of techniques aimed at performing affect recognition from data, in different modalities and at different granularity scales. Sentiment analysis, for example, performs coarse-grained affect recognition, as it is usually considered a binary classification task (positive versus negative), while emotion recognition performs fine-grained affect recognition, as it aims to classify data according to a large set of emotion labels.

In this paper, we focus on these two kinds of affect recognition by specifically identifying datasets (Table 1, 2) and works (Table 5, 6) in both fields. While there is a fixed taxonomy for sentiment which is bound within positive, negative and neutral sentiments, the taxonomy for emotions is diverse. Philosophical studies on emotions date back to ancient Greeks and Romans. Following the early Stoics, for example, Cicero enumerated and organized the emotions into four basic categories: *metus* (fear), *aegritudo* (pain), *libido* (lust), and *laetitia* (pleasure). Studies on the evolutionary theory of emotions, in turn, were initiated in the late 19th century by Darwin [10]. His thesis was that emotions evolved via natural selection and, therefore, have cross-culturally universal counterparts. In the early 1970s, Ekman found evidence that humans share six basic emotions: happiness, sadness, fear, anger, disgust, and sur-

Table 1
Multimodal emotion analysis datasets.

Dataset	References	Modality	Speakers	Features	Sentiment	Annotators	Availability
HUMAINE	Cowie et al. [28]	A+V	50	Emotion words, authenticity, core affect dimensions,context labels	NA	6 (only 16 clips labeled)	Publicly available http://emotion-research.net/download/pilot-db
Belfast database	Cowie et al. [29]	A+V	125 (31 M, 94 F)	Wide range of emotions	NA	7	On registration http://belfast-naturalistic-db.sspnet.eu
SEMAINE	McKeown et al. [30]	A+V	150	Angry, happy, fear, disgust, disgust, sadness, contempt and amusement	NA	6	On registration http://semaine-db.eu
IEMOCAP	Busso et al. [32]	A+V	10 (5 M, 5 F)	Happiness, anger, sadness, frustration and neutral state	NA	3 for each emotion category	On request http://sail.usc.edu/iemocap
eNTERFACE	Martin et al. [33]	A+V	42(34 M, 8 F)	Happiness, anger, sadness, surprise, disgust and fear	NA	2	Publicly available http://enterface.net

Legenda: A=Audio; V=Video

Table 2
Multimodal sentiment analysis datasets.

Dataset	References	Modality	Speakers	Features	Sentiment	Annotators	Availability
ICT-MMMO	Wollmer et al. [27]	A+T+V	370	1000 linguistic +1941 acoustic +20 visual	Strongly Negative, Weakly Negative, Neutral, Weakly Positive and Strongly Positive	3	By sending mail to Giota Stratou (stratou@ict.usc.edu)
MOUD	Rosas et al. [26]	A+T+V	80 (65 F, 15 M)	28 acoustic +40 visual	Positive, Negative, and Neutral	2	Publicly available http://web.eecs.umich.edu/mihalcea/downloads.html
YouTube dataset	Morency et al. [6]	A+T+V	47 (20 F, 27 M)	1000 linguistic +1941 acoustic +20 visual	Polarized words, smile, look away, Pauses and Pitch	3	By sending mail to Giota Stratou (stratou@ict.usc.edu)

Legenda: A=Audio; T=Text; V=Video

Table 3
State of the art of multimodal affect recognition where the text modality has been used.

References	Data Type	Modality	Fusion Type
Chuang & Wu (2004) [192]	act	A+T	dec
Forbes-Riley&Litman (2004) [193]	nat	A+T	feat
Litman&Forbes-Riley (2004) [194]	nat	A+T	feat
Rigoll et al. (2005) [195]	act	A+T	dec
Litman&Forbes-Riley (2006) [196]	nat	A+T	feat
Seppi et al. (2008) [197]	ind	A+T	feat
Eyben et al. (2010) [121]	ind	A+T	model
Schuller (2011) [198]	nat	A+T	feat
Wu and Liang (2011) [97]	act	A+T	dec
Rozgic et al. (2012) [199]	act	A+T+V	feat
Savran et al. (2012) [200]	ind	A+T+V	model
Rosas et al. (2013) [4]	nat	A+T+V	feat
Wollmer et al. (2013) [27]	nat	A+T+V	hybrid
Sarkar et al. (2014) [201]	nat	A+T+V	feat
Alam et al. (2014) [202]	nat	A+T+V	dec
Ellis et al. (2014) [203]	nat	A+T+V	dec
Poria et al. (2014) [202]	act	A+T+V	feat
Siddiquie et al. (2015) [204]	nat	A+T+V	hybrid
Poria et al. (2015) [205]	nat	A+T+V	dec
Poria et al. (2015) [189]	nat	A+T+V	feat
Cai et al. (2015) [206]	nat	T+V	dec
Ji et al. (2015) [207]	nat	T+V	model
Yamasaki et al. (2015) [208]	nat	A+T	model
Poria et al. (2016) [94]	nat	A+T+V	feat

Legenda: Data Type (act=Acted, ind=Induced, nat=Natural); Modality (V=Video, A=Audio, T=Text); Fusion Type (feat=Feature; dec=Decision).

prise [11]. Few tentative efforts to detect non-basic affective states, such as fatigue, anxiety, satisfaction, confusion, or frustration, have been also made [12–14].

In 1980, Averill put forward the idea that emotions cannot be explained strictly on the basis of physiological or cognitive terms. Instead, he claimed that emotions are primarily social constructs; hence, a social level of analysis is necessary to truly understand the nature of emotions. The relationship between emotions and language (and the fact that the language of emotions is considered a vital part of the experience of emotions) has been used by social constructivists and anthropologists to question the universality of Ekman's studies, arguably because the language labels he used to code emotions are somewhat US-centric. In addition, other cultures might have labels that cannot be literally translated to English (e.g., some languages do not have a word for fear [15]).

For their deep connection with language and for the limitedness of the emotional labels used, all such categorical approaches usually fail to describe the complex range of emotions that can occur in daily communication. The dimensional approach [16], in turn, represents emotions as coordinates in a multi-dimensional space. For both theoretical and practical reasons, an increasing number of researchers prefer to define emotions according to two or more dimensions. An early example is Russell's circumplex model [17], which uses the dimensions of arousal and valence to plot 150 affective labels.

Similarly, Whissell considers emotions as a continuous 2D space whose dimensions are evaluation and activation [18]. The evaluation dimension measures how a human feels, from positive to negative. The activation dimension measures whether humans are more or less likely to take some action under the emotional state, from active to passive. In her study, Whissell assigns a pair of values < activation, evaluation> to each of the approximately 9,000 words with affective connotations that make up her Dictionary of Affect in Language.

Another bi-dimensional model is Plutchik's wheel of emotions, which offers an integrative theory based on evolutionary principles [19]. Following Darwin's thought, the functionalist approach to emotions holds that emotions have evolved for a particular function, such as to keep the subject safe [20]. Emotions are adaptive as they have a complexity born of a long evolutionary history and, although we conceive emotions as feeling states, Plutchik says the feeling state is part of a process involving both cognition and be-

Table 4
State of the art of visual-audio affect recognition

References	Data Type	Modality	Fusion Type
Busso et al. (2004) [211]	act	V+A	feat
Chen et al. (2005) [212]	act	V+A	feat
Gunes & Piccardi (2005) [213]	act	V+B	feat
Hoch et al. (2005) [214]	act	V+A	dec
Kapoor & Picard (2005) [215]	nat	V+B+C	model
Kim et al. (2005) [216]	ind	A+Pp	feat
Wang & Guan (2005) [217]	act	V+A	feat
Zeng et al. (2005) [218]	act	V+A	model
Gunes & Piccardi (2005) [219]	act	V+B	feat
Pal et al. (2006) [220]	nat	V+A	dec
Sebe et al. (2006) [221]	act	V+A	model
Zeng et al. (2006) [222]	nat	V+A	model
Caridakis et al. (2006) [223]	ind	V+A+B	model
D'Mello & Graesser (2007) [224]	nat	B+C	feat
Gong et al. (2007) [225]	act	V+B	feat
Han et al. (2007) [226]	act	V+A	dec
Joo et al. (2007) [227]	act	V+A	dec
Karpouzis et al. [228] (2007)	ind	V+A	feat
Kim (2007) [216]	ind	A+Pp	feat
Schuller et al. (2007) [229]	nat	V+A	feat
Shan et al. (2007) [230]	act	V+B	feat
Zeng et al. (2007) [231]	act	V+A	model
Haq et al. (2008) [232]	act	V+A	feat
Kanluan et al. (2008) [233]	nat	V+A	dec
Metallinou et al. (2008) [234]	act	V+A	dec
Wang & Guan (2008) [217]	act	V+A	feat
Wimmer et al. (2008) [235]	ind	V+A	feat
Bailenson et al. (2008) [236]	ind	V+Pp	feat
Castellano et al. (2008)[237]	act	V+A+B	feat
Chetty & Wagner (2008) [238]	act	V+A	hybrid
Castellano et al. (2009) [237]	nat	V+C	feat
Emerich et al. (2009) [239]	act	V+A	feat
Gunes & Piccardi (2009) [240]	act	V+B	feat
Haq & Jackson (2009) [241]	act	V+A	dec
Khalali and Moradi (2009) [242]	ind	Cp+Pp	feat
Paleari et al. (2009) [243]	act	V+A	model
Rabie et al. (2009) [244]	act	V+A	model
D'Mello and Graesser (2010) [245]	nat	V+A+C	feat
Dy et al. (2010) [246]	nat	V+A	dec
Gajsek et al. (2010) [247]	act	V+A	dec
Kessous et al. (2010) [248]	act	V+A+B	feat
Kim and Lingenfelser (2010) [249]	ind	A+Pp	dec
Mansoorizadeh and Charkari (2010) [250]	act	V+A	hybrid
Mansoorizadeh and Charkari (2010) [250]	act	V+A	hybrid
Wollmer et al. (2010) [251]	act	V+A	feat
Glodek et al. (2011) [252]	ind	V+A	dec
Banda and Robinson (2011) [253]	act	V+A	dec
Chanel et al. (2011) [254]	nat	Cp+Pp	dec
Cueva et al. (2011) [255]	act	V+A	dec
Datcu and Rothkrantz (2011) [256]	act	V+A	feat
Jiang et al. (2011) [257]	act	V+A	model
Lingenfelser et al. (2011) [258]	act	V+A	dec
Lingenfelser et al. (2011) [258]	ind	V+A	dec
Nicolaou et al. (2011) [259]	ind	V+A+B	model
Vu et al. (2011) [260]	act	A+B	dec
Wagner et al. (2011) [261]	act	V+A+B	dec
Walter et al. (2011) [262]	ind	A+Pp	dec
Hussain et al. (2012) [263]	ind	V+Pp	dec
Koelstra et al. (2012) [264]	ind	Cp+C+Pp	dec
Lin et al. (2012) [265]	act	V+A	model
Lin et al. (2012) [265]	ind	V+A	model
Lu and Jia (2012) [266]	act	V+A	model
Metallinou et al. (2012) [267]	act	V+A	model
Monkaresi et al. (2012) [209]	ind	V+Pp	feat
Park et al. (2012) [268]	act	V+A	dec
Rashid et al. (2012) [269]	act	V+A	dec
Soleymani et al. (2012) [270]	ind	Cp+Gaze	dec
Tu et al. (2012) [271]	act	V+A	dec
Baltrusaitis et al. (2013) [272]	ind	V+A	model
Dobrisek et al. (2013) [273]	act	V+A	dec
Glodek et al. (2013) [274]	ind	V+A	dec
Hommel et al. (2013) [275]	act	V+A	dec
Krell et al. (2013) [276]	ind	V+A	dec
Wollmer et al. (2013a) [277]	ind	V+A	model
Chen et al. (2014) [278]	act	V+A	dec
Wang et al. (2014) [210]	ind	Cp+C	feat

Legenda: Data Type (act=Acted, ind=Induced, nat=Natural); Modality (V=Video, A=Audio, B=Body, Pp=Peripheral physiology, CP=Central physiology, Content=Content/context); Fusion Type (feat=Feature, dec=Decision).

Table 5
Comparative table: key studies on multimodal emotion analysis datasets.

Datasets	Reference	Summary	Performance
SEMAINE	Gunes et al. [290]	V	0.094 (MSE)
	Valstar et al. [291]	V	68.10% (Acc)
	Eyben et al. [295]	A+V	0.190 (MLE)
HUMAINE	Chang et al. [293]	A	93.60% (Acc)
	Castellano et al. [237]	A+V	78.30% (Acc)
	Eyben et al. [121]	A+T	0.55 (CC)
eNTERFACE	Eyben et al. [107]	A	75.20% (Acc)
	Chetty et al. [238]	T+V	86.10% (Acc)
	Zhang et al. [294]	A+V	66.51% (WAA)
	Paleari et al. [243]	A+V	43.00% (WAA)
	Dobrivsek et al. [273]	A+V	77.20% (UNWAA)
	Poria et al. [205]	A+T+V	87.95% (WAA)
IEMOCAP	Rahaman et al. [296]	A	72.80%(A)
	Jio et al. [297]	A+T	69.20% (WAA)
	Metallinou et al. [234]	A+V	75.45%(UWAA)
	Rozgic et al. [298]	A+V	69.50%(WAA)
	Poria et al. [94]	A+T+V	76.85%(UNWAA)

Legenda: A=Audio; T=Text; V=Video; MSE=Mean Squared Error; MLE=Maximum Likelihood Estimate; Acc=Accuracy; WAA=Weighted Average Accuracy; UWAA=Unweighted Average Accuracy; CC=Co-relation Coefficient

havior and containing several feedback loops. In 1980, he created a wheel of emotions, which consisted of 8 basic emotions and 8 advanced emotion,s each composed of 2 basic ones. In such model, the vertical dimension represents intensity and the radial dimension represents degrees of similarity among emotions.

Besides bi-dimensional approaches, a commonly used framework for emotion representation is the < arousal, valence, dominance> set, which is known in the literature by different names, including < evaluation, activation, power> and < pleasure, arousal, dominance> [21]. Recent evidence suggests there should be a fourth dimension: Fontaine et al. reported consistent results from various cultures where a set of four dimensions is found in user studies, namely <valence, potency, arousal, unpredictability> [22]. Dimensional representations of affect are attractive mainly because they provide a way of describing emotional states that is more tractable than using words. This is of particular importance when dealing with naturalistic data, where a wide range of emotional states occur. Similarly, they are better equipped to deal with non-discrete emotions and variations in emotional states over time [23], since in such cases changing from one universal emotion label to another would not make much sense in real life scenarios.

Dimensional approaches, however, have a few limitations. Although the dimensional space allows comparison of affect words according to their reciprocal distance, it usually does not enable operations between these, e.g., for studying compound emotions. Most dimensional representations, moreover, do not model the fact that two or more emotions may be experienced at the same time. Eventually, all such approaches work at word level, which makes them unable to grasp the affective valence of multiple-word concepts.

All such limitations are overcome by the Hourglass of Emotions [24], a new affective representation based on Plutchik's model, which represents affective states both through labels, and through four independent but concomitant affective dimensions, which can potentially describe the full range of emotional experiences that are rooted in any of us (Fig. 3). By leveraging on multiple (polarized) activation levels for each affective dimension, the Hourglass of Emotions covers cases where up to four emotions can be expressed at the same time and allows for reasoning on them in an algebraic manner. The model also allows for affective common sense reasoning on both single words and multiple-word expressions [25] and provides a formula to calculate polarity based on

Table 6
Study characteristics of recent papers on multimodal analysis.

References	Modality	Speakers/Datasets	Model	Features	Fusion type
DeVault et al. [299]	A+T+V	351 (217 M, 132 F and 2 did not not report the gender)	4 statistically trained utterance classifiers	Smile intensity, 3D head position and orientation, intensity or lack of facial expressions like anger, disgust and joy, speaking fraction, dynamics, speech dynamics, gaze direction, etc.	MF
Alam et al. [202]	A+T+V	404 YouTube vloggers (194 M, 210 F)/ YouTube personality dataset	SMO for SVM	A-V, lexical, POS psycholinguistic, emotional and traits	D
Sarkar et al. [201]	A+T+V	404 YouTube vloggers/ YouTube personality dataset	LR model with ridge estimator	A-V, text, demographic and sentiment	N/A
Poria et al. [205]	A+T+V	42/ ISEAR, CK++, eNTERFACE dataset	SVM	66 FCP using Luxand software, JAudio software, BOC, Sentic features and Negation	F
Poria et al. [279]	A+T+V	47 (27 M and 20 F)/YouTube dataset SenticNet	ELM	Softwares using Luxand FSDK 1.7 and GAVAM, openEAR and Concept-gram and SenticNet-based features	F and D
Poria et al. [189]	A+T+V	47 (27 M 20 F)/YouTube dataset	Multiple Kernel Learning	Softwares using CLM-Z and GAVAM, openEAR and using CNN	D
Siddiquie et al. [204]	A+T+V	230 videos/ Rallying a Crowd (RAC) dataset	RBF SVM and LR classifier	Softwares using CAFFEE and features (prosody, MFCC, or spectrogram) and using SATSVM and DCM	F and D

Legenda: A=Audio; V=Video; T=Text; ML=Machine Learning; SMO=Sequential Minimal Optimization; LR=Logistic Regression; MKL=Multiple Kernel Learning; D=Decision; F=Feature; MF=Multisense Framework

emotions, which represents the first explicit attempt to bridge sentiment analysis and emotion recognition.

3. Available datasets

As the primary purpose of this paper is to inform readers on recent advances in multimodal affect recognition, in this section we describe widely-used datasets for multimodal emotion recognition and sentiment analysis. However, we do not cover unimodal datasets, for example facial expression recognition from image datasets (e.g., CK++), as they are outwith the scope of the paper.

In the literature, we found two main methodologies for dataset collection: natural videos and video recordings of subjects acting based on pre-decided scripts. To curate the latter, subjects were provided affect-related scripts and asked to act. It is observed that such datasets can suffer from inaccurate actions by subjects, leading to corrupted samples or inaccurate information for the training dataset. According to D'Mello et al. [8], even though, in literature, a multimodal framework achieved performance improvement over unimodal systems, improvement was much lower when it was trained on natural data (4.59% improvement) versus acted data (12.7% improvement).

There are several other drawbacks associated with this method, e.g., the time taken to create the dataset and biased labeling. Due to these problems, a model trained on these datasets may suffer from poor generalization capability. To overcome such problems, Morency et al. [6] proposed a method of dataset collection in which the product review videos were crawled from popular social websites and later labeled with emotion and sentiment labels.

A common feature amongst both approaches is that they are labeled at utterance level, i.e., for each utterance there is an associated emotion or sentiment label. Utterance-level labeling scheme is particularly important to track the emotion and sentiment dynamics of the subject's mindset in a video.

3.1. Datasets for multimodal sentiment analysis

Available datasets for multimodal sentiment analysis have mostly been collected from product reviews available on different online video sharing platforms, e.g., YouTube. The publicly available multimodal emotion recognition and sentiment analysis datasets are summarized in Tables 1 and 2, respectively.

YouTube Dataset. This dataset was developed in 2011 by Morency et al. [6]. The idea behind its development is to capture the data present in the increasing number of videos posted online every day. The authors take pride in developing the first publicly available dataset for tri-modal sentiment analysis, by combining visual, audio and textual modalities. The dataset was created by collecting videos from YouTube that are diverse and multimodal and have ambient noises. The keywords used for the collection of videos are opinion, review, best perfume, tooth paste, business, war, job, I hate and I like. Finally, a dataset of 47 videos was created, out of which 20 were from female speakers and the rest male, with their ages ranging from 14 to 60 years. All speakers expressed their views in English and belonged to different cultures. The videos were set to .mp4 format with a size of 360x480. The 47 videos in the dataset were further annotated with one of three sentiment labels: positive, negative or neutral. This annotation task led to 13 positively, 12 negatively and 22 neutrally labeled videos. For qualitative and statistical analysis of the dataset, the authors used polarized words in text, 'smile' and 'look away' in visual, and pauses and pitch in aural modality, as the main features.

MOUD Dataset. The Multimodal Opinion Utterances Dataset (MOUD) was developed in 2013 by Perez-Rosas et al. [26]. This is a dataset of utterances, with all videos recorded in Spanish. A final set of 80 videos was selected, out of which 65 were from female speakers and 15 from male speakers, with age ranging from 20 to 60 years. A multimodal dataset of 498 utterances was eventually created with an average duration of 5 seconds and a standard deviation of 1.2 seconds. The dataset was annotated using Elan, an annotator tool used for video and audio sources, along with two other annotators. The annotation task led to 182 positive, 231 negative and 85 neutral labeled utterances. There were 28 features considered for computation in total, including: prosody features, energy features, voice probabilities and spectral features. This trimodality dataset is said to produce an error rate reduction of 10.5% compared to the best unimodality set. The authors also experimentally demonstrated an interesting fact, that a 'istressed brow' is the strongest feature for segment classification, with a smile being a close second.

ICT-MMMO Database. The Institute for Creative Technologies Multi-Modal Movie Opinion (ICT-MMMO) database was developed in 2013 by Wollmer et al. [27]. This dataset is a collection of online videos obtained from YouTube and ExpoTV reviewing movies in English. The authors used keywords such as movie, review,

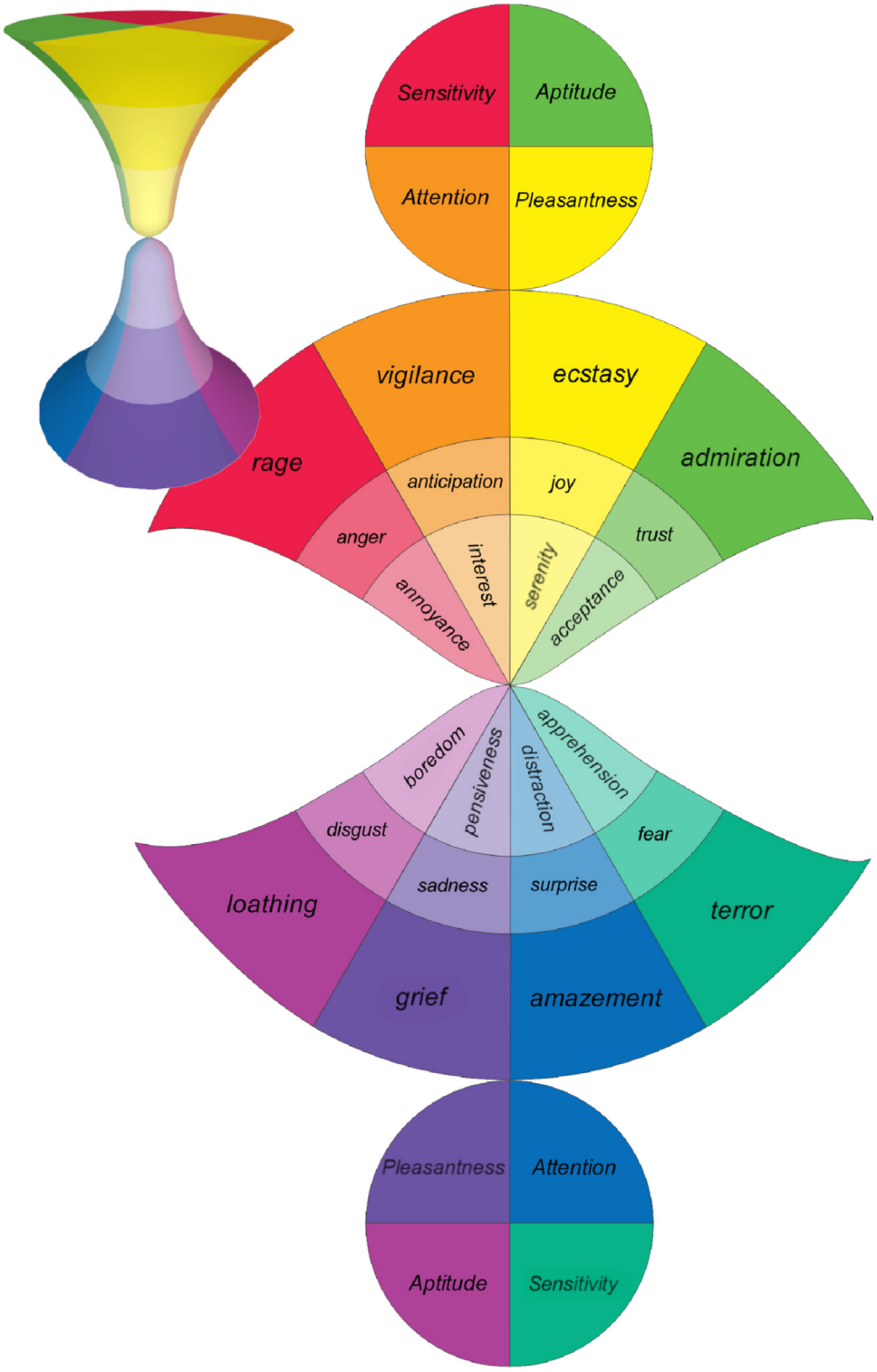


Fig. 3. The Hourglass of Emotions.

videos and opinions, and the names of recent movies as listed by imdb.com, as search keywords. The authors collected 308 YouTube videos, out of which 228 were annotated as positive, 57 as negative and 23 as neutral. They also gathered 78 movie review videos from ExpoTV, from which 62 were annotated as negative, 14 as neutral and 2 as positive. The final dataset comprised a total of 370 videos, which included all 308 videos from YouTube and 62 negative movie review videos from ExpoTV. The annotation task was performed by two annotators for YouTube videos and one annotator for ExpoTV videos. In contrast with other datasets, this dataset had five sentiment labels: strongly positive, weakly positive, neutral, strongly negative and weakly negative.

3.2. Datasets for multimodal emotion recognition

We describe the datasets currently available for multimodal emotion recognition below. To the best of our knowledge, all available datasets for multimodal emotion recognition are acted.

HUMAINE Database. This dataset was developed in 2007 by Douglas-Cowie et al. [28]. The database provides naturalistic clips of pervasive emotions from multiple modalities and labels the best ones describing them. It consists of 50 clips from both naturalistic and induced data, spanning a broad emotional space, covering cues from body gestures, face, voice and words and representing speakers from different genders and cultures. Labels describing both signs and emotional content are designed to be time aligned rather than global, as timing appears to be an important factor in many areas.

The Belfast Database. This dataset was developed in 2000 by Douglas-Cowie et al. [29]. The database consists of audiovisual data of people discussing emotional subjects and are taken from TV chat shows and religious programs. It comprises 100 speakers and 239 clips, with 1 neutral and 1 emotional clip for each speaker. Two types of descriptors were provided for each clip – dimensional and categorical. Activation and evaluation are dimensions that are known to discriminate effectively between emotional states. Activation values indicate the dynamics of a state and evaluation values provide a global indication of the positive or negative feelings associated with the emotional state. Categorical labels describe the emotional content of each state.

The SEMAINE Database. This dataset was developed in 2007 by McKeown et al. [30]. It is a large audiovisual database created for building agents that can engage a person in a sustained and emotional conversation using a Sensitive Artificial Listener (SAL) [31] paradigm. SAL is an interaction involving two parties: a 'human' and an 'operator' (either machine or a person simulating a machine). The interaction is based on two qualities: one is low sensitivity to preceding verbal context (the words the user used that do not dictate whether to continue the conversation) and the second is conduciveness (response to a phrase by continuing the conversation). There were 150 participants, 959 conversations, each lasting 5 minutes. There were 6-8 annotators per clip, who eventually traced 5 affective dimensions and 27 associated categories. For the recordings, the participants were asked to talk in turn to four emotionally stereotyped characters. The characters are Prudence, who is even-tempered and sensible; Poppy, who is happy and outgoing; Spike, who is angry and confrontational; and Obadiah, who is sad and depressive. Videos were recorded at 49.979 frames per second at a spatial resolution of 780 x 580 pixels and 8 bits per sample, while audio was recorded at 48 kHz with 24 bits per sample. To accommodate research in audio-visual fusion, the audio and video signals were synchronized with an accuracy of 25micro-seconds.

Interactive Emotional Dyadic Motion Capture Database (IEMOCAP). IEMOCAP dataset was developed in 2008 by Busso et al. [32]. 10 actors were asked to record their facial expressions in front of cameras. Facial markers, and head and hand gesture trackers were applied in order to collect facial expressions, and head and hand gestures. In particular, the dataset contains a total of 10 hours recording of dyadic sessions, each of them expressing one of the following emotions: happiness, anger, sadness, frustration and neutral state. The recorded dyadic sessions were later manually segmented at utterance level (defined as continuous segments when one of the actors was actively speaking). The acting was based on some scripts, hence, it was easy to segment the dialogs for utterance detection in the textual part of the recordings. Busso et al. [32] used two famous emotion taxonomies in order to manually label the dataset at utterance level: discrete categorical-based annotations (i.e., labels such as happiness, anger, and sadness), and continuous attribute-based annotations (i.e., activation, valence and dominance). To assess the emotion categories of the recordings, six human evaluators were appointed. Having two different annotation schemes can provide complementary information in human-machine interaction systems. The evaluation sessions were organized so that three different evaluators assessed each utterance. Self-assessment manikins (SAMs) were also employed to evaluate the corpus in terms of the attributes valence [1-negative, 5-positive], activation [1-calm, 5-excited], and dominance [1-weak, 5-strong]. Two more human evaluators were asked to estimate the emotional content in recordings using the SAM system. These two types of emotional descriptors facilitate the complementary insights about the emotional expressions of humans, emotional communications between people which can further help develop better human-machine interfaces by automatically recognizing and synthesizing emotional cues expressed by humans.

The eNTERFACE database. This dataset was developed in 2006 by Martin et al. [33]. It is an audiovisual developed for use as a reference database for testing and evaluating video, audio or joint audio-visual emotion recognition algorithms. This database elicited universal emotions of happiness, sadness, surprise, anger, disgust and fear with the help of 42 speakers, from 14 different nationalities.

4. Unimodal features for affect recognition

Unimodal systems act as the primary building blocks for a well-performing multimodal framework. In this section, we describe the literature of unimodal affect analysis primarily focusing on visual, audio and textual modalities. The following section focuses on multimodal fusion. This particularly benefits the readers as they can refer to this section for unimodal affect analysis literature while the following section will inform on how to integrate the output of unimodal systems, with the final goal of developing a multimodal affect analysis framework.

4.1. Visual modality

Facial expressions are primary cues for understanding emotions and sentiments. Across the ages of people involved, and the nature of conversations, facial expressions are the primary channel for forming an impression of the subject's present state of mind. Ekman et al. [11], considered pioneers in this research, argued that it is possible to detect six basic emotions, e.g., Anger, Joy, Sadness, Disgust and Surprise from cues of facial expressions. In this section, we present various studies on the use of visual features for multimodal affect analysis.

4.1.1. Facial action coding system

As facial cues gained traction in discerning emotions, a number of observer-based measurement systems for facial expressions were developed [34–36]. Out of these systems, the Facial Action Coding System (FACS) developed by Ekman and Friesen [34] has been widely used. FACS is based on the reconstruction of facial expressions in terms of Action Units (AU). The facial muscles of all humans are almost identical and AUs are based on movements of these muscles, which consist of three basic parts: AU number, FACS name, and muscular basis. FACS only distinguishes facial actions and gives no inference on emotions. FACS codes are used to infer emotions using a variety of available resources such as FACS Investigators' Guide [37], the FACS interpretive database [38], and a large body of empirical research [39].

These resources use combinations of AUs for specifying emotions. In 1992, the seventh emotion 'contempt' was added to the universal set of emotions [40], as it expresses *disrespect* which is equally important when compared to the six basic emotions. In

2002, an updated version of FACS was introduced where the description of each AU, and AU combinations were refined. Furthermore, details on head movements and eye positions were also added [37]. In addition to emotion and sentiment analysis, FACS is also used in the field of neuroscience [41], computer vision [42], computer graphics [43] and animation [44], and face encoding for digital signal processing [45].

Ekman's work inspired many researchers to employ image and video processing methods in order to analyze facial expressions. Yacoob et al. [46] and Black et al. [47] used high gradient points on the face, and tracked head and facial movements to recognize facial expressions. Geometrical features [48] with a multi-scale, multi-orientation Gabor Wavelet-based representation was used to identify expressions. Kalman Filter and probabilistic principal component analysis (PCA) [49] was used to track the pupils, in order to enhance the features. A stochastic gradient descent based technique [50] and active appearance model (AAM) [51] were used to recover the face shape and texture parameters, for facial features.

A comparison of several techniques [52], such as optical flow, PCA, independent component analysis (ICA), local feature analysis and Gabor wavelet, for recognition of action units, found that, Gabor wavelet representation and ICA performed better on most datasets. Considering every part of the face as an important feature, a multi-state face component model [53], was introduced to exploit both permanent and transient features. Permanent features are those which remain the same through ages, which include opening and closing of lips and eyes, pupil location, eyebrows and cheek areas.

Transient features are observed only at the time of facial expressions, such as contraction of the corrugator muscle that produces vertical furrows between the eyebrows. Texture features of the face have also been considered for facial expression analysis in a number of feature extraction methods, including: image intensity [42], image difference [54], edge detection [53], and Gabor wavelets [55]. In order to recognize and model facial expressions in terms of emotions and sentiments, numerous classifiers have been used, such as Nearest Neighbor [54], Neural Networks [53], support vector machine (SVM) [56], Bayesian Networks [57], and AdaBoost classifiers [58].

4.1.2. Main facial expression recognition techniques

Some of the important facial expression recognition techniques, face tracking and feature extraction methods are briefly described below:

Active Appearance Models (AAM) [59] are well-known algorithms for modeling deformable objects. The models decouple the shape and texture of objects, using a gradient-based model fitting approach. Most popular applications of AAM include recognition, tracking, segmentation and synthesis.

Optical flow models [46] are used to calculate the motion of the objects or the motion of two image frames, based on gradients. These methods are also termed differential methods as they are calculated using Taylor series.

Active Shape Models (ASM) [60] are statistical models that deform to fit the data or object in an image in ways consistent with the training data provided. These models are used mainly to enhance automatic analysis of images under noisy or cluttered environments.

3D Morphable Models (3DMM) [61] are models that are used for facial feature analysis by modeling 3D faces, that are immune to pose and illumination. Thus, these models are used for automatic 3D face registration by computing dense one-to-one correspondences, and adjusting the naturalness of modeled faces.

Muscle-based models [62] are models that consist of facial feature points corresponding to facial muscles, for tracking motion of facial components, such as eyebrows, eyes and mouth, thus recognizing facial expressions.

3D wireframe models [63] are 3-dimensional models of an object where the edges or vertices are connected using straight lines or curves. Once the model is designed for a given face, the head motion and local deformations of the facial features such as eyebrows, eyes and mouth can be tracked.

Elastic net model [64] represents facial expressions as motion vectors of the deformed net, from a facial edge image.

Geometry-based shape models [65] are models that represent expression changes in a face through geometry-based high-dimensional 2D shape transformations, which are then used to register regions of a face with expressions, to those defined on the template face.

3D Constrained Local Model (CLM-Z) [66] is a non-rigid face tracking model used to track facial features under varying poses, by including both depth and intensity information. Non-rigid face tracking refers to points of interest in an image, for example, nose tip, corners of eyes and lips. The CLM-Z model can be described by the parameters $p = [s, R, q, t]$, where s is a scale factor, R is object rotation, t represents 2D translation and q is the vector describing non-rigid variation of the q.

Generalized Adaptive View-based Appearance Model (GAVAM) [67] is a probabilistic framework combining dynamic or motion-based approaches to track the position and orientation of the head through video sequences, and employs static user-independent approaches to detect head pose from an image. GAVAM is considered a high-precision, user-independent real-time head pose tracking algorithm.

In other works, the CLM-Z and GAVAM models were integrated for rigid and non-rigid facial tracking to improve pose estimation accuracy for both 2D and 3D cases [66].

4.1.3. Extracting temporal features from videos

Although we have addressed some of the key works on recognizing facial expressions from images, most of those methods do not work well for videos as they do not model temporal information. In this paragraph, we discuss a few methods which used temporal information [56,68–70], Motion-Units (MU) (otherwise called facial motion) [63] and features in terms of duration, content and valence [71] for affect recognition from videos.

An important facet in video-based methods is maintaining accurate tracking throughout the video sequence. A wide range of deformable models, such as muscle-based models [62], 3D wireframe modesl [63], elastic net models [64] and geometry-based shape models [65,72], have been used to track facial features in videos. Thus, deformable models have demonstrated an improvement in both facial tracking and facial expression analysis accuracy, [73]. Following this, many automatic methods for detection of facial features and facial expressions were proposed [74–76], both image-based and video-based.

4.1.4. Body gestures

Though most research works have concentrated on facial feature extraction for emotion and sentiment analysis, there are some contributions based on features extracted from body gestures. Research in psychology suggests that body gestures provide a significant source of features for emotion and sentiment recognition. In [77], a detailed study was carried out on how body gestures are related to emotions and how various combinations of body gesture dimensions and qualities can be found in different emotions. It was also shown how basic emotions can be automatically distinguished from simple statistical measures of motion's dynamics induced by body movements [78].

Based on these groundbreaking studies, a set of body gesture features for emotion recognition were extracted to help autistic

children [79]. Inspired by these pioneering findings, an automatic emotion recognition framework was proposed from body gestures, using a set of postural, kinematic, and geometrical features extracted from sequences of 3D skeletal movements, which were fed to a multiclass SVM classifier for emotion recognition [80].

In [81], a mathematical model was developed to analyze the dynamics of body gestures for emotion expressiveness. Some of the extracted motion cues to understand the subject's temporal profile included: initial and final slope of the main peak, ratio between the maximum value and the duration of the main peak, ratio between the absolute maximum and the biggest following relative maximum, centroid of the energy, symmetry index, shift index of the main peak, and number of peaks.

In [82], both facial and hand gesture features were used to perform emotion analysis and the creation of moving skin masks in order to estimate user's movement by tracking the centroid of skin masks over the person under experimentation.

4.1.5. New era: deep learning to extract visual features

In the last two sections, we described the use of handcrafted feature extraction from a visual modality and mathematical models for facial expression analysis. With the advent of deep learning, we can now extract features automatically without prior intervention. The deep learning framework enables robust and accurate feature learning, which in turn produces benchmark performance on a range of applications, including digit recognition [83], image classification [84], feature learning [85], visual recognition [86], musical signal processing [87] and NLP [88]. Both academia and industries have invested a huge amount of effort in building powerful deep neural networks. These demonstrate the potential of deep learning to develop robust features, in both supervised and unsupervised settings. Even though deep neural networks may be trapped in local optima [89], different optimization techniques can be effectively employed to enhance their performance in many challenging fields. Inspired by the recent success of deep learning, emotion and sentiment analysis tasks have also been enhanced by the adoption of deep learning algorithms, e.g., convolutional neural network (CNN).

In [90], a novel visual sentiment prediction framework was designed to understand images using CNN. The framework is based on transfer learning from a CNN pre-trained on large scale data for object recognition, which in turn is used for sentiment prediction. The main advantage of the proposed framework is that there is no requirement of domain knowledge for visual sentiment prediction.

Motivated by the need for processing increasingly large and noisy data in the field of image sentiment analysis, CNN has been employed in [91], coupled with a progressive strategy to fine tune deep learning networks to filter out noisy training data and use of domain transfer learning to enhance performance.

In [92], emotion recognition for user generated videos is performed through the extraction of deep convolution network features and through zero-shot emotion learning, a method that predicts emotions not observed in the training set. To implement this task, image transfer encoding (ITE) is proposed to encode the extracted features and generate video representations.

More recently, deep 3D convolutional networks (C3D) (Fig. 4) have been proposed for spatio-temporal feature learning [93]. The C3D network comprises 8 convolution layers, 5 pooling layers, 2 fully connected layers, and a softmax output layer. The network has been shown to be more amenable for spatio-temporal feature learning, in comparison with 2D convolution networks. The 3x3 convolution kernels in all layers were found to create the best performing architecture, with learned features using a simple classifier outperforming existing state-of-the-art methods.

Poria et al. [94] have developed a convolutional recurrent neural network (RNN) to extract visual features. In their study, a CNN and RNN have been stacked and trained together (Fig. 5). On both multimodal sentiment analysis and emotion recognition datasets, their approach outperformed the state of the art.

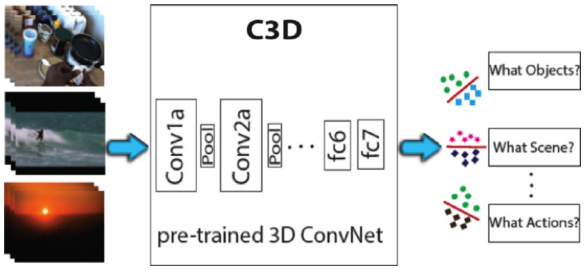


Fig. 4. C3D for extracting spatio-temporal generic video features.

4.2. Audio modality

Similar to text and visual feature analysis, emotion and sentiment analysis through audio features has specific components. Several prosodic and acoustic features have been used in the literature to teach machines how to detect emotions [95–98]. Since emotional characteristics are more prominent in prosodic features, these features are widely used in the literature [99,100].

Researchers started targeting affective reactions to everyday sounds [101], which have ultimately led to enormous applications to date, both in unimodal and multimodal analysis. The current trend is to understand affect in naturalistic videos [102–104], e.g., spontaneous dialogs, audio recordings collected in call centers, interviews, etc. Early research on extraction of audio features focused on the phonetic and acoustic properties of spoken language. With the help of psychological studies related to emotion, it was found that vocal parameters, especially pitch, intensity, speaking rate and voice quality play an important role in recognition of emotion and sentiment analysis [98].

Further studies showed that acoustic parameters change not only through oral variations, but are also dependent on personality traits. Various works have been carried out based on the types of features that are needed for better analysis [105,106]. Researchers have found pitch and energy related features playing a key role in affect recognition. Other features that have been used by some researchers for feature extraction include formants, mel frequency cepstral coefficients (MFCC), pause, teager energy operated based features, log frequency power coefficients (LFPC) and linear prediction cepstral coefficients (LPCC).

Some of the important audio features are described briefly below:

- *Mel Frequency Cepstral Coefficients (MFCC)* are coefficients that collectively form a mel-frequency cepstrum (MFC). The MFC is a short-term power spectrum of a sound or an audio clip, which approximates the human auditory system more closely than any other available linearly-spaced frequency band distribution. This feature is calculated based on the linear cosine transform of a log power spectrum, on a mel-frequency scaling.
- *Spectral centroid* indicates the center of mass of the magnitude spectrum, which simply provides an indication of the brightness of a sound.
- *Spectral flux* defines how quickly the power spectrum of a signal is changing. This feature is usually calculated by taking the Euclidean distance between two normalized spectra.
- *Beat Histogram* is a histogram showing the strength of different rhythmic periodicities in a signal. It is typically calculated by taking the RMS of 256 windows and then taking the FFT of the output.

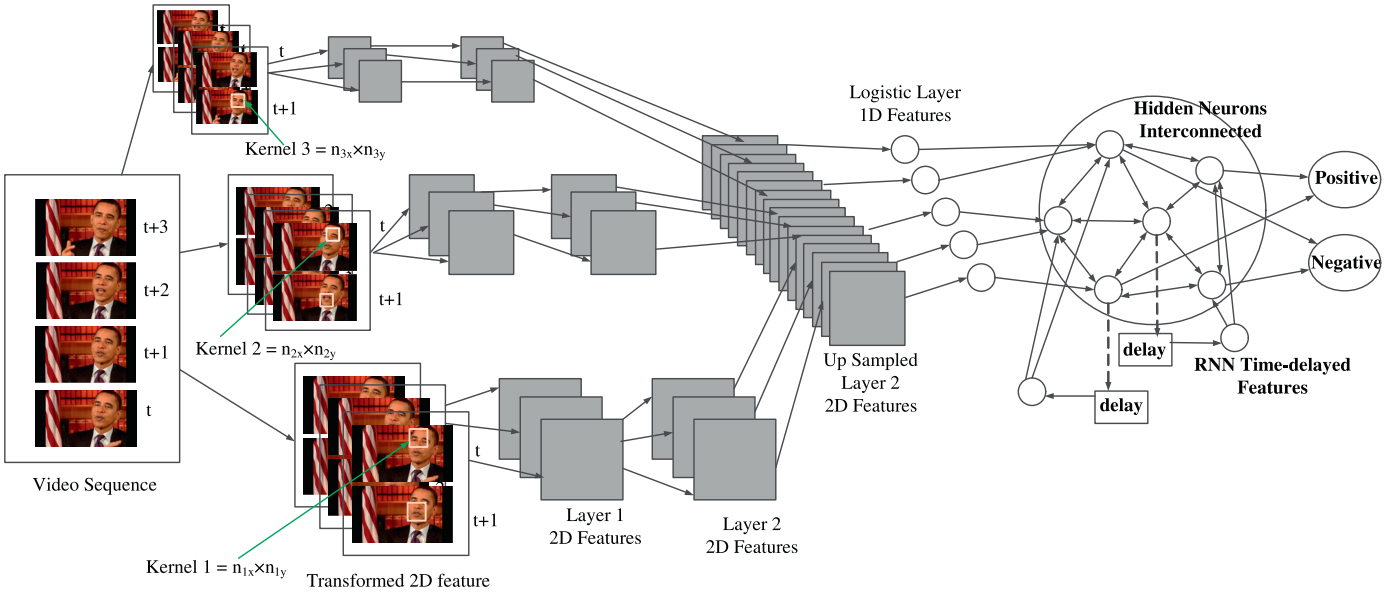


Fig. 5. CNN for visual sentiment recognition as proposed by Poria et al. [94].

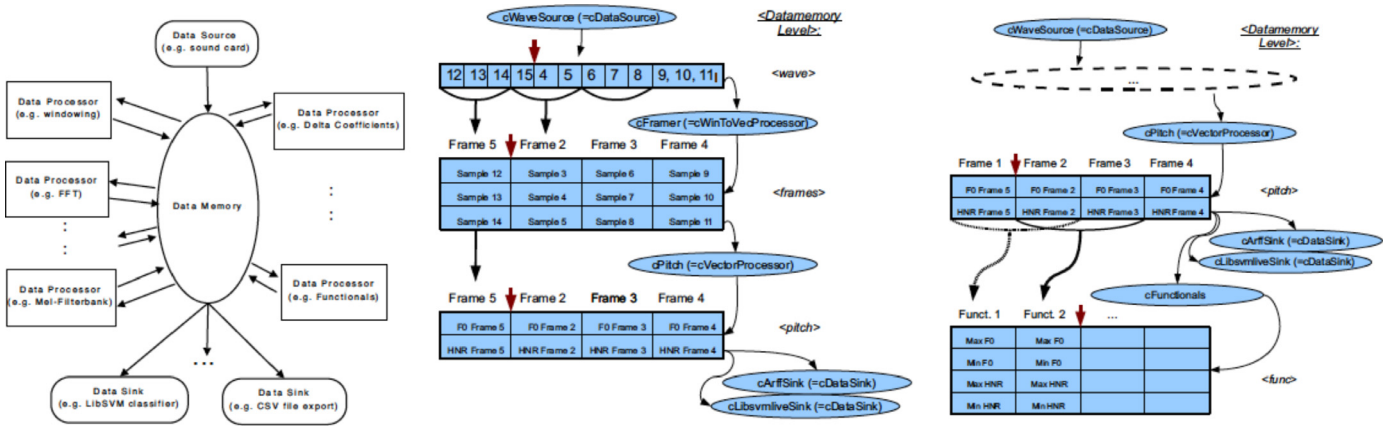


Fig. 6. The OpenSMILE framework.

- *Beat sum* is used to find regular beats in a signal. It is calculated as the sum of all entries in the beat histogram.
- *Strongest beat* is the strongest beat in a signal and is found by identifying the strongest bin the beat histogram.
- *Pause duration* is the time the speaker is silent in an audio segment.
- *Pitch* is the quality of a sound governed by the rate of vibrations producing it; the degree of highness or lowness of a tone.
- *The Perceptual Linear Predictive Coefficients (PLP)* are created from linear prediction coefficients (LPC) by performing perceptual processing before autoregressive modeling and followed by cepstral conversion.

OpenSMILE [107] is a popular audio feature extraction toolkit which is able to extract all the key features as elaborated above. This framework is shown in Fig. 6. The affective reactions to sound have been classified as discrete feeling states and states based on dimensions [101,108]. Discrete feeling states are defined as emotions that are spontaneous, uncontrollable or, in other words, universal emotions. The states based on dimension are hedonic valence (pleasantness), arousal (activation, intensity) and dominance. Recent studies on speech-based emotion analysis [98,109–111] have focused on identifying several acoustic features such as fundamental frequency (pitch), intensity of utterance [112], bandwidth, and duration. The speaker-dependent approach often gives much better results than the speaker-independent approach, as shown by benchmark results of Navas et al. [113], where about 98% accuracy was achieved using the Gaussian mixture model (GMM) as a classifier, with prosodic, voice quality as well as MFCCs employed as speech features. However, the speaker-dependent approach is not feasible in many practical applications that deal with a very large number of users.

4.2.1. Local features vs. Global features

Audio affect classification is also classified into local features and global features. The common approach to analyze audio modality is to segment each audio/utterance into either overlapped or non-overlapped segments and examine them. Within a segment the signal is considered to be stationary. The features extracted from these segments are called local features.

In speech production, there are several utterances and, for each utterance, the audio signal can be divided into several segments. Global features are calculated by measuring several statistics, e.g., average, mean, deviation of the local features. Global features are the most commonly used features in the literature. They are fast to compute and, as they are fewer in number compared to local features, the overall speed of computation is enhanced [114]. However, there are some drawbacks of calculating global features, as

some of them are only useful to detect affect of high arousal, e.g., anger and disgust. For lower arousals, global features are not that effective, e.g., global features are less prominent to distinguish between anger and joy. Global features also lack temporal information and dependence between two segments in an utterance.

4.2.2. Speaker-independent applications

To the best of our knowledge, for speaker-independent applications, the best classification accuracy achieved to-date is 81% [115], obtained on the Berlin Database of Emotional Speech (BDES) [116] 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 [117]. As demonstrated in the analysis by Scherer et al. [118], human ability to recognize emotions from speech audio is about 60%. Their study showed that sadness and anger are detected more easily from speech, while the recognition of joy and fear is less reliable. Caridakis et al. [81] obtained 93.30% and 76.67% accuracy to identify anger and sadness, respectively, from speech, using 377 features based on intensity, pitch, MFCCs, Bark spectral bands, voiced segment characteristics, and pause length.

4.2.3. Audio features extraction using deep networks

As for computer vision, deep learning is also gaining increasing attention in audio classification research. In the context of audio emotion classification, autoencoder followed by a CNN has been used in [119]. Authors trained CNN on the features extracted from all time frames. These types of models are usually incapable of modeling temporal information.

To overcome this problem, Long Short Term Memory (LSTM) [120], and bi-directional LSTM [121] have been commonly used on hand-extracted acoustic features. In computer vision, deep networks are frequently used for automatic feature extraction. A possible research question is whether deep networks can be replicated for automatic feature extraction from aural data. As shown in a pilot study [122], CNN can be used to extract features from audio, which can subsequently be used in a classifier for the final emotion classification task. Generalized discriminant analysis (GerDA) based deep neural networks are also a very popular approach in the literature for automatic feature extraction from raw audio data. However, most deep learning approaches in audio emotion classification literature rely on handcrafted features [123].

Recently, researchers have applied audio emotion and sentiment analysis in many fields, in particular to one of the most active and prominent areas in recent years: human-computer interaction [124,125].

4.3. Textual modality

In this section, we present the state of the art of both emotion recognition and sentiment analysis from text. The task of automatically identifying fine-grained emotions, such as anger, joy, surprise, fear, disgust, and sadness, explicitly or implicitly expressed in text has been addressed by several researchers [126,127]. So far, approaches to text-based emotion and sentiment recognition rely mainly on rule-based techniques, bag of words (BoW) modeling using a large sentiment or emotion lexicon [128], or statistical approaches that assume the availability of a large dataset annotated with polarity or emotion labels [129].

Several supervised and unsupervised classifiers have been built to recognize emotional content in text [130]. The SNoW architecture [131] is one of the most useful frameworks for text-based emotion recognition. In the past decade, researchers have mainly focused on sentiment extraction from texts of different genres, such as news [132], blogs [133], and customer reviews [134] to name a few.

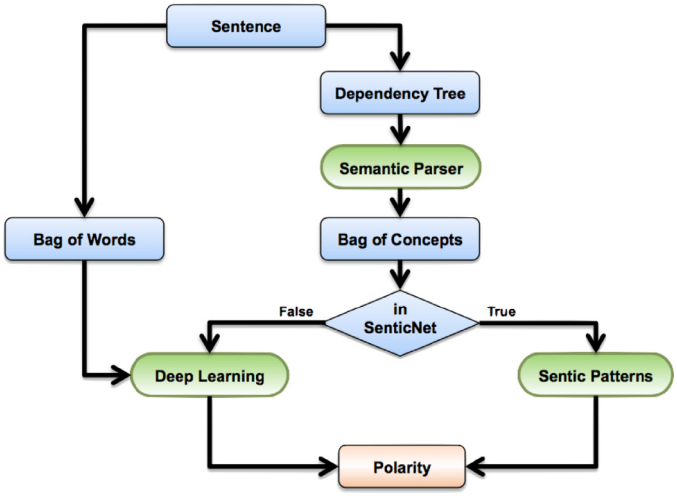


Fig. 7. Sentic computing framework.

Sentiment analysis systems can be broadly categorized into knowledge-based and statistics-based systems [135]. While initially the use of knowledge bases was more popular for the identification of emotions and polarity in text, recently sentiment analysis researchers have been increasingly using statistics-based approaches, with a special focus on supervised statistical methods. For example, Pang et al. [136] compared the performance of different machine learning algorithms on a movie review dataset and obtained 82.90% accuracy using a large number of textual features. A recent approach by Socher et al. [137] obtained even better accuracy (85%) on the same dataset using a recursive neural tensor network (RNTN). Yu and Hatzivassiloglou [138] used semantic orientation of words to identify polarity at sentence level. Melville et al. [139] developed a framework that exploits word-class association information for domain-dependent sentiment analysis.

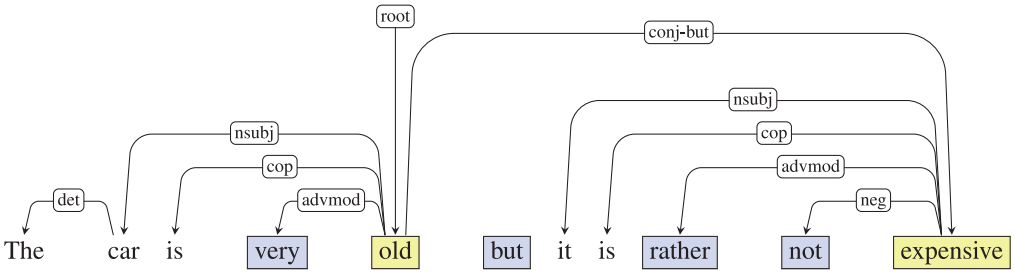
Other unsupervised or knowledge-based approaches to sentiment analysis include: Turney et al. [140], who used seed words to calculate polarity and semantic orientation of phrases; Melville et al. [141], who proposed a mathematical model to extract emotional clues from blogs and used them for sentiment recognition; Gangemi et al. [142], who presented an unsupervised frame-based approach to identify opinion holders and topics; and sentic computing [143], a hybrid approach to sentiment analysis that exploits an ensemble of deep learning, commonsense reasoning, and linguistics to better grasp semantics and sentics (i.e., denotative and connotative information) associated with natural language concepts (Fig. 7).

4.3.1. Single- vs. Cross-domain

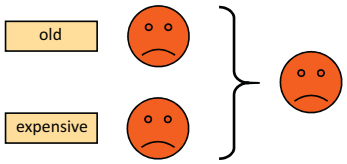
Sentiment analysis research can also be categorized as single-domain [136,140,144] versus cross-domain [145]. The work presented in [146] discusses spectral feature alignment to group domain-specific words from different domains into clusters. Authors first incorporated domain-independent words to aid the clustering process and then exploited the resulting clusters to reduce the gap between domain-specific words of two domains.

Bollegala et al. [147] developed a sentiment-sensitive distributional thesaurus by using labeled training data from a source domain and unlabeled training data from both source and target domains. Sentiment sensitivity was obtained by including documents' sentiment labels into the context vector. At the time of training and testing, this sentiment thesaurus was used to expand the feature vector.

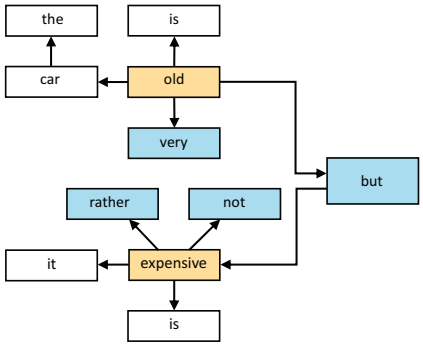
Some recent approaches used SenticNet [148], a domain-independent resource for sentiment analysis containing 50,000



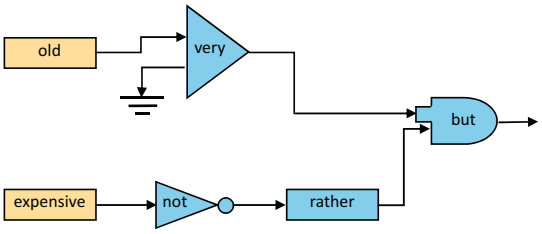
(a) Dependency tree of a sentence.



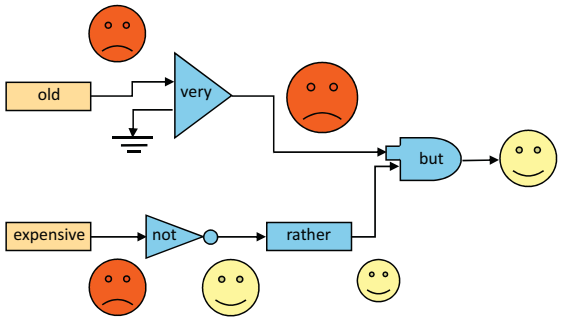
(b) The old way: averaging over a bag of sentiment words. The overall polarity of a sentence is given by the algebraic sum of the polarity values associated with each affect word, divided by the total number of words.



(c) The dependency tree of a sentence resembles an electronic circuit: words shown in blue can be thought as sort of "boolean operations" acting on other words.



(d) The electronic circuit metaphor: sentiment words are "sources" while other words are "elements", e.g., *very* is an amplifier, *not* is a logical complement, *rather* is a resistor, *but* is an OR-like element that gives preference to one of its inputs.



(e) The final sentiment data flow of the "signal" in the "circuit".

Fig. 8. Example of how sentic patterns work on the sentence "The car is very old but it is rather not expensive".

commonsense concepts, for tasks such as opinion holder detection [142], knowledge expansion [149], subjectivity detection [150], event summarization [151], short text message classification [152], sarcasm detection [153], Twitter sentiment classification [154], deception detection [155], user profiling [156], emotion visualization [157], and business intelligence [158].

4.3.2. Use of linguistic patterns

Whilst machine learning methods, for supervised training of the sentiment analysis system, are predominant in literature, a number of unsupervised methods such as linguistic patterns can also be found. In theory, sentence structure is key to carry out sentiment analysis, as a simple change in the word order can flip the polarity of a sentence. Linguistic patterns aim to better understand sentence structure based on its lexical dependency tree, which can be used to calculate sentiment polarity.

In 2014, [159] proposed a novel sentiment analysis framework which incorporates computational intelligence, linguistics, and commonsense computing [160] in an attempt to better understand sentiment orientation and flow in natural language text. Fig. 8 shows how linguistic patterns can function like logic gates in a circuit. One of the earliest works in the study of linguistic patterns for sentiment analysis was carried out by [161], where a corpus and some seed adjective sentiment words were used to find additional sentiment adjectives in the corpus. Their technique exploited a set of linguistic rules on connectives ('and', 'or', 'but', 'either/or', 'neither/nor') to identify sentiment words and their orientations. In this way, they defined the idea of *sentiment consistency*.

Kanayama et al. [144] extended the approach by introducing definitions of intra-sentential (within a sentence) and inter-sentential (between neighboring sentences) sentiment consistency. Negation plays a major role in detecting the polarity of sentences. In [162], Jia et al. carried out an experiment to identify negations in text using linguistic clues and showed a significant performance improvement over the state of the art. However, when negations are implicit, e.g., cannot be recognized by an explicit negation identifier, sarcasm detection needs to be considered as well.

In [163], three conceptual layers, each of which consists of 8 textual features, was proposed to grasp implicit negations. A method to exploit discourse relations for the detection of tweets

polarity was proposed in [164]. The authors showed how conjunctions, connectives, modals and conditionals might affect sentiments in tweets. In [165], a discourse parser combined information of sentential syntax, semantics and lexical information to build a tree that served as a representation of the discourse structure.

Wolf et al. [166] presented a method to represent discourse coherence, using contentful conjunctions to illustrate coherence relations. Discourse coherence relations have also been explored in [167] and, in [168], discourse connectives identification is applied to biomedical text. Liu et al. [169] proposed a collection of opinion rules implying positive or negative sentiment. First, the rules at a conceptual level are described without considering how they may be expressed in actual sentences, i.e., without considering context. Next, an inspection at expression level combines more than one input-constituent expression to derive an overall sentiment orientation for the composite expression. Moilanen et al. [170] introduced the notion of sentiment conflict, which is used when opposite sentiment words occur together, e.g., 'terribly good'. Conflict resolution is achieved by ranking the constituents on the basis of relative weights assigned to them, considering which constituent is more important with respect to sentiment. In [171], a holistic lexicon-based approach was used to evaluate the semantic orientations of opinions expressed on product features in reviews, by exploiting external evidence and linguistic conventions of natural language expressions. This approach, implemented in a system called Opinion Observer, allows for handling opinion words that are context-dependent. The authors found that both aspect and sentiment expressing words are important and proposed using the pair (*aspect, sentiment word*) as an *opinion context*.

In a more recent work, Poria et al. [172] presented the first deep learning approach to aspect-based sentiment analysis. Authors used a 7-layer deep convolutional neural network to tag each word in opinionated sentences as either aspect or non-aspect words. They also developed a set of linguistic patterns for the same purpose and combined them with the neural network. The resulting ensemble classifier, coupled with a word-embedding model for sentiment analysis, allowed their approach to obtain significantly better accuracy than state-of-the-art methods.

4.3.3. Bag of words versus bag of concepts

Text representation is a key task for any text classification framework. BoW looks for surface word forms and does not consider semantic and contextual clues in text. Most of the well-known techniques have focused on BoW representation for text classification [173–175]. To overcome the problem of limited capability in grasping semantic clues, some existing related works relied on using knowledge bases [176,177].

The bag of concepts (BoC) model leverages on representing text as a conceptual vector rather than relying on terms in text. For example, if a text contains "red" and "orange" then, BoC models them as the concept "color", e.g., BoC looks for hyponym. The BoC model was first proposed by Sahlgren et al. [178] to enhance the performance of SVM in text categorization tasks. According to their method, concepts are synonym sets of BoW. Among recent approaches adopting the BoC model, Wang et al. [179] presented the idea of concept as a set of entities in a given domain, e.g., words belonging to similar classes have similar representation.

If a document contains "Jeep" and "Honda" then both of these words can be conceptualized by "Car". On the basis of their study, we identify two major advantages of the BoC model:

- *Replacement of surface matching with semantic similarity:* the BoC model calculates semantic similarity between words and multi-word expressions at a higher level.
- *Tolerance with new terms:* once concepts related to a category are modeled, the BoC model is able to handle new words under that category.

In [180], Zhang et al. discussed semantic classification on a disease corpus. Though their approach does not focus on the BoC model, they attempted to capture semantic information from text at a higher level. According to their study, use of contextual semantic features along with the BoW model can be very useful for semantic text classification. Wu et al. [181] built a sentiment lexicon using a commonsense knowledge base. Under the hypothesis that concepts pass their sentiment intensity to neighbors based on the relations connecting them, they constructed an enriched sentiment lexicon able to perform better on sentiment polarity classification tasks.

Concept-based approaches to sentiment analysis focus on a semantic analysis of text through the use of web ontologies or semantic networks, which allow the aggregation of conceptual and affective information associated with natural language opinions. By relying on large semantic knowledge bases, such approaches step away from the blind use of keywords or word co-occurrence counts and, instead, rely on implicit features associated with natural language concepts [182]. Unlike syntactical techniques, concept-based approaches are also able to detect sentiments that are expressed in a subtle manner, e.g., through the analysis of concepts that do not explicitly convey any emotion, but are implicitly linked to other concepts that do so.

The analysis at concept level is intended to infer the semantic and affective information associated with natural language opinions and, hence, to enable a comparative fine-grained aspect-based sentiment analysis. Rather than gathering isolated opinions about a whole item (e.g., iPhone7), users are generally more interested in comparing different products according to their specific features (e.g., iPhone7's vs Galaxy S7's touchscreen), or even sub-features (e.g., fragility of iPhone7's vs Galaxy S7's touchscreen). In this context, the construction of comprehensive common and commonsense knowledge bases is key for aspect extraction and polarity detection, respectively. Commonsense, in particular, is necessary to appropriately deconstruct natural language text into sentiments; for example, to appraise the concept 'small room' as negative and 'small queue' as positive, or the concept 'go read the book' as positive for a book review but negative for a movie review.

4.3.4. Contextual subjectivity

Wilson et al. [183] reported that, although a word or phrase in a lexicon is marked positive or negative, in the context of the sentence expression it may have no sentiment or even have opposite sentiment.

In their work, subjective expressions were first labeled, with the goal of the work aimed at classifying the contextual sentiment of the given expressions. The authors employed a supervised learning approach based on two steps: first, it determined whether the expression is subjective or objective, second it determined whether the subjective expression is positive, negative, or neutral.

In [184], authors presented an analysis of opinions based on a lexical semantic analysis of a wide class of expressions coupled together, inspecting how clauses involving these expressions were related to each other within a discourse. Narayanan et al. [185] aimed to analyze the sentiment polarity of conditional sentences, studying the linguistic structure of such sentences and applying supervised learning models for dynamic classification.

4.3.5. New era of NLP: emergence of deep learning

Deep-learning architectures and algorithms have already made impressive advances in fields such as computer vision and pattern recognition. Following this trend, recent NLP research is now increasingly focusing on the use of new deep learning methods. As

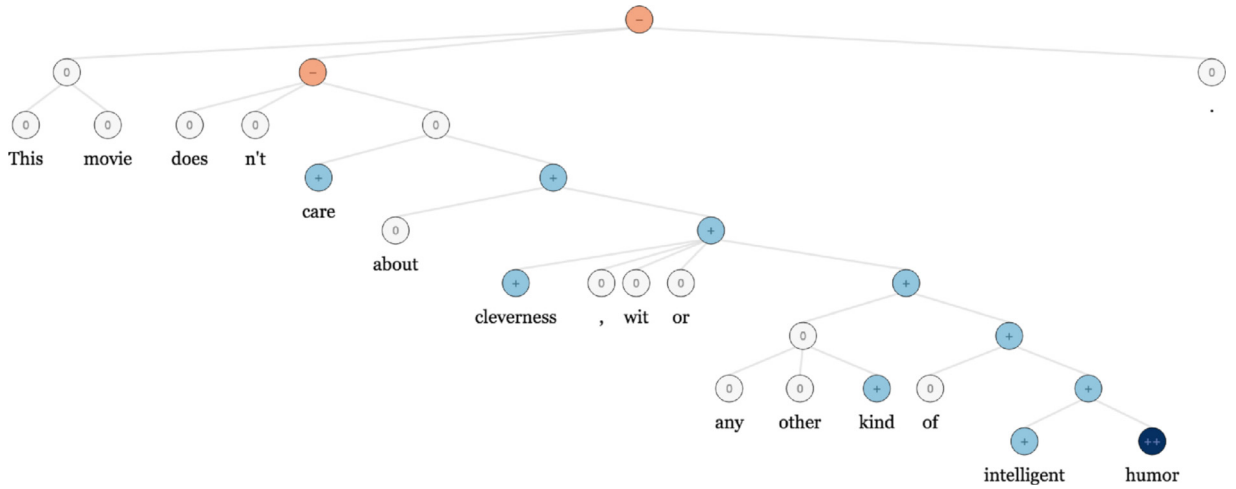


Fig. 9. RNTN applied on the dependency tree of the sentence "This movie doesn't care about cleverness, wit or any other kind of intelligent humor".

demonstrated in [186], a simple deep learning framework outperforms most state-of-the-art approaches, in several NLP tasks such as named entity recognition (NER), sequential role labeling (SRL), and part of speech (POS) tagging. Alternative approaches have exploited the fact that many short n-grams are neutral while longer phrases are well distributed among positive and negative subjective sentence classes. Thus, matrix representations for long phrases and matrix multiplication to model composition, are also being used to evaluate sentiment.

In such models, sentence composition is modeled using deep neural networks such as recursive auto-associated memories [187,188]. Recursive neural networks predict the sentiment class at each node in the parse tree and attempt to capture the negation and its scope in the entire sentence. In the standard configuration, each word is represented as a vector and it is first determined which parent has already computed its children. Next, the parent is computed via a composition function over child nodes, which depends on words being combined and, hence, is linguistically motivated. However, the number of possible composition functions is exponential, hence, in [137], a RNTN was introduced (Fig. 9), which uses a single tensor composition function to define multiple bilinear dependencies between words.

More recently, a new trend has emerged [189,190] focusing on the use of word embeddings pre-trained on a large corpus [191]. In such methods, word vectors are typically concatenated to form a sentence or document vector and then fed to a deep network for training. Studies show that these methods outperform state-of-the-art feature extraction based opinion mining methods, thus establishing themselves as new state-of-the-art benchmarks [190].

5. Multimodal affect recognition

Multimodal affect analysis has already created a lot of buzz in the field of affective computing. This field has now become equally important and popular among the computer scientists [9]. In the previous section, we have discussed state-of-the-art methods which used either of the Visual, Audio or Text modalities for affect recognition. In this section, we discuss the approaches to solve the multimodal affect recognition problem.

5.1. Information fusion techniques

Multimodal affect recognition can be seen as the fusion of information from different modalities. Multimodal fusion is the process of combining data collected from various modalities for analysis tasks. It has gained increasing attention from researchers in diverse fields, due to its potential for innumerable applications, including but not limited to: sentiment analysis, emotion recognition, semantic concept detection, event detection, human tracking, image segmentation, video classification, etc. The fusion of multimodal data can provide surplus information with an increase in accuracy [8] of the overall result or decision. As the data collected from modalities comes in various forms, it is also necessary to consider the period of multimodal fusion in different levels. To date, there are mainly two levels or types of fusion studied by researchers: feature-level fusion or early fusion, and decision-level fusion or late fusion. These have also been employed by some researchers as part of a hybrid fusion approach. Furthermore, there is 'model-level fusion', a type of multimodal fusion designed by researchers as per their application requirements.

Feature-level or early fusion [4,201,205,209,210] fuses the features extracted from various modalities such as visual features, text features, audio features, etc., as a general feature vector and the combined features are sent for analysis. The advantage of feature-level fusion is that the correlation between various multimodal features at an early stage can potentially provide better task accomplishment. The disadvantage of this fusion process is time synchronization, as the features obtained belong to diverse modalities and can differ widely in many aspects, so before the fusion process takes place, the features must be brought into the same format.

Decision-level or late fusion. [202,206,208,273,274] In this fusion process, the features of each modality are examined and classified independently and the results are fused as a decision vector to obtain the final decision. The advantage of decision-level fusion is that, the fusion of decisions obtained from various modalities becomes easy compared to feature-level fusion, since the decisions resulting from multiple modalities usually have the same form of data. Another advantage of this fusion process is that, every modality can utilize its best suitable classifier or model to learn its features. As different classifiers are used for the analysis task, the learning process of all these classifiers at the decision-level fusion stage, becomes tedious and time consuming. Our survey of fusion methods used to date has shown that, more recently, researchers have tended to prefer decision-level fusion over feature-level fusion. Among the notable decision-level fusion methods, Kalman filter has been in [274] as method to fuse classifiers. They considered video as a time dynamics or series and the prediction scores (between 0 and 1) of the base classifiers were fused using Kalman filter. On the other hand, Dobrivsek et al. [273] employed weight sum and weighted product rule for fusion. On the eNTERFACE dataset

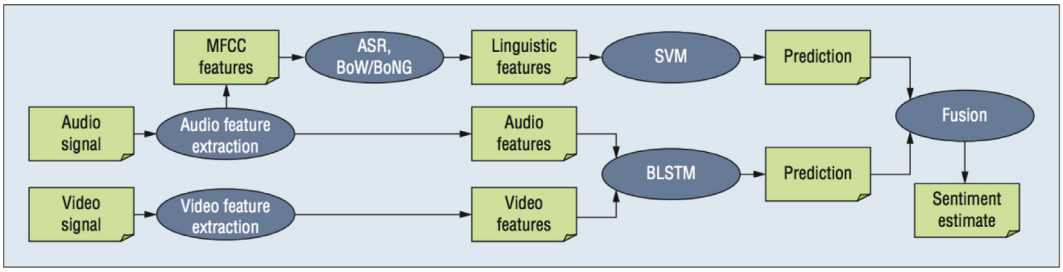


Fig. 10. Hybrid fusion for multimodal sentiment analysis in YouTube videos as proposed by [27].

weighted product (accuracy: 77.20%) rule gave better result than weighted sum approach (accuracy: 75.90%).

Hybrid multimodal fusion. [27,189,250,279] This type of fusion is the combination of both feature-level and decision-level fusion methods. In an attempt to exploit the advantages of both feature and decision-level fusion strategies and overcome the disadvantages of each, researchers opt for hybrid fusion. One such hybrid fusion proposed by Wollmer et al. [27] is shown in Fig. 10. As we can see in Fig. 10, in their method, audio and visual features were fused using BLSTM at feature level. The result of that fusion were then fused with the prediction of the textual classifier using decision-level fusion.

Model-level fusion. [265–267,272,277] It is a technique that uses the correlation between data observed under different modalities, with a relaxed fusion of the data. Researchers built models satisfying their research needs and the problem space. Song et al. [280] used a tripled Hidden Markov Model (HMM) to model the correlation properties of three component HMM's based on audio-visual streams. Zeng et al. [281] proposed a Multi-stream Fused Hidden Markov Model (MFHMM) for audio-visual affect recognition. The MFHMM builds an optimal connection between various streams based on maximum entropy and maximum mutual information principle. Caridakis et al. [223] and Petridis et al. [282] proposed neural networks to combine audio and visual modalities for emotion recognition. Sebe et al. [221] proposed the Bayesian network topology to recognize emotions from audio-visual modalities, by combining the two modalities in a probabilistic manner. According to Atrey et al. [283], fusion can be classified into three categories: rule-based, classification-based and estimation-based methods. The categorization is based on the basic nature of the methods and the problem space, as outlined next.

Rule-based fusion methods. [284,285] As the name suggests, multimodal information is fused by statistical rule based methods such as linear weighted fusion, majority voting and custom-defined rules. The linear weighted fusion method uses sum or product operators to fuse features obtained from different modalities or decision obtained from a classifier. Before the fusion of multimodal information takes place, normalized weights are assigned to every modality under consideration. Thus, the linear weighted fusion method is computationally less expensive compared to other methods, however the weights need to be normalized appropriately for optimal execution. The drawback is that the method is sensitive to outliers. Majority voting fusion is based on the decision obtained by a majority of the classifiers. Custom-defined rules are application specific, in that, the rules are created depending on the information collected from various modalities and the final outcome expected, in order to achieve optimized decisions.

Classification-based fusion methods. [286,287] In this method, a range of classification algorithms are used to classify the multimodal information into pre-defined classes. Various methods used under this category include: SVMs, Bayesian inference, Dempster-Shafer theory, dynamic bayesian networks, neural networks and maximum entropy models. SVM is probably the most widely used supervised learning method for data classification tasks. In this method, input data vectors are classified into predefined learned classes, thus solving the pattern classification problem in view of multimodal fusion. The method is usually applicable for decision-level and hybrid fusion. The Bayesian inference fusion method fuses multimodal information based on rules of probability theory. In this method, the features from various modalities or the decisions obtained from various classifiers are combined and an implication of the joint probability is derived. The Dempster–Shafer evidence theory generalizes Bayesian theory of subjective probability. This theory allows union of classes and also represents both uncertainty and imprecision, through the definition of belief and plausibility functions. The Dempster—Shafer theory is a statistical method and is concerned with fusing independent sets of probability assignments to form a single class, thus relaxing the disadvantage of the Bayesian inference method. The Dynamic Bayesian Network (DBN) is an extension of the Bayesian inference method to a network of graphs, where the nodes represent different modalities and the edges denote their probabilistic dependencies. The DBN is termed by different names in the literature such as Probabilistic Generative Models, graphical models, etc. The advantage of this network over other methods is that the temporal dynamics of multimodal data can easily be integrated. The most popular form of DBN is the Hidden Markov Model (HMM). The Maximum entropy model is a statistical model classifier which follows an information-theoretic approach and provides probability of observed classes. Finally, the other widely used method is Neural Networks. A typical neural network model consists of input, hidden and output nodes or neurons. The input to the network can be features of different modality or decisions from various classifiers. The output provides fusion of data under consideration. The hidden layer of neurons provides activation functions to produce the expected output, and the number of hidden layers and neurons are chosen to obtain the desired accuracy of results. The connections between neurons have specific weights which can be appropriately tuned for the learning process of the neural network, to achieve the target performance accuracy.

Estimation-based fusion methods. [288,289] This category includes kalman filter, extended kalman filter and particle filter based fusion methods. These methods are usually employed to estimate the state of moving object using multimodal information, especially audio and video. The kalman filter is used for real-time dynamic, low-level data and provides state estimates for the system. This model does not require storage of the past of the object under observation, as the model only needs the state estimate of the previous time stamp. However the kalman filter model is restricted to linear systems. Thus, for systems with non-linear characteristics, extended kalman filter is used. Particle filters, also known as Sequential Monte Carlo model, is a simulation-based method used to

obtain the state distribution of non-linear and non-Gaussian state-space models.

5.2. Recent Results

In this section, we describe recent key works in multimodal affect recognition. We summarize state-of-the-art methods, their results and categorize the works based on the datasets described in Section 3.

5.2.1. Multimodal sentiment analysis

MOUD Dataset. The work by Perez et al. [26] focused on multimodal sentiment analysis using the MOUD dataset based on visual, audio and textual modalities. FACS and AUs were used as visual features and openEAR was used for extracting acoustic, prosodic features. Simple unigrams were used for textual feature construction. The combination of these features were then fed to an SVM for fusion and 74.66% accuracy was obtained.

In 2015, Poria et al. [189] proposed a novel method for extraction of features from short texts using a deep CNN. The method was used for detection of sentiment polarity with all three modalities (audio, video and text) under consideration in short video clips of a person uttering a sentence. In this paper, a deep CNN was trained, however, instead of using it as a classifier, values from its hidden layer were used as features for input to a second stage classifier, leading to further improvements in accuracy. The main novelty of this paper was using deep CNN to extract features from text and multiple kernel learning for classification of the multimodal heterogeneous fused feature vectors. For the visual modality, CLM-Z based features were used and openEAR was employed on the audio data for feature extraction.

YouTube Dataset. Morency et al. [6] extracted facial features like smile detection and duration, look away and audio features like pause duration for sentiment analysis on the YouTube dataset. As textual features, two lexicons containing positive and negative words were developed from the MPQA corpus distribution. They fused and fed those features to a Hidden Markov Model (HMM) for final sentiment classification. However, the accuracy was relatively lower (55.33%). Possible future work would be to use more advanced classifiers, such as SVM, CNN, coupled with the use of complex features.

Poria et al. [279] proposed a similar approach where they extracted FCPs using CLM-Z, and used the distances between those FCPs as features. Additionally, they used GAVAM to extract head movement and other rotational features. For audio feature extraction, the state of the art openEAR was employed. Concept-based methods and resources like SenticNet [148] were used to extract textual features. To this end, both feature-level and decision-level fusion were used to obtain the final classification result.

ICT-MMMO Dataset. Wollmer et al. [27] used the same mechanism as [6] for audio-visual feature extraction. In particular, OKAO vision was used to extract visual features which were then fed to CFS for feature selection. In the case of audio feature extraction, they used openEAR. Simple Bag-Of-Words were utilized as text features. Audio-visual features were fed to a Bidirectional-LSTM (BLSTM) for early feature-level fusion and SVM was used to obtain the class label of the textual modality. Finally, the output of BLSTM and SVM were fused at the decision level, using a weighted summing technique.

5.2.2. Multimodal emotion recognition

Recent Works on the SEMAINE Dataset. Gunes et al. [290] used visual aspects which aimed to predict dimensional emotions from spontaneous head gestures. Automatic detection of head nods and shakes was based on 2-dimensional (2D) global head motion estimation. In order to determine the magnitude and direction of the 2D head motion, optical flow was computed between two consecutive frames. It was applied to a refined region (i.e., resized and smoothed) within the detected facial area to exclude irrelevant background information. Directional codewords were generated by the visual feature extraction module, and fed into a HMM for training a nodHMM and a shakeHMM. A Support Vector Machine for Regression (SVR) was used for dimensional emotion prediction from head gestures. The final feature set was scaled in the range [-1, +1]. The parameters of SVR, for each coder-dimension combination, were optimized using 10-fold cross-validation for a subset of the data at hand. The MSE for detection of valence, arousal and other axis was found to be 0.1 on average, as opposed to 0.115 resulting from human annotators.

Valstar et al. [291] focussed on FACS Acton Units detection and intensity estimation, and derived its datasets from SEMAINE and BP4D-Spontaneous database. The training partition (SEMAINE database) consisted of 16 sessions, the development partition had 15 sessions, and the test partition 12 sessions. There were a total of 48,000 images in total in the training partition, 45,000 in development, and 37,695 in testing (130,695 frames in total). For SEMAINE, one-minute segments of the most facially-expressive part of each selected interaction were coded. For the baseline system for this task, two types of features were extracted: two-layer appearance features (Local Binary Gabor Patterns) and geometric features derived from tracked facial point locations, which were then fed into a linear SVM. The average MSE on AU in BP4D datasets was around 0.8, while similar techniques were not applied on SEMAINE. [290,291] took into consideration all the frames of videos, which in turn made the training more time taking.

Nicolaou et al. [292] developed an algorithm for automatically segmenting videos into data frames, in order to show the transition of emotions. To ensure one-to-one correspondence between timestamps of accorder, annotations were binned according to video frames. The crossing over from one emotional state to the other was detected by examining the valence values and identifying the points where the sign changed. The crossovers were then matched across coders. The crossover frame decision was made and the start frame of the video segment decided. The ground truth values for valence were retrieved by incrementing the initial frame number where each crossover was detected by the coders. The procedure of determining combined average values continued until the valence value crossed again to a non-negative valence value. The endpoint of the audio-visual segment was then set to the frame including the offset, after crossing back to a non-negative valence value. Discerning dimensional emotions from head gestures proposed a string-based audiovisual fusion which achieved better results for dimensions valence and expectation as compared to feature-based fusion. This approach added video-based events like facial expression action units, head nods, shakes as ‘words’ to string of acoustic events.

The non-verbal visual events were extracted similar to the unimodal analysis illustrated in [290] (use of nodeHMM and shakeHMM). For detection of facial action units, a local binary patterns descriptor was used and tested on the MMI facial Expression Database. For verbal and non-verbal acoustic events, emotionally relevant keywords derived from automatic speech recognition (ASR) transcripts of SEMAINE, were used. Key words were detected using the multi-stream large vocabulary continuous speech recognition (LVCSR) engine on recognizer’s output, rather than ground truth labels. Finally, a SVR with linear kernel was trained. The event fusion was performed at the string level per segment, by joining all events where more than half of the event overlapped with the segment in a single string. The events could thus be seen as ‘words’. The resulting strings were converted to a feature vec-

tor representation through a binary bag-of-words (BOW) approach. This lead to an average correlation coefficient of 0.70 on Activation, Valence and Intensity which nearly matches human accuracy for the same task.

Recent Works on the HUMAINE Dataset. Chang et al. [293] worked on the vocal part of the HUMAINE dataset information to analyze emotion, mood and mental state, eventually combining it into low footprint C library as AMMON for phones. Sound processing starts with segmenting the audio stream from the microphone into frames with fixed duration (200 ms) and fixed stepping duration (80 ms). The features selected were LLDs (ZCR, RMS, MFCC, etc.) and functions (Mean, SD, skewness). AMMON was developed by extending an ETSI (European Tele-communications Standards Institute) front-end feature extraction library. It included features to describe glottal vibrational cycles which is a promising feature for monitoring depression. They performed a 2-way classification task to separate clips with positive emotions from those with negative emotions. A feature vector was extracted from each clip using AMMON without glottal timings. Finally, the use of SVM with these feature vectors produced 75% accuracy on BELFAST (The naturalistic dataset of HUMAINE).

Castellano et al. [237] aimed to integrate information from facial expressions, body movement, gestures and speech, for recognition of eight basic emotions. The facial features were extracted by generating feature masks, which were then used to extract feature points, comparing them to a neutral frame to produce FAPs as in the previous research. Body tracking was performed using the EyesWeb platform which tracked silhouettes and blobs, extracting motion and fluidity as main expressive cues. The speech feature extraction focused on intensity, pitch, MFCC, BSB and pause length. These were then independently fed into a Bayesian classifier and integrated at decision-level fusion. While the unimodal analysis led to an average of 55% accuracy, feature-level fusion produced a significantly higher accuracy of 78.3%. Decision-level fusion results did not vary much over feature-level fusion.

Another interesting work in [121] aims to present a novel approach to online emotion recognition from visual, speech and text data. For video labeling, temporal information was exploited, which is known to an important issue, i.e., one utterance at time (t) depends on the utterance at time t. The audio features used in the study included: signal energy, pitch, voice quality, MFCC, spectral energy and time signal, which were then modeled using a LSTM. The spoken content knowledge was incorporated at frame level via early fusion, wherein negative keywords were used for activation, and positive for valence. Subsequently, frame-based emotion recognition with unimodal and bimodal feature sets, and turn-based emotion recognition with an acoustic feature set were performed as evaluations. Finally, whilst a SVR was found to outperform a RNN in recognizing activation features, the RNN performed better in recognition of valence from frame-based models. The inclusion of linguistic features produced no monotonic trend in the system.

Recent Works on the eNTERFACE dataset. eNTERFACE dataset is one the most widely used datasets in multimodal emotion recognition. Though in this discussion we mainly focus on multimodalities, we also explain some of the notable unimodal works which have impacted this research field radically.

Among unimodal experiments reported on this dataset, one of the notable works was carried out by Eyben et al. [107]. They pioneered the openEAR, a toolkit to extract speech related features for affect recognition. Several LLDs like Signal Energy, FFT-spectrum, MFCC, Pitch and their functionals were used as features. Multiple Data Sinks were used in the feature extractor, feeding data to different classifiers (K-Nearest Neighbor, Bayes and Support-Vector based classification and regression using the freely available LibSVM). The experiments produced a benchmark accuracy of 75% on the eNTERFACE dataset.

The study by Chetty et al. [238] aimed to develop an audiovisual fusion approach at multiple levels to resolve the misclassification of emotions that occur at unimodal level. The method was tested on two different acted corpora, DaFEx and eNTERFACE. Facial deformation features were identified using singular value decomposition (SVD) values (positive for expansion and negative for contraction) and were used to determine movement of facial regions. Marker-based audio visual features were obtained by dividing the face into several sectors, and making the nose marker the local center for each frame. PCA was used to reduce the number of features per frame to a 10-dimensional vector for each area. LDA optimized SVDF and VDF feature vectors and an SVM classifier was used for evaluating expression quantification, as High, Medium and Low. The unimodal implementation of audio features led to an overall performance accuracy of around 70% on DaFEx and 60% on eNTERFACE corpus, but the sadness-neutral pair and happiness-anger pair were confused significantly. The overall performance accuracy for visual only features was found to be around 82% for the eNTERFACE corpus and only slightly higher on the DaFEx corpus, however, a significant confusion value on neutral-happiness and sadness-anger pairs was found. Audiovisual fusion led to an improvement of 10% on both corpus, significantly decreasing the misclassification probability.

Another attempt [294] at merging audio-visual entities led to 66.5% accuracy on the eNTERFACE dataset (Anger being the highest at 81%). They adopted LBP for facial image representations for facial expression recognition. The process of LBP features extraction generally consists of three steps: first, a facial image is divided into several non-overlapping blocks. Second, LBP histograms are computed for each block. Finally, the block LBP histograms are concatenated into a single vector. As a result, the facial image is represented by the LBP code. For audio features, prosody features like pitch, intensity and quality features like HNR, jitter and MFCC are extracted. These features were fed into a SVM with the radial basis function kernel. While unimodal analysis produced an accuracy of 55% (visual at 63%), multi-modal analysis increased this to 66.51%, demonstrating support for the convergence idea.

While the previous two papers focused on late fusion-based emotion recognition, SAMMI [243] was built to focus on real-time extraction, taking into account low quality videos and noise. A module called 'Dynamic Control' can be used to adapt the various fusion algorithms and content-based concept extractors to the quality of input signals. For example, if sound quality was detected to be low, the relevance of the vocal emotional estimation, with respect to video emotional estimation, was reduced. This was an important step to make the system more reliable and lose some constraints. The visual part was tested on two approaches: (a) Facial FP absolute movements (b) Relative movements of couples of facial FP. For low cost benefits, authors used the Tomasi implementation of the Lukas Kanade (LK) algorithm (embedded in the Intel OpenCV library). The vocal expressions extracted were similar to those reported in other papers (HNR, jiter, intensity, etc.). The features were fed as one second window interval definitions, into two classifiers: SVM and a conventional Neural Network (NN). Finally, SAMMI performed fusion between estimations resulting from the different classifiers or modalities. The output of such a module significantly enhanced the system performance. Since the classification step is computationally efficient with both NN and SVM classifiers, multiple classifiers can be employed at the same time without adversely impacting the system performance. Though the NN was found to improve the CR+ value in fear and sadness, an overall Bayesian network performed equally well with a CR+ of 0.430.

Poria et al. [205] proposed an intelligent multimodal emotion recognition framework that adopts an ensemble feature extraction by exploiting the joint use of text, audio and video features. They trained visual classifier on CK++ dataset, textual classifier on ISEAR dataset and tested on the eNTERFACE dataset. Audio features were extracted using openAudio and cross-validated on the eNTERFACE dataset. Training on the CK++ and ISEAR datasets improved the generalization capability of the corresponding classifier through cross-validated performance on both datasets. Finally, we used feature-level fusion for evaluation and a 87.95% accuracy was achieved, which exceeded all earlier benchmarks.

Recent Works on the IEMOCAP dataset. In multimodal emotion recognition, IEMOCAP dataset is the most popular dataset and numerous works have reported its use as a benchmark. Below, we outline some of the recent key works. A summary and comparison of these studies are shown in Table 5.

Rehman and Busso [296] developed a personalized emotion recognition system using an unsupervised feature adaption scheme by exploiting the audio modality. The OpenSMILE toolkit with the INTERSPEECH 2009 Emotion Challenge feature set was used to extract a set of common acoustic and prosodic features. A linear kernel SVM with sequential minimal optimization (SMO) was used as the emotion detector. The purpose of normalizing acoustic features was to reduce speaker variability, while preserving the discrimination between emotional classes. The iterative feature normalization approach iteratively estimated the normalizing parameters from an unseen speaker. It thus served as a suitable framework for personalized emotion recognition system. In the IFN scheme, an emotion recognition system was used to iteratively identify neutral speech of the unseen speaker. Next, it estimated the normalization parameters using only this subset (relying on the detected labels). These normalization parameters were then applied to the entire data, including the emotional samples. To estimate the performance, the study used leave-one-speaker out, 10-fold cross validation. The results on the IEMOCAP database indicated that the accuracy of the proposed system was 2% (absolute) higher than the one achieved by the baseline, without the feature adaptation scheme. The results on uncontrolled recordings (i.e., speech downloaded from a video-sharing website) revealed that the feature adaptation scheme significantly improved the unweighted and weighted accuracies of the emotion recognition system.

While most papers have focused on audio-visual fusion, Qio Jio [297] reported emotion recognition with acoustic and lexical features. For acoustic features, low-level acoustic features were extracted at frame level on each utterance and used to generate feature representation of the entire dataset, using the OpenSMILE toolkit. The features extracted were grouped into three categories: continuous, qualitative and cepstral. Low-level feature vectors were then turned into a static feature vector. For each emotional utterance, a GMM was built via MAP adaptation using the features extracted in the same utterance. Top 600 words from each of the four emotion classes respectively were selected and merged to form a basic word vocabulary of size 2000. A new lexicon for each emotion class (in which each word has a weight indicating its inclination for expressing this emotion) was constructed. The new emotion lexicon not only collected words that appeared in one emotion class but also assigned a weight indicating its inclination for expressing this emotion. This emotion lexicon was then used to generate a vector feature representation for each utterance. Two types of fusion schemes were experimented with: early fusion (feature concatenation) and late fusion (classification score fusion). The SVM with linear kernel was used as emotion classifier. The system based on early fusion of Cepstral-BoW and GSV-mean acoustic features combined with ACO-based system, Cepstrum-based system, Lex-BoW-based system, and Lex-eVector-based system through late fusion achieves the best weighted emotion recognition accuracy of 69.2%.

Continuing with bimodal systems, Metallinou et al. [234] carried out emotion recognition using audio-visual modalities by exploiting Gaussian Mixture Models (GMMs). Markers were placed on the faces of actors to collect spatial information of these markers for each video frame in IEMOCAP. Facial markers were separated into six blocks, each of which defined a different facial region. A GMM was trained for each of the emotional states examined; angry (ANG), happy (HAP), neutral (NEU) and sad (SAD). The marker point coordinates were used as features for the training of Gaussian mixture models. The frame rate of the markers was 8.3 ms. The feature vector for each facial region consisted of 3-D coordinates of the markers belonging to that region plus their first and second derivatives. GMM with 64 mixtures was chosen as it was shown to achieve good performance. MFCCs are used for vocal analysis. The feature vector comprised 12 MFCCs and energy, their first and second derivatives; constituting a 39-dimensional feature vector. The window length for the MFCC extraction was 50ms and the overlap set to 25ms, to match the window of the facial data extraction. Similar to facial analysis, a GMM was trained for each emotion along with an extra one for background noise. Here, a GMM with 32 mixtures was chosen. Two different classifier combination techniques were explored: the first a Bayesian approach for multiple cue combination, and the second an ad-hoc method utilizing SVMs with radial basis kernels that used post classification accuracies as features. Anger and happiness were found to have better recognition accuracies in the face-based classifier compared to emotional states with lower levels of activation, such as sadness and neutrality; while anger and sadness demonstrated good accuracy in voice-based classifiers. A support vector classifier (SVC) was used to combine the separate face and voice model decisions. The Bayesian classifier and SVC classifiers were found to perform comparably, with neutral being the worst recognized emotional state, and anger/sadness being the best.

While previous works focused on bimodality, the work in [298] aims to classify emotions using audio, visual and textual information by attaching probabilities to each category based on automatically generated trees, with SVMs acting as nodes. There were several acoustic features used, ranging from jitter and shimmer for negative emotions to intensity and voicing statistics per frame. Instead of representing the non-stationary MFCC features using statistical functionals as in previous works, they use a set of model-based features obtained by scoring all MFCC vectors in a sentence using emotion-dependent Gaussian mixture models (GMM). The lexical features were summarized using LIWC and GI systems represented by bag of word stems. The visual features encapsulated facial animation parameters representing nose, mouth and chin markers, eyebrow angle, etc. A randomized tree is generated using the set of all classifiers whose performance is above a threshold parameter. The experiments were conducted in leave-one-speaker-out fashion. The unimodal feature set achieved an accuracy of around 63% whereas their combination led to an increase of around 8%.

5.2.3. Other multimodal cognitive research

DeVault et al. [299] introduced SimSensei Kiosk, a virtual human interviewer named Ellie, for automatic assessment of distress indicators among humans. Distress indicators are verbal and non-verbal behaviors correlated with depression, anxiety or post-traumatic stress disorder (PSTD). The SimSensei Kiosk was developed in such a way the user feels comfortable talking and sharing information, thus providing clinicians an automatic assessment of psychological distress in a person. The evaluation of the kiosk was carried out by the Wizard-of-Oz prototype system, which had two human operators for deciding verbal and non-verbal responses.

This development of SimSensei kiosk was carried out over a period of two years with 351 participants, out of which 217 were male, 132 were female and 2 did not report the gender. In this work, termed the Multisense framework, a multimodal real-time sensing system was used, for synchronized capture of different modalities, real-time tracking and fusion process. The multimodal system was also integrated with GAVAM head tracker, CLM-Z face tracker, SHORE face detector, and more. The SimSensei Kiosk used 4 statistically trained utterance classifiers to capture the utterance meaning of the users and Cerebella, a research platform for realization of the relation between mental states and human behavior.

Alam et al. [202], proposed an automatic personality trait recognition framework using the YouTube personality dataset. The dataset consists of videos by 404 YouTube bloggers (194 male and 204 female). The features used for this task were linguistic, psycholinguistic, emotional features and audio-visual features. Automatic recognition of personality traits is an important topic in the field of NLP, particularly aimed at processing the interaction between human and virtual agents. High dimensional features were selected using the relief algorithm and classification models were generated using SMO for the SVM. At the final stage, decision-level fusion for classification of personality traits was used.

Other notable work in personality recognition was carried out by Sarkar et al. [201], who used the YouTube personality dataset and a logistic regression model with ridge estimator, for classification purposes. They divided features into five categories, i.e., audio-visual features, text features, word statistics features, sentiment features and gender features. A total of 1079 features were used, with 25 audiovisual features, 3 word statistics feature, 5 sentiment feature, 1 demographic feature and 1045 text features. In conclusion, their in-depth feature analysis showcased helpful insights for solving the multimodal personality recognition task.

Siddiquie et al. [204], introduced the task of exploiting multimodal affect and semantics for automatic classification of politically persuasive web videos. Rallying A Crowd (RAC) dataset was used for experimentation with 230 videos. The approach was executed by extraction of audio, visual and textual features to capture affect and semantics in the audio-video content and sentiment in the viewers' comments. For the audio domain, several grades of speech arousal and related semantic categories such as crowd reaction and music were detected. For the visual domain, visual sentiment and semantic content were detected. The research employed both feature-level and decision-level fusion methods. In the case of decision-level fusion, the author used both conventional- and learning-based decision fusion approaches to enhance the overall classification performance.

In Table 6, some key research works in multimodal sentiment analysis and opinion mining are summarized and categorized based on their proposed method.

6. Available APIs

In this section, we list 20 popular APIs for emotion recognition from photos, videos, text and speech. The main categories of emotions that are detected using the APIs are Joy, Anger, Contempt, Fear, Surprise, Sadness and Disgust.

Sentiment analysis is also explored by some of the APIs, in addition to emotion recognition, to determine whether the expressed emotion is positive or negative.

- *Emotient*[1] detects Attention, Engagement and Sentiment from facial expressions. These factors are considered key performance indicators for adding business value to advertising, media, consumer packaged goods and other industries, which need consumers' feedbacks to improve the quality of their products.
- *Imotions*[2] combines Emotient face expression technology to extract emotions and sentiments from various observed bodily cues. It can also be easily combined with other technologies such as EEG, eye tracking, Galvanic Skin response, etc., to improve emotion recognition accuracy.
- *EmoVu*[3] by Eyeris is a comprehensive face analytics API that employs deep learning for emotion recognition. The API also provides vision software to support ambient intelligence and is also useful for detecting age and gender identification, eye tracking and gaze estimation.
- *nViso*[4] uses 3D facial imaging technology for emotion recognition from facial expressions in real time. The software is completely automated and received the IBM award for smarter computing in 2013.
- *Alchemy API*[5] is also powered by IBM Watson. The API performs sentiment analysis on large and small documents, news articles, blog posts, product reviews, comments and tweets.
- *Kairos*[6] provides an API for analyzing facial expressions and features for emotion recognition, gender and age detection and attention management. It provides applications for various industries such as advertising, market research, health, financial services, retail, etc.
- *Tone API*[7] provides emotional insights from written text. It focuses mainly on marketers and writers to improve their content on the basis of emotional insights.
- *Project Oxford*[8] by Microsoft provides APIs for categories such as Vision, Speech, Language, Knowledge and Search.
- *Face reader*[9] by Noldus is widely used for academic purposes. It is a facial expression analysis software for analyzing universal emotions in addition to neutral and contempt. The software is also used to observe gaze direction and head orientation.
- *Sightcorp*[10] is a facial expression analysis API and is also used for eye tracking, age and gender estimation, head pose estimation, etc.
- *SkyBiometry*[11] is a face recognition and face detection, cloud biometrics API. This API is used to detect emotions such as happy, sad, angry, surprise, disgust, scared and neutral from faces.
- *CrowdEmotions*[12] detects the dynamics of six basic emotions: happiness, surprise, anger, disgust, fear and sadness. It also captures people's engagement, emotions and body language, towards a particular event.
- *Affectiva*[13] is an API for emotion recognition using deep learning. It is said to have nearly 4 million faces as emotion database in order to provide great accuracy.
- *The Tone Analyzer*[14] is an API, powered by IBM Watson, for analyzing emotional states in text.
- *Repustate API*[15] is used for sentiment analysis in text. This API is based on linguistic theory and review cues based on POS tagging, lemmatization, prior polarity and negations.

[1] emotient.com

[2] imotions.com
[3] emovu.com
[4] nviso.ch
[5] alchemyapi.com
[6] kairos.com
[7] toneapi.com
[8] microsoft.com/cognitive-services/en-us/apis
[9] noldus.com/facereader
[10] sightcorp.com
[11] skybiometry.com
[12] crowdemotion.co.uk
[13] affectiva.com
[14] tone-analyzer-demo.mybluemix.net
[15] repustate.com/sentiment-analysis

- *Receptiviti API*[16] is used to analyze texts, tweets, emails, chats, surveys and voice data to provide insights into various aspects of people's personal lives, such as personality, emotion, tone and relationships.
- *Bitext*[17] is a text analysis API that is used for sentiment analysis, categorization, entity extraction and concept extraction. It is mainly focused for market research specialists.
- *Mood Patrol*[18] is used to detect emotions from given text. It was developed by Soul Hackers Lab and works reasonably well on small documents.
- *Synesketch*[19] is an open source software used for textual emotion recognition, sentiment recognition and visualization. It analyzes text in terms of emotions such as happiness, sadness, anger, fear, disgust, and surprise, and the intensity of emotion and sentiment, such as positive or negative.
- *Sentic API*[20] is a free API for emotion recognition and sentiment analysis providing semantics and sentics associated with 50,000 commonsense concepts in 40 different languages.

7. Discussion

Timely surveys are rudimentary for any field of research. In this survey, we not only discuss the state of the art but also collate available datasets and illustrate key steps involved in a multimodal affect analysis framework. We have covered around 100 papers in our study. In this section, we describe some of our major findings from this survey.

7.1. Major findings

The multimodal analysis of affective content now a days is as popular as the unimodal analysis. This is due to the need of mining information from the growing amount of videos posted the social media and the advancement of human-computer interaction agents. As discussed in [8], the trends in multimodal affect analysis can be classified into two timelines. Till 2003, "the use of basic signal processing and machine learning techniques, independently applied to still frames (but occasionally to sequences) of facial or vocal data, to detect exaggerated context free expressions of a few basic affective states, that are acted by a small number of individuals with no emphasis on generalizability" – in other words, mainly unimodal and in some cases bimodal, i.e., audio-visual clues, were used for affect analysis. In 2016, the trend leans towards using more than one modality for affect recognition of videos using machine learning techniques. In particular, there has been a growing interest in using deep learning techniques and a number of fusion methods. There is a significant amount of work that has been done in multimodal sentiment analysis in the past 3 years. The types of dataset are radically changing, wherein the past, acted data were being used, but presently, videos are being crawled from YouTube and used for experimentation in research.

Though visual and audio modalities have been used for multimodal affect recognition in many studies since 2004 (Table 4), it is worth mentioning that although most of the reported works use audio and visual information for affect recognition, recent advancements in text affect analysis [159] have led to increasing use of text modality in these works, particularly from 2010 onwards [6,189]. For example, from the Table 3 it can be seen that from 2010 onwards, text modality has been considered in many research works on multimodal affect analysis. More recent works, such as Poria et al. [189] have used a CNN for automatic feature extraction from text and fused with the visual and audio features. The effectiveness of this approach is evident from Fig. 11 where we plot the percentage of research works on multimodal affect analysis with or without text modality, reported over recent years. As can be seen from the Figure, most of the research studies on multimodal affect analysis report that the use of text modality boosts the performance of both unimodal and bimodal affect detectors.

In our literature survey, we have found more than 90% of studies reported visual modality as superior to audio and other modalities. Audio modality often suffers from the presence of noise in the signal. However, recent studies on multimodal sentiment analysis by Perez et al. [26] and Poria et al. [189] have demonstrated that an efficient and intelligent text analysis engine can outperform other unimodal classifiers, i.e., visual and audio. In both these independent studies, text modality was found to play the most vital role in multimodal sentiment analysis, and, furthermore, when fused with audio-visual features, it was shown to improve the performance significantly. On the MOUD dataset Perez et al. [26] obtained accuracies of 70.94%, 67.31% and 64.85% respectively for textual, visual and audio modalities. On the other hand, Poria et al. [189] obtained 79.77%, 76.38% and 74.22% accuracies for the respective modalities, on the same dataset.

In summary, as noted earlier, there are several concerns that need to be addressed in this research field. Firstly, the amount of trust that should be placed in acted corpora is debatable. The primary question that arises is if they appropriately replicate the natural characteristics of a spontaneous expression. For example, in acted data, people rarely smile while acting as a frustrated person, whereas studies [300] show that in 90% of cases in real-life situations, people smile while expressing their frustration. Such errors eventually lead to poor generalization of any multimodal sentiment analysis system.

Apart from replication, the taxonomy of the emotions and sentiments is never set in stone. Though for sentiment analysis, it is relatively straight forward and practically convenient to use positive, negative and neutral sentiment dimensions, in the case of emotions, the number of emotional dimensions to use is unclear. In the literature, most studies use Ekman's six basic emotions, i.e., Anger, Disgust, Joy, Surprise, fear and Sad, for experimentation, however people often tend to use complex emotions like love, frustration, etc., in their day to day conversations.

For text modality, deep learning is now a days the most popular methods. Compare to the commonly used bag of words method bag of concepts based methods are also developed by researchers [179]. On the other hand, also for visual modality the trend has shifted from the use of different complex image processing methods to the development of complex deep networks. With the advent of CNN [90], C3D [93] the video classification performance using the deep networks has overshadowed existing image processing algorithms like Optical flow, ASM, AAM. Though deep learning based methods are quite popular in text and visual affect recognition, not many works have been proposed in the literature for audio classification using deep networks. So, for audio classification the hand-crafted feature computation methods, e.g., OpenSMILE [229], are still very popular and widely used in the audio affect classification research. The degree to which a multimodal system can be generalized is also a crucial factor in determining its practical implementation. For example, it is particularly difficult to determine whether a developed approach is subject independent and can work well with any context, and to what extent the system should be trained on diverse contextual data.

To date, the most widely used fusion method is feature-level fusion, which consumes a lot of time and requires effective feature selection methods. Since 2010, multimodal fusion has drawn increasing attention of researchers, and a number of decision-level

[16] receptiviti.ai
[17] bitext.com/text-analysis-api
[18] market.mashape.com/soulhackerslabs/moodpatrol
[19] krcadinac.com/synesketch
[20] sentic.net/api

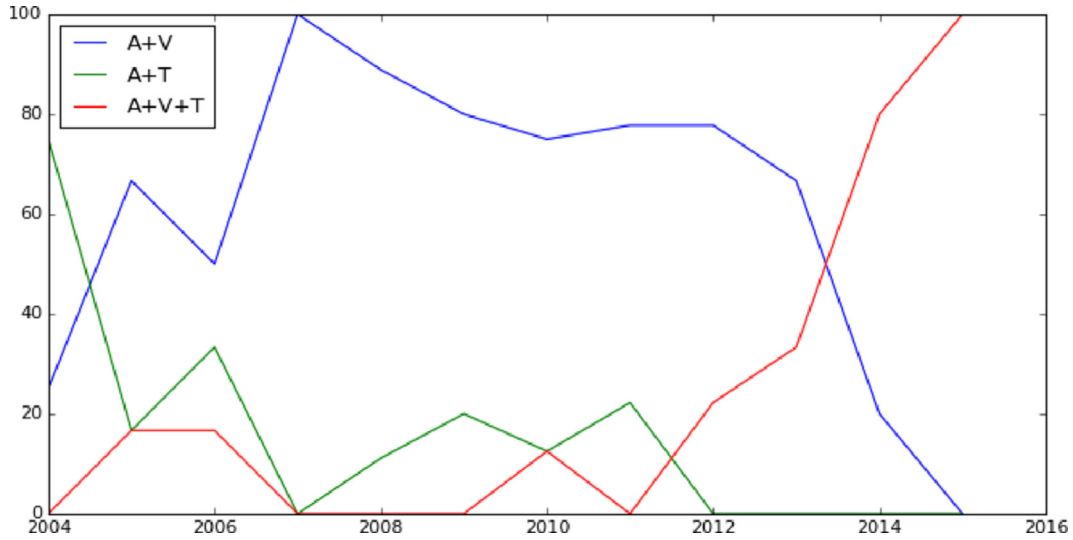


Fig. 11. Percentage of research works done using different modalities for affect recognition over the years (Legenda: A=Audio; T= Text; V=Video).

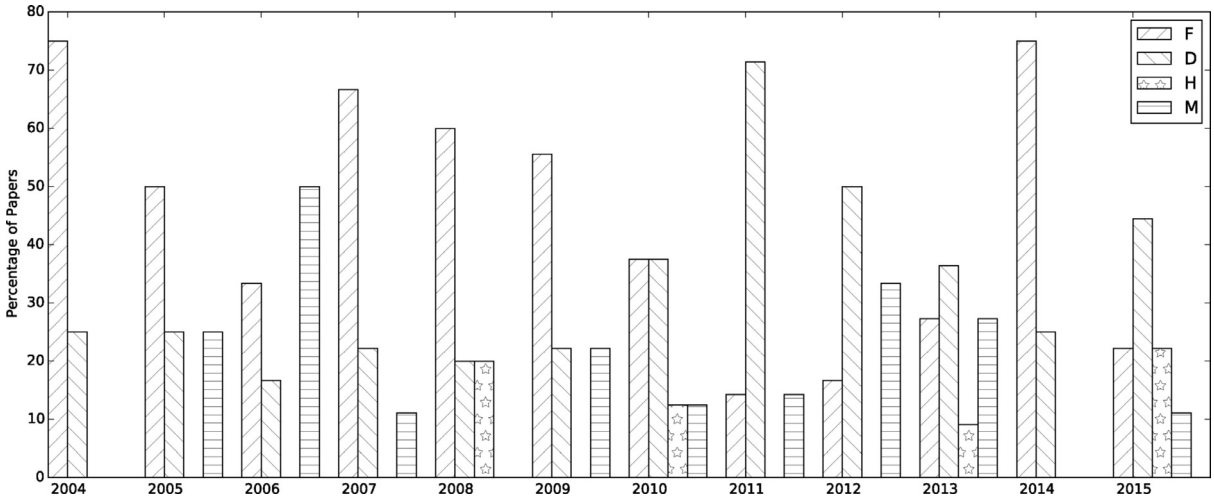


Fig. 12. Percentage of research articles in multimodal affect analysis using different fusion methods over the years (Legenda: F=Feature-Level Fusion; D=Decision-Level Fusion; H=Hybrid Fusion; M=Model-Based Fusion).

fusion methods have been recently reported, as can be seen from Fig. 12.

7.2. Future Directions

As this survey paper has demonstrated, there are significant research challenges outstanding in this multi-disciplinary field. One important area of future research is to investigate novel approaches for advancing our understanding of the temporal dependency between utterances, i.e., the effect of utterance at time *t* on the utterance at time *t+1*. Complex deep networks like 3D CNNs, RNNs have already been used to measure this temporal dependency, however these need to be further evaluated comparatively, on a range of real benchmark multi-modal datasets. Textual modality is often ignored in video classification. The progress in text classification research can play a major role in future of the multimodal affect analysis research.

With the advent of deep learning research, it is now a viable question whether to use deep features or low-level manually-extracted features for the video classification. Future research should focus on answering this question. A valid question is, can the ensemble application of deep learning and handcrafted features improve the performance of the video affect recognition? The use of deep learning for multimodal fusion can also be an important future work.

Extensive research is also required on videos containing spontaneous expressions rather than acted expressions. Furthermore, real-time fusion methods for fusing the information extracted from the multimodal data are also an interesting area of research, which can be enhanced through: 1) Testing the method on several datasets, 2) Increase in generalizability.

Another aspect of future research worthy of exploring, is to understand affect in a conversation. In such conversations, emotion expressed by a person can impact other persons in that conversation. If the multimodal system can model the inter person emotional dependency, that would lead to major advances in multimodal affect research.

8. Conclusion

In this paper, we carried out a first of its kind review of the fundamental stages of a multimodal affect recognition framework. We started by discussing available benchmark datasets, followed by an overview of the state of the art in audio-, visual- and textual-based affect recognition. In particular, we highlighted prominent studies in unimodal affect recognition, which we consider crucial

components of a multimodal affect detector framework. For example, without efficient unimodal affect classifiers or feature extractors, it is not possible to build a well-performing multimodal affect detector. Hence, if one is aware of the state of the art in unimodal affect recognition, which has been thoroughly reviewed in this paper, it would facilitate the construction of an appropriate multimodal framework.

Our survey has confirmed other researchers' findings that multimodal classifiers can outperform unimodal classifiers. Furthermore, text modality plays an important role in boosting the performance of an audio-visual affect detector. On the other hand, the use of deep learning is increasing in popularity, particularly for extracting features from modalities. Although feature-level fusion is widely used for multimodal fusion, there are other fusion methods developed in the literature. However, since fusion methods are, in general, not being used widely by the sentiment analysis and related NLP research communities, there are significant and timely opportunities for future research in the multi-disciplinary field of multimodal fusion. As identified in this review, some of the other key outstanding challenges in this exciting field include: estimating noise in unimodal channels, synchronization of frames, voice and utterance, reduction of multi-modal Big Data dimensionality to meet real-time performance needs, etc. These challenges suggest we are still far from producing a real-time multimodal affect detector which can effectively and affectively communicate with humans, and feel our emotions.

References

[1] J. Balazs, J. Velásquez, Opinion mining and information fusion: a survey, Inf. Fusion 27 (95-110) (2016).
[2] S. Sun, C. Luo, J. Chen, A review of natural language processing techniques for opinion mining systems, Inf. Fusion 36 (10-25) (2017).
[3] E. Cambria, H. Wang, B. White, Guest editorial: big social data analysis, Knowl.-Based Syst. 69 (2014) 1–2.
[4] V. Rosas, R. Mihalcea, L.-P. Morency, Multimodal sentiment analysis of spanish online videos, IEEE Intell. Syst. 28 (3) (2013) 38–45.
[5] H. Qi, X. Wang, S.S. Iyengar, K. Chakrabarty, Multisensor data fusion in distributed sensor networks using mobile agents, in: Proceedings of 5th International Conference on Information Fusion, 2001, pp. 11–16.
[6] L.-P. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis: harvesting opinions from the web, in: Proceedings of the 13th international conference on multimodal interfaces, ACM, 2011, pp. 169–176.
[7] S. Shimojo, L. Shams, Sensory modalities are not separate modalities: plasticity and interactions, Curr.Opin.Neurobiol. 11 (4) (2001) 505–509.
[8] S.K. D'mello, J. Kory, A review and meta-analysis of multimodal affect detection systems, ACM Comput. Surv. 47 (3) (2015) 43–79.
[9] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions, IEEE Trans. pattern Anal.Mach.Intell. 31 (1) (2009) 39–58.
[10] C. Darwin, The Expression of the Emotions in Man and Animals, John Murray, London, 1872.
[11] P. Ekman, D. Keltner, Universal facial expressions of emotion, California Mental Health Res. Digest 8 (4) (1970) 151–158.
[12] W.G. Parrott, Emotions in Social Psychology: Essential Readings, Psychology Press, 2001.
[13] T. Dalgleish, M.J. Power, Handbook of Cognition and Emotion, Wiley Online Library, 1999.
[14] J.J. Prinz, Gut Reactions: A Perceptual Theory of Emotion, Oxford University Press, 2004.
[15] J.A. Russell, Core affect and the psychological construction of emotion., Psychol.Rev. 110 (1) (2003) 145.
[16] C.E. Osgood, The nature and measurement of meaning., Psychol.Bull. 49 (3) (1952) 197–237.
[17] J.A. Russell, Affective space is bipolar., J.Personality Social Psychol. 37 (3) (1979) 345–356.
[18] C. Whissell, The dictionary of affect in language, Emotion 4 (113-131) (1989) 94.
[19] R. Plutchik, Emotion: A Psychoevolutionary Synthesis, Harpercollins College Division, 1980.
[20] A. Freitas, E. Castro, Facial expression: the effect of the smile in the treatment of depression. empirical study with Portuguese subjects, Emotional Expression (2009) 127–140.
[21] A. Mehrabian, Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament, Curr. Psychol. 14 (4) (1996) 261–292.
[22] J.R. Fontaine, K.R. Scherer, E.B. Roesch, P.C. Ellsworth, The world of emotions is not two-dimensional, Psychol.Sci. 18 (12) (2007) 1050–1057.
[23] T. Cochrane, Eight dimensions for the emotions, Social Sci. Inf. 48 (3) (2009) 379–420.
[24] E. Cambria, A. Livingstone, A. Hussain, The hourglass of emotions, in: A. Esposito, A. Vinciarelli, R. Hoffmann, V. Muller (Eds.), Cognitive Behavioral Systems, Lecture Notes in Computer Science, 7403, Springer, Berlin Heidelberg, 2012, pp. 144–157.
[25] E. Cambria, J. Fu, F. Bisio, S. Poria, AffectiveSpace 2: enabling affective intuition for concept-level sentiment analysis, in: Proceedings of AAAI, 2015, pp. 508–514. Austin
[26] V. Pérez-Rosas, R. Mihalcea, L.-P. Morency, Utterance-level multimodal sentiment analysis., in: ACL (1), 2013, pp. 973–982.
[27] M. Wollmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, L.-P. Morency, Youtube movie reviews: Sentiment analysis in an audio-visual context, Intell. Syst. IEEE 28 (3) (2013) 46–53.
[28] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, et al., The humaine database: addressing the collection and annotation of naturalistic and induced emotional data, in: Affective Computing Intell.Interact., Springer, 2007, pp. 488–500.
[29] E. Douglas-Cowie, R. Cowie, M. Schroder, A new emotion database: considerations, sources and scope, in: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, 2000, pp. 39–44.
[30] G. McKeown, M. Valstar, R. Cowie, M. Pantic, M. Schroder, The semaine database: annotated multimodal records of emotionally colored conversations between a person and a limited agent, Affective Comput. IEEE Trans.. 3 (1) (2012) 5–17.
[31] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amier, D. Heylen, The sensitive artificial listner: an induction technique for generating emotionally coloured conversation, LREC Workshop Corpora Res. Emot. Affect (2008).
[32] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, Iemocap: interactive emotional dyadic motion capture database, Lang.Resour.Eval. 42 (4) (2008) 335–359.
[33] O. Martin, I. Kotsia, B. Macq, I. Pitas, The enterface'05 audio-visual emotion database, in: Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on, IEEE, 2006. 8–8
[34] E. Paul, W. Friesen, Facial action coding system investigator's guide, 1978.
[35] C.E. Izard, L.M. Dougherty, E.A. Hembree, A System for Identifying Affect Expressions by Holistic Judgments (AFFEX), Instructional Resources Center, University of Delaware, 1983.
[36] A.M. Kring, D. Sloan, The facial expression coding system (faces): a users guide, Unpublished manuscript (1991).
[37] P. Ekman, W.V. Friesen, J.C. Hager, Facs investigator's guide, A Human Face (2002).
[38] P. Ekman, E. Rosenberg, J. Hager, Facial action coding system affect interpretation dictionary (facsaid), 1998.
[39] P. Ekman, E.L. Rosenberg, What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS), Oxford University Press, USA, 1997.
[40] D. Matsumoto, More evidence for the universality of a contempt expression, Motivation Emotion 16 (4) (1992) 363–368.
[41] W.E. Rinn, The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions., Psychol.Bull. 95 (1) (1984) 52.
[42] M.S. Bartlett, J.C. Hager, P. Ekman, T.J. Sejnowski, Measuring facial expressions by computer image analysis, Psychophysiology 36 (02) (1999) 253–263.
[43] M. Breidt, C. Wallraven, D.W. Cunningham, H. Bulthoff, Facial animation based on 3d scans and motion capture, Siggraph'03 Sketches and Applications (2003).
[44] F.I. Parke, K. Waters, Computer facial animation, CRC Press, 2008.
[45] H. Tao, H.H. Chen, W. Wu, T.S. Huang, Compression of mpeg-4 facial animation parameters for transmission of talking heads, Circuits Syst. Video Technol. IEEE Trans. 9 (2) (1999) 264–276.
[46] Y. Yacoob, L. Davis, Computing spatio-temporal representations of human faces, in: Computer Vision and Pattern Recognition, IEEE, 1994, pp. 70–75.
[47] M.J. Black, Y. Yacoob, Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion, in: Computer Vision, 1995. Proceedings., Fifth International Conference on, IEEE, 1995, pp. 374–381.
[48] Z. Zhang, Feature-based facial expression recognition: Sensitivity analysis and experiments with a multilayer perceptron, Int.J.Pattern Recognit.Artif. Intell. 13 (06) (1999) 893–911.
[49] A. Haro, M. Flickner, I. Essa, Detecting and tracking eyes by using their physiological properties, dynamics, and appearance, in: Computer Vision and Pattern Recognition, 1, IEEE, 2000, pp. 163–168.
[50] M.J. Jones, T. Poggio, Multidimensional morphable models, in: Computer Vision, 1998. Sixth International Conference on, IEEE, 1998, pp. 683–688.
[51] T.F. Cootes, G.J. Edwards, C.J. Taylor, et al., Active appearance models, IEEE Trans.Pattern Anal.Mach.Intell. 23 (6) (2001) 681–685.
[52] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, T.J. Sejnowski, Classifying facial actions, Pattern Anal. Mach. Intell. IEEE Trans. 21 (10) (1999) 974–989.
[53] Y.-l. Tian, T. Kanade, J.F. Cohn, Recognizing action units for facial expression analysis, Pattern Anal. Mach. Intell. IEEE Trans. 23 (2) (2001) 97–115.
[54] B. Fasel, J. Luettin, Recognition of asymmetric facial action unit activities and intensities, in: Pattern Recognition, 2000. Proceedings. 15th International Conference on, 1, IEEE, 2000, pp. 1100–1103.

[55] M.J. Lyons, J. Budynek, S. Akamatsu, Automatic classification of single facial images, IEEE Trans. Pattern Anal. Mach. Intell. 21 (12) (1999) 1357–1362.
[56] G. Littlewort, M.S. Bartlett, I. Fasel, J. Susskind, J. Movellan, Dynamics of facial expression extracted automatically from video, Image Vision Comput. 24 (6) (2006) 615–625.
[57] I. Cohen, N. Sebe, A. Garg, L.S. Chen, T.S. Huang, Facial expression recognition from video sequences: temporal and static modeling, Comput. Vision Image Underst. 91 (1) (2003) 160–187.
[58] Y. Wang, H. Ai, B. Wu, C. Huang, Real time facial expression recognition with adaboost, in: Proceedings of the 17th International Conference on, 3, IEEE, 2004, pp. 926–929.
[59] A. Lanitis, C.J. Taylor, T.F. Cootes, Automatic face identification system using flexible appearance models, Image Vision Comput. 13 (5) (1995) 393–401.
[60] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models-their training and application, Comput.Vision Image Underst. 61 (1) (1995) 38–59.
[61] V. Blanz, T. Vetter, A morphable model for the synthesis of 3d faces, in: Proceedings of the 26th annual conference on Computer graphics and interactive techniques, ACM Press/Addison-Wesley Publishing Co., 1999, pp. 187–194.
[62] H. Ohta, H. Saji, H. Nakatani, Recognition of facial expressions using muscle--based feature models, in: Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on, 2, IEEE, 1998, pp. 1379–1381.
[63] I. Cohen, N. Sebe, F. Gozman, M.C. Cirelo, T.S. Huang, Learning bayesian network classifiers for facial expression recognition both labeled and unlabeled data, in: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, 1, IEEE, 2003, pp. I–595.
[64] S. Kimura, M. Yachida, Facial expression recognition and its degree estimation, in: Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, IEEE, 1997, pp. 295–300.
[65] R. Verma, C. Davatzikos, J. Loughead, T. Indersmitten, R. Hu, C. Kohler, R.E. Gur, R.C. Gur, Quantification of facial expressions using high-dimensional shape transformations, J.Neurosci.Methods 141 (1) (2005) 61–73.
[66] T. Baltrušaitis, P. Robinson, L.-P. Morency, 3d constrained local model for rigid and non-rigid facial tracking, in: Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2610–2617.
[67] L.-P. Morency, J. Whitehill, J. Movellan, Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation, in: Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on, IEEE, 2008, pp. 1–8.
[68] M. Yeasin, B. Bullot, R. Sharma, From facial expression to level of interest: a spatio-temporal approach, in: Computer Vision and Pattern Recognition, 2, IEEE, 2004, pp. II–922.
[69] J.J.-J. Lien, T. Kanade, J.F. Cohn, C.-C. Li, Detection, tracking, and classification of action units in facial expression, Robo. Auton. Syst. 31 (3) (2000) 131–146.
[70] Y. Chang, C. Hu, M. Turk, Probabilistic expression analysis on manifolds, in: Computer Vision and Pattern Recognition, 2, IEEE, 2004, pp. II–520.
[71] A.M. Kring, D.M. Sloan, The facial expression coding system (faces): development, validation, and utility., Psychol.Assess. 19 (2) (2007) 210.
[72] C. Davatzikos, Measuring biological shape using geometry-based shape transformations, Image Vision Comput. 19 (1) (2001) 63–74.
[73] Z. Wen, T.S. Huang, Capturing subtle facial motions in 3d face tracking, in: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, IEEE, 2003, pp. 1343–1350.
[74] M. Pantic, L.J. Rothkrantz, Expert system for automatic analysis of facial expressions, Image Vision Comput. 18 (11) (2000a) 881–905.
[75] M. Pantic, L.J. Rothkrantz, Automatic analysis of facial expressions: The state of the art, Pattern Anal. Mach. Intell. IEEE Trans. 22 (12) (2000b) 1424–1445.
[76] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recognit. 36 (1) (2003) 259–275.
[77] M. De Meijer, The contribution of general features of body movement to the attribution of emotions, J. Nonverbal Behav. 13 (4) (1989) 247–268.
[78] A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis, P.F. Driessen, Gesture-based affective computing on motion capture data, in: Affective Computing and Intelligent Interaction, Springer, 2005, pp. 1–7.
[79] S. Piana, A. Staglianò, A. Camurri, F. Odone, A set of full-body movement features for emotion recognition to help children affected by autism spectrum condition, IDGEI International Workshop, 2013.
[80] S. Piana, A. Stagliano, F. Odone, A. Verri, A. Camurri, Real-time automatic emotion recognition from body gestures, arXiv preprint arXiv:1402.5047 (2014).
[81] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, K. Karpouzis, Multimodal emotion recognition from expressive faces, body gestures and speech, in: Artificial intelligence and innovations 2007: From theory to applications, Springer, 2007, pp. 375–388.
[82] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, S. Kollias, Emotion analysis in man-machine interaction systems, in: Machine learning for multimodal interaction, Springer, 2004, pp. 318–328.
[83] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[84] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.
[85] Y. LeCun, K. Kavukcuoglu, C. Farabet, et al., Convolutional networks and applications in vision., in: ISCAS, 2010, pp. 253–256.
[86] K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, Y.L. Cun, Learning convolutional feature hierarchies for visual recognition, in: Advances in neural information processing systems, 2010, pp. 1090–1098.
[87] P. Hamel, D. Eck, Learning features from music audio with deep belief networks., in: ISMIR, Utrecht, The Netherlands, 2010, pp. 339–344.
[88] I. Chaturvedi, Y.-S. Ong, I. Tsang, R. Welsch, E. Cambria, Learning word dependencies in text by means of a deep recurrent belief network, Knowl.-Based Syst. 108 (2016) 144–154.
[89] G. Hinton, A practical guide to training restricted boltzmann machines, Momentum 9 (1) (2010) 926.
[90] C. Xu, S. Cetintas, K.-C. Lee, L.-J. Li, Visual sentiment prediction with deep convolutional neural networks, arXiv preprint arXiv:1411.5731 (2014).
[91] Q. You, J. Luo, H. Jin, J. Yang, Robust image sentiment analysis using progressively trained and domain transferred deep networks, arXiv preprint arXiv:1509.06041 (2015).
[92] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, L. Sigal, Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization, arXiv preprint arXiv:1511.04798 (2015).
[93] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, arXiv preprint arXiv:1412.0767 (2014).
[94] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional MKL based multimodal emotion recognition and sentiment analysis, Proceedings of ICDM, Barcelona, 2016.
[95] C.-H. Wu, J.-F. Yeh, Z.-J. Chuang, Emotion perception and recognition from speech, in: Affective Information Processing, Springer, 2009, pp. 93–110.
[96] D. Morrison, R. Wang, L.C. De Silva, Ensemble methods for spoken emotion recognition in call-centres, Speech Commun. 49 (2) (2007) 98–112.
[97] C.-H. Wu, W.-B. Liang, Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels, IEEE Trans. Affective Comput. 2 (1) (2011) 10–21.
[98] I.R. Murray, J.L. Arnott, Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion, J. Acoust. Soc. Am. 93 (2) (1993) 1097–1108.
[99] I. Luengo, E. Navas, I. Hernáez, J. Sánchez, Automatic emotion recognition using prosodic parameters., in: Interspeech, 2005, pp. 493–496.
[100] S.G. Koolagudi, N. Kumar, K.S. Rao, Speech emotion recognition using segmental level prosodic analysis, in: Devices and Communications (ICDeCom), 2011 International Conference on, IEEE, 2011, pp. 1–5.
[101] D. Västfjäll, M. Kleiner, Emotion in product sound design, Proceedings of Journées Design Sonore (2002).
[102] A. Batliner, K. Fischer, R. Huber, J. Spilker, E. Nöth, How to find trouble in communication, Speech Commun. 40 (1) (2003) 117–143.
[103] C.M. Lee, S.S. Narayanan, Toward detecting emotions in spoken dialogs, IEEE Trans.Speech Audio Process. 13 (2) (2005) 293–303.
[104] J. Hirschberg, S. Benus, J.M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis, et al., Distinguishing deceptive from non-deceptive speech., in: INTERSPEECH, 2005, pp. 1833–1836.
[105] L. Devillers, L. Vidrascu, L. Lamel, Challenges in real-life emotion annotation and machine learning based detection, Neural Netw. 18 (4) (2005) 407–422.
[106] T. Vogt, E. André, Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition, in: Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on, IEEE, 2005, pp. 474–477.
[107] F. Eyben, M. Wöllmer, B. Schuller, Openear—introducing the munich open–source emotion and affect recognition toolkit, in: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, IEEE, 2009, pp. 1–6.
[108] R.W. Levenson, Human emotion: a functional view, Nat.Emotion 1 (1994) 123–126.
[109] D. Datcu, L. Rothkrantz, Semantic audio-visual data fusion for automatic emotion recognition, Euromedia'2008 (2008).
[110] F. Dellaert, T. Polzin, A. Waibel, Recognizing emotion in speech, in: Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, 3, IEEE, 1996, pp. 1970–1973.
[111] T. Johnstone, Emotional speech elicited using computer games, in: Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, 3, IEEE, 1996, pp. 1985–1988.
[112] L.S.-H. Chen, Joint Processing of Audio-Visual Information for the Recognition of Emotional Expressions in Human-Computer Interaction, Citeseer, 2000 Ph.D. thesis.
[113] E. Navas, I. Hernaez, I. Luengo, An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional tts, Audio Speech Lang. Process. IEEE Transac. 14 (4) (2006) 1117–1127.
[114] M. El Ayadi, M.S. Kamel, F. Karray, Survey on speech emotion recognition: features, classification schemes, and databases, Pattern Recognit. 44 (3) (2011) 572–587.
[115] H. Atassi, A. Esposito, A speaker independent approach to the classification of emotional vocal expressions, in: Tools with Artificial Intelligence, 2008. ICTAI'08. 20th IEEE International Conference on, 2, IEEE, 2008, pp. 147–152.
[116] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, A database of german emotional speech., in: Interspeech, 5, 2005, pp. 1517–1520.
[117] P. Pudil, F. Ferri, J. Novovicova, J. Kittler, Floating search methods for feature selection with nonmonotonic criterion functions, in: Pattern Recognition, 1994. Vol. 2-Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on, 2, IEEE, 1994, pp. 279–283.
[118] K.R. Scherer, Adding the affective dimension: a new look in speech analysis and synthesis., ICSLP, 1996.

[119] Z. Huang, M. Dong, Q. Mao, Y. Zhan, Speech emotion recognition using cnn, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2014, pp. 801–804.
[120] A. Graves, S. Fernández, J. Schmidhuber, Bidirectional lstm networks for improved phoneme classification and recognition, in: Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005, Springer, 2005, pp. 799–804.
[121] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, R. Cowie, On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues, J. Multimodal User Interfaces 3 (1-2) (2010) 7–19.
[122] N. Anand, P. Verma, Convoluted feelings convolutional and recurrent nets for detecting emotion from audio data, Technical Report, Stanford University, 2015.
[123] K. Han, D. Yu, I. Tashev, Speech emotion recognition using deep neural network and extreme learning machine., in: Interspeech, 2014, pp. 223–227.
[124] A. Tajadura-Jiménez, D. Västfjäll, Auditory-induced emotion: a neglected channel for communication in human-computer interaction, in: Affect and Emotion in Human-Computer Interaction, Springer, 2008, pp. 63–74.
[125] T. Vogt, E. André, J. Wagner, Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation, in: Affect and emotion in human-computer interaction, Springer, 2008, pp. 75–91.
[126] C. Strapparava, A. Valitutti, Wordnet affect: an affective extension of wordnet., in: LREC, 4, 2004, pp. 1083–1086.
[127] C.O. Alm, D. Roth, R. Sproat, Emotions from text: machine learning for text-based emotion prediction, in: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2005, pp. 579–586.
[128] G. Mishne, et al., Experiments with mood classification in blog posts, in: Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access, 19, Citeseer, 2005, pp. 321–327.
[129] L. Oneto, F. Bisio, E. Cambria, D. Anguita, Statistical learning theory and ELM for big social data analysis, IEEE Comput. Intell. Mag. 11 (3) (2016) 45–55.
[130] C. Yang, K.H.-Y. Lin, H.-H. Chen, Building emotion lexicon from weblog corpora, in: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, 2007, pp. 133–136.
[131] F.-R. Chaumartin, Upar7: a knowledge-based system for headline sentiment tagging, in: Proceedings of the 4th International Workshop on Semantic Evaluations, Association for Computational Linguistics, 2007, pp. 422–425.
[132] A. Esuli, F. Sebastiani, Sentiwordnet: a publicly available lexical resource for opinion mining, in: Proceedings of LREC, 6, Citeseer, 2006, pp. 417–422.
[133] K.H.-Y. Lin, C. Yang, H.-H. Chen, What emotions do news articles trigger in their readers? in: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2007, pp. 733–734.
[134] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2004, pp. 168–177.
[135] E. Cambria, Affective computing and sentiment analysis, IEEE Intell. Syst. 31 (2) (2016) 102–107.
[136] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, in: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, 2002, pp. 79–86.
[137] R. Socher, A. Perelygin, J.Y. Wu, J. Chuang, C.D. Manning, A.Y. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the conference on empirical methods in natural language processing (EMNLP), 1631, Citeseer, 2013, p. 1642.
[138] H. Yu, V. Hatzivassiloglou, Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences, in: Proceedings of the 2003 conference on Empirical methods in natural language processing, Association for Computational Linguistics, 2003, pp. 129–136.
[139] P. Melville, W. Gryc, R.D. Lawrence, 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, ACM, 2009, pp. 1275–1284.
[140] P.D. Turney, Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 417–424.
[141] X. Hu, J. Tang, H. Gao, H. Liu, Unsupervised sentiment analysis with emotional signals, in: Proceedings of the 22nd international conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2013, pp. 607–618.
[142] A. Gangemi, V. Presutti, D. Reforgiato Recupero, Frame-based detection of opinion holders and topics: a model and a tool, Comput. Intell. Mag. IEEE 9 (1) (2014) 20–30, doi:10.1109/MCI.2013.2291688.
[143] E. Cambria, A. Hussain, Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis, Springer, Cham, Switzerland, 2015.
[144] H. Kanayama, T. Nasukawa, Fully automatic lexicon expansion for domain-oriented sentiment analysis, in: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2006, pp. 355–363.
[145] J. Blitzer, M. Dredze, F. Pereira, et al., Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification, in: ACL, 7, 2007, pp. 440–447.
[146] S.J. Pan, X. Ni, J.-T. Sun, Q. Yang, Z. Chen, Cross-domain sentiment classification via spectral feature alignment, in: Proceedings of the 19th international conference on World wide web, ACM, 2010, pp. 751–760.
[147] D. Bollegala, D. Weir, J. Carroll, Cross-domain sentiment classification using a sentiment sensitive thesaurus, Knowl. Data Eng. IEEE Trans. 25 (8) (2013) 1719–1731.
[148] E. Cambria, S. Poria, R. Bajpai, B. Schuller, SenticNet 4: a semantic resource for sentiment analysis based on conceptual primitives, in: Proceedings of COLING, 2016, pp. 2666–2677.
[149] H.-H. Wu, A.C.-R. Tsai, R.T.-H. Tsai, J.Y.-j. Hsu, Sentiment value propagation for an integral sentiment dictionary based on commonsense knowledge, in: Technologies and Applications of Artificial Intelligence (TAAI), 2011 International Conference on, IEEE, 2011, pp. 75–81.
[150] J.M. Chenlo, D.E. Losada, An empirical study of sentence features for subjectivity and polarity classification, Inf. Sci. 280 (2014) 275–288.
[151] R. Shah, Y. Yu, A. Verma, S. Tang, A. Shaikh, R. Zimmermann, Leveraging multimodal information for event summarization and concept-level sentiment analysis, Knowl.-Based Syst. 108 (2016) 102–109.
[152] G. Gezici, R. Dehkharghani, B. Yanikoglu, D. Tapucu, Y. Saygin, Su-sentilab: a classification system for sentiment analysis in twitter, in: International Workshop on Semantic Evaluation, 2013, pp. 471–477.
[153] S. Poria, E. Cambria, D. Hazarika, P. Vij, A deeper look into sarcastic tweets using deep convolutional neural networks, in: Proceedings of COLING, 2016, pp. 1601–1612.
[154] F. Bravo-Marquez, M. Mendoza, B. Poblete, Meta-level sentiment models for big social data analysis, Knowl.-Based Syst. 69 (2014) 86–99.
[155] M. Jaiswal, S. Tabibu, R. Bajpai, The truth and nothing but the truth: multimodal analysis for deception detection, ICDM, 2016.
[156] H. Xie, X. Li, T. Wang, R. Lau, T.-L. Wong, L. Chen, F.-L. Wang, Q. Li, Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy, Inf. Process. Manage. 52 (2016) 61–72.
[157] A. Scharl, A. Hubmann-Haidvogel, A. Jones, D. Fischl, R. Kamolov, A. Weichselbraun, W. Rafelsberger, Analyzing the public discourse on works of fiction – detection and visualization of emotion in online coverage about hbo's game of thrones, Inf. Process. Manage. 52 (1) (2016) 129–138.
[158] M. Egger, D. Schoder, Consumer-oriented tech mining: Integrating the consumer perspective into organizational technology intelligence – the case of autonomous driving, in: Hawaii International Conference on System Sciences, 2017.
[159] S. Poria, E. Cambria, G. Winterstein, G.-B. Huang, Sentic patterns: dependency-based rules for concept-level sentiment analysis, Knowl.-Based Syst. 69 (2014) 45–63.
[160] E. Cambria, A. Hussain, C. Havasi, C. Eckl, Common sense computing: from the society of mind to digital intuition and beyond, in: J. Fierrez, J. Ortega, A. Esposito, A. Drygajlo, M. Faundez-Zanuy (Eds.), Biometric ID Management and Multimodal Communication, Lecture Notes in Computer Science, 5707, Springer, Berlin Heidelberg, 2009, pp. 252–259.
[161] V. Hatzivassiloglou, K.R. McKeown, Predicting the semantic orientation of adjectives, in: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 1997, pp. 174–181.
[162] L. Jia, C. Yu, W. Meng, The effect of negation on sentiment analysis and retrieval effectiveness, in: Proceedings of the 18th ACM conference on Information and knowledge management, ACM, 2009, pp. 1827–1830.
[163] A. Reyes, P. Rosso, On the difficulty of automatically detecting irony: beyond a simple case of negation, Knowl. Inf. Syst. 40 (3) (2014) 595–614.
[164] K. Chawla, A. Ramteke, Iitb-sentiment-analysts: Participation in sentiment analysis in twitter semeval 2013 task, in: Second Joint Conference on Lexical and Computational Semantics (* SEM), 2, Citeseer, 2013, pp. 495–500.
[165] L. Polanyi, C. Culy, M. Van Den Berg, G.L. Thione, D. Ahn, Sentential structure and discourse parsing, in: Proceedings of the 2004 ACL Workshop on Discourse Annotation, Association for Computational Linguistics, 2004, pp. 80–87.
[166] F. Wolf, E. Gibson, Representing discourse coherence: a corpus-based study, Comput. Ling. 31 (2) (2005) 249–287.
[167] B. Wellner, J. Pustejovsky, C. Havasi, A. Rumshisky, R. Sauri, Classification of discourse coherence relations: an exploratory study using multiple knowledge sources, in: Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, Association for Computational Linguistics, 2009, pp. 117–125.
[168] B.P. Ramesh, H. Yu, Identifying discourse connectives in biomedical text, in: AMIA Annual Symposium Proceedings, 2010, American Medical Informatics Association, 2010, p. 657.
[169] B. Liu, Sentiment analysis and opinion mining, Synth. Lect. Human Lang. Technol. 5 (1) (2012) 1–167.
[170] K. Moilanen, S. Pulman, Sentiment composition, in: Proceedings of the Recent Advances in Natural Language Processing International Conference, 2007, pp. 378–382.
[171] X. Ding, B. Liu, P.S. Yu, A holistic lexicon-based approach to opinion mining, in: Proceedings of the 2008 International Conference on Web Search and Data Mining, ACM, 2008, pp. 231–240.
[172] S. Poria, E. Cambria, A. Gelbukh, Aspect extraction for opinion mining with a deep convolutional neural network, Knowl.-Based Syst. 108 (2016) 42–49.

[173] H.T. Ng, W.B. Goh, K.L. Low, Feature selection, perceptron learning, and a usability case study for text categorization, in: ACM SIGIR Forum, 31, ACM, 1997, pp. 67–73.

[174] Y.-H. Kim, S.-Y. Hahn, B.-T. Zhang, Text filtering by boosting naive bayes classifiers, in: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2000, pp. 168–175.

[175] A. Jordan, On discriminative vs. generative classifiers: Aa comparison of logistic regression and naive bayes, Adv.Neural Inf.Process.Syst. 14 (2002) 841.

[176] Y. Li, D. McLean, Z.A. Bandar, J.D. O'shea, K. Crockett, Sentence similarity based on semantic nets and corpus statistics, Knowl. Data Eng. IEEE Trans. 18 (8) (2006) 1138–1150.

[177] X.-H. Phan, L.-M. Nguyen, S. Horiguchi, Learning to classify short and sparse text & web with hidden topics from large-scale data collections, in: Proceedings of the 17th international conference on World Wide Web, ACM, 2008, pp. 91–100.

[178] M. Sahlgren, R. Cöster, Using bag-of-concepts to improve the performance of support vector machines in text categorization, in: Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics, 2004, p. 487.

[179] F. Wang, Z. Wang, Z. Li, J.-R. Wen, Concept-based short text classification and ranking, in: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, ACM, 2014, pp. 1069–1078.

[180] Y. Zhang, B. Liu, Semantic text classification of emergent disease reports, in: Knowledge Discovery in Databases: PKDD 2007, Springer, 2007, pp. 629–637.

[181] C.-E. Wu, R.T.-H. Tsai, Using relation selection to improve value propagation in a conceptnet-based sentiment dictionary, Knowl.-Based Syst. 69 (2014) 100–107.

[182] E. Cambria, B. White, Jumping NLP curves: a review of natural language processing research, IEEE Comput. Intell. Mag. 9 (2) (2014) 48–57.

[183] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics, 2005, pp. 347–354.

[184] N. Asher, F. Benamara, Y.Y. Mathieu, Appraisal of opinion expressions in discourse, Lingvisticæ Investigationes 32 (2) (2009) 279–292.

[185] R. Narayanan, B. Liu, A. Choudhary, Sentiment analysis of conditional sentences, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, Association for Computational Linguistics, 2009, pp. 180–189.

[186] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, J. Mach. Learn. Res. 12 (2011) 2493–2537.

[187] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, CoRR abs/1404.2188 (2014).

[188] X. Glorot, A. Bordes, Y. Bengio, Domain adaptation for large-scale sentiment classification: a deep learning approach, in: In Proceedings of the Twenty-eight International Conference on Machine Learning, ICML, 2011, pp. 513–520.

[189] S. Poria, E. Cambria, A. Gelbukh, Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis, in: Proceedings of EMNLP, 2015, pp. 2539–2544.

[190] Y. Kim, Convolutional neural networks for sentence classification, arXiv preprint arXiv:1408.5882 (2014).

[191] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[192] Z.-J. Chuang, C.-H. Wu, Multi-modal emotion recognition from speech and text, Comput. Ling. Chinese Lang.Process. 9 (2) (2004) 45–62.

[193] K. Forbes-Riley, D.J. Litman, Predicting emotion in spoken dialogue from multiple knowledge sources., in: HLT-NAACL, Citeseer, 2004, pp. 201–208.

[194] D.J. Litman, K. Forbes-Riley, Predicting student emotions in computer-human tutoring dialogues, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2004, p. 351.

[195] G. Rigoll, R. Müller, B. Schuller, Speech emotion recognition exploiting acoustic and linguistic information sources, Proc. SPECOM, Patras, Greece (2005) 61–67.

[196] D.J. Litman, K. Forbes-Riley, Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors, Speech Commun. 48 (5) (2006) 559–590.

[197] D. Seppi, A. Batliner, B. Schuller, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, V. Aharonson, Patterns, prototypes, performance: classifying emotional user states., in: INTERSPEECH, 2008, pp. 601–604.

[198] B. Schuller, Recognizing affect from linguistic information in 3d continuous space, Affective Comput. IEEE Trans. 2 (4) (2011) 192–205.

[199] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, R. Prasad, Speech language & multimedia technol., raytheon bbn technol., cambridge, ma, usa, in: Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific, IEEE, 2012, pp. 1–4.

[200] A. Savran, H. Cao, M. Shah, A. Nenkova, R. Verma, Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering, in: Proceedings of the 14th ACM international conference on Multimodal interaction, ACM, 2012, pp. 485–492.

[201] C. Sarkar, S. Bhatia, A. Agarwal, J. Li, Feature analysis for computational personality recognition using youtube personality data set, in: Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition, ACM, 2014, pp. 11–14.

[202] F. Alam, G. Riccardi, Predicting personality traits using multimodal information, in: Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition, ACM, 2014, pp. 15–18.

[203] J.G. Ellis, B. Jou, S.-F. Chang, Why we watch the news: a dataset for exploring sentiment in broadcast video news, in: Proceedings of the 16th International Conference on Multimodal Interaction, ACM, 2014, pp. 104–111.

[204] B. Siddiquie, D. Chisholm, A. Divakaran, Exploiting multimodal affect and semantics to identify politically persuasive web videos, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 203–210.

[205] S. Poria, E. Cambria, A. Hussain, G.-B. Huang, Towards an intelligent framework for multimodal affective data analysis, Neural Netw. 63 (2015) 104–116.

[206] G. Cai, B. Xia, Convolutional neural networks for multimedia sentiment analysis, in: National CCF Conference on Natural Language Processing and Chinese Computing, Springer, 2015, pp. 159–167.

[207] R. Ji, D. Cao, D. Lin, Cross-modality sentiment analysis for social multimedia, in: Multimedia Big Data (BigMM), 2015 IEEE International Conference on, IEEE, 2015, pp. 28–31.

[208] T. Yamasaki, Y. Fukushima, R. Furuta, L. Sun, K. Aizawa, D. Bollegala, Prediction of user ratings of oral presentations using label relations, in: Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia, ACM, 2015, pp. 33–38.

[209] H. Monkaresi, M.S. Hussain, R.A. Calvo, Classification of affects using head movement, skin color features and physiological signals, in: Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on, IEEE, 2012, pp. 2664–2669.

[210] S. Wang, Y. Zhu, G. Wu, Q. Ji, Hybrid video emotional tagging using users' eeg and video content, Multimedia tools Appl. 72 (2) (2014) 1257–1283.

[211] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, S. Narayanan, Analysis of emotion recognition using facial expressions, speech and multimodal information, in: Proceedings of the 6th international conference on Multimodal interfaces, ACM, 2004, pp. 205–211.

[212] C.-Y. Chen, Y.-K. Huang, P. Cook, Visual/acoustic emotion recognition, in: 2005 IEEE International Conference on Multimedia and Expo, IEEE, 2005, pp. 1468–1471.

[213] H. Gunes, M. Piccardi, Fusing face and body display for bi-modal emotion recognition: single frame analysis and multi-frame post integration, in: Affective Computing and Intelligent Interaction, Springer, 2005, pp. 102–111.

[214] S. Hoch, F. Althoff, G. McGlaun, G. Rigoll, Bimodal fusion of emotional data in an automotive environment, in: Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on, 2, IEEE, 2005, pp. ii–1085.

[215] A. Kapoor, R.W. Picard, Multimodal affect recognition in learning environments, in: Proceedings of the 13th annual ACM international conference on Multimedia, ACM, 2005, pp. 677–682.

[216] J. Kim, Bimodal Emotion Recognition using Speech and Physiological Changes, INTECH Open Access Publisher, 2007.

[217] Y. Wang, L. Guan, Recognizing human emotional state from audiovisual signals*, Multimedia IEEE Trans. 10 (5) (2008) 936–946.

[218] Z. Zeng, J. Tu, M. Liu, T.S. Huang, Multi-stream confidence analysis for audio-visual affect recognition, in: Affective Computing and Intelligent Interaction, Springer, 2005, pp. 964–971.

[219] H. Gunes, M. Piccardi, Affect recognition from face and body: early fusion vs. late fusion, in: 2005 IEEE international conference on systems, man and cybernetics, 4, IEEE, 2005, pp. 3437–3443.

[220] P. Pal, A.N. Iyer, R.E. Yantorno, Emotion detection from infant facial expressions and cries, in: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, 2, IEEE, 2006. II–II

[221] N. Sebe, I. Cohen, T. Gevers, T.S. Huang, Emotion recognition based on joint visual and audio cues, in: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, 1, IEEE, 2006, pp. 1136–1139.

[222] Z. Zeng, Y. Hu, Y. Fu, T.S. Huang, G.I. Roisman, Z. Wen, Audio-visual emotion recognition in adult attachment interview, in: Proceedings of the 8th international conference on Multimodal interfaces, ACM, 2006, pp. 139–145.

[223] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, K. Karpouzis, Modeling naturalistic affective states via facial and vocal expressions recognition, in: Proceedings of the 8th international conference on Multimodal interfaces, ACM, 2006, pp. 146–154.

[224] S. D'mello, A. Graesser, Mind and body: dialogue and posture for affect detection in learning environments, Frontiers Artif. Intell.Appl. 158 (2007) 161.

[225] S. Gong, C. Shan, T. Xiang, Visual inference of human emotion and behaviour, in: Proceedings of the 9th international conference on Multimodal interfaces, ACM, 2007, pp. 22–29.

[226] M.-J. Han, J.-H. Hsu, K.-T. Song, F.-Y. Chang, A new information fusion method for svm-based robotic audio-visual emotion recognition, in: Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on, IEEE, 2007, pp. 2656–2661.

[227] J. Jong-Tae, S. Sang-Wook, K. Kwang-Eun, S. Kwee-Bo, Emotion recognition method based on multimodal sensor fusion algorithm, ISIS, Sokcho-City (2007).

[228] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, S. Kollias, Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition, in: Artifical intelligence for human computing, Springer, 2007, pp. 91–112.
[229] B. Schuller, R. Müeller, B. Höernler, A. Höethker, H. Konosu, G. Rigoll, Audiovisual recognition of spontaneous interest within conversations, in: Proceedings of the 9th international conference on Multimodal interfaces, ACM, 2007, pp. 30–37.
[230] C. Shan, S. Gong, P.W. McOwan, Beyond facial expressions: learning human emotion from body gestures., in: BMVC, 2007, pp. 1–10.
[231] Z. Zeng, J. Tu, M. Liu, T.S. Huang, B. Pianfetti, D. Roth, S. Levinson, Audio-visual affect recognition, Multimedia IEEE Trans. 9 (2) (2007) 424–428.
[232] S. Haq, P.J. Jackson, J. Edge, Audio-visual feature selection and reduction for emotion classification, in: Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'08), Tangalooma, Australia, 2008.
[233] I. Kanluan, M. Grimm, K. Kroschel, Audio-visual emotion recognition using an emotion space concept, in: Signal Processing Conference, 2008 16th European, IEEE, 2008, pp. 1–5.
[234] A. Metallinou, S. Lee, S. Narayanan, Audio-visual emotion recognition using gaussian mixture models for face and voice, in: Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on, IEEE, 2008, pp. 250–257.
[235] M. Wimmer, B. Schuller, D. Arsic, G. Rigoll, B. Radig, Low-level fusion of audio, video feature for multi-modal emotion recognition., in: VISAPP (2), 2008, pp. 145–151.
[236] J.N. Bailenson, E.D. Pontikakis, I.B. Mauss, J.J. Gross, M.E. Jabon, C.A. Hutcherson, C. Nass, O. John, Real-time classification of evoked emotions using facial feature tracking and physiological responses, Int.J.Human-Comput.Stud. 66 (5) (2008) 303–317.
[237] G. Castellano, L. Kessous, G. Caridakis, Emotion recognition through multiple modalities: face, body gesture, speech, in: Affect and emotion in human-computer interaction, Springer, 2008, pp. 92–103.
[238] G. Chetty, M. Wagner, R. Goecke, A multilevel fusion approach for audiovisual emotion recognition., in: AVSP, 2008, pp. 115–120.
[239] S. Emerich, E. Lupu, A. Apatean, Emotions recognition by speech and facial expressions analysis, in: Proceedings of the 17th European Signal Processing Conference (EUSIPCO'09), 2009, pp. 1617–1621.
[240] H. Gunes, M. Piccardi, Automatic temporal segment detection and affect recognition from face and body display, Syst. Man, Cybern. Part B 39 (1) (2009) 64–84.
[241] S. Haq, P.J. Jackson, J. Edge, Speaker-dependent audio-visual emotion recognition., in: AVSP, 2009, pp. 53–58.
[242] Z. Khalili, M.H. Moradi, Emotion recognition system using brain and peripheral signals: using correlation dimension to improve the results of eeg, in: 2009 International Joint Conference on Neural Networks, IEEE, 2009, pp. 1571–1575.
[243] M. Paleari, R. Benmokhtar, B. Huet, Evidence theory-based multimodal emotion recognition, in: International Conference on Multimedia Modeling, Springer, 2009, pp. 435–446.
[244] A. Rabie, B. Wrede, T. Vogt, M. Hanheide, Evaluation and discussion of multi-modal emotion recognition, in: Computer and Electrical Engineering, 2009. ICCEE'09. Second International Conference on, 1, IEEE, 2009, pp. 598–602.
[245] S.K. D'Mello, A. Graesser, Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features, User Model. User-Adapted Interact. 20 (2) (2010) 147–187.
[246] M.L.I.C. Dy, I.V.L. Espinosa, P.P.V. Go, C.M.M. Mendez, J.W. Cu, Multimodal emotion recognition using a spontaneous filipino emotion database, in: Human-Centric Computing (HumanCom), 2010 3rd International Conference on, IEEE, 2010, pp. 1–5.
[247] R. Gajsek, V. Štruc, F. Mihelic, Multi-modal emotion recognition using canonical correlations and acoustic features, in: Proceedings of the 2010 20th International Conference on Pattern Recognition, IEEE Computer Society, 2010, pp. 4133–4136.
[248] L. Kessous, G. Castellano, G. Caridakis, Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis, J. Multimodal User Interfaces 3 (1-2) (2010) 33–48.
[249] J. Kim, F. Lingenfelser, Ensemble approaches to parametric decision fusion for bimodal emotion recognition., in: BIOSIGNALS, 2010, pp. 460–463.
[250] M. Mansoorizadeh, N.M. Charkari, Multimodal information fusion application to human emotion recognition from face and speech, Multimedia Tools Appl. 49 (2) (2010) 277–297.
[251] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, S.S. Narayanan, et al., Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling., in: INTERSPEECH, 2010, pp. 2362–2365.
[252] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, et al., Multiple classifier systems for the classification of audio-visual emotional states, in: Affective Computing and Intelligent Interaction, Springer, 2011, pp. 359–368.
[253] N. Banda, P. Robinson, Noise analysis in audio-visual emotion recognition, in: International Conference on Multimodal Interaction, Alicante, Spain, Citeseer, 2011, pp. 1–4.
[254] G. Chanel, C. Rebetez, M. Bétrancourt, T. Pun, Emotion assessment from physiological signals for adaptation of game difficulty, Syst. Man Cybern. Part A 41 (6) (2011) 1052–1063.
[255] D.R. Cueva, R.A. Gonçalves, F. Cozman, M.R. Pereira-Barretto, Crawling to improve multimodal emotion detection, in: Advances in Soft Computing, Springer, 2011, pp. 343–350.
[256] D. Datcu, L.J. Rothkrantz, Emotion recognition using bimodal data fusion, in: Proceedings of the 12th International Conference on Computer Systems and Technologies, ACM, 2011, pp. 122–128.
[257] D. Jiang, Y. Cui, X. Zhang, P. Fan, I. Ganzalez, H. Sahli, Audio visual emotion recognition based on triple-stream dynamic bayesian network models, in: Affective Computing and Intelligent Interaction, Springer, 2011, pp. 609–618.
[258] F. Lingenfelser, J. Wagner, E. André, A systematic discussion of fusion techniques for multi-modal affect recognition tasks, in: Proceedings of the 13th international conference on multimodal interfaces, ACM, 2011, pp. 19–26.
[259] M.A. Nicolaou, H. Gunes, M. Pantic, Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space, Affective Comput. IEEE Trans. 2 (2) (2011) 92–105.
[260] H.A. Vu, Y. Yamazaki, F. Dong, K. Hirota, Emotion recognition based on human gesture and speech information using rt middleware, in: Fuzzy Systems (FUZZ), 2011 IEEE International Conference on, IEEE, 2011, pp. 787–791.
[261] J. Wagner, E. Andre, F. Lingenfelser, J. Kim, Exploring fusion methods for multimodal emotion recognition with missing data, Affective Comput. IEEE Trans. 2 (4) (2011) 206–218.
[262] S. Walter, S. Scherer, M. Schels, M. Glodek, D. Hrabal, M. Schmidt, R. Böck, K. Limbrecht, H.C. Traue, F. Schwenker, Multimodal emotion classification in naturalistic user behavior, in: Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments, Springer, 2011, pp. 603–611.
[263] M. Hussain, H. Monkaresi, R.A. Calvo, Combining classifiers in multimodal affect detection, in: Proceedings of the Tenth Australasian Data Mining Conference-Volume 134, Australian Computer Society, Inc., 2012, pp. 103–108.
[264] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: a database for emotion analysis; using physiological signals, Affective Comput. IEEE Trans. 3 (1) (2012) 18–31.
[265] J.-C. Lin, C.-H. Wu, W.-L. Wei, Error weighted semi-coupled hidden markov model for audio-visual emotion recognition, Multimedia, IEEE Trans. 14 (1) (2012) 142–156.
[266] K. Lu, Y. Jia, Audio-visual emotion recognition with boosted coupled hmm, in: Pattern Recognition (ICPR), 2012 21st International Conference on, IEEE, 2012, pp. 1148–1151.
[267] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, S. Narayanan, Context-sensitive learning for enhanced audiovisual emotion classification, Affective Comput. IEEE Trans. 3 (2) (2012) 184–198.
[268] J.-S. Park, G.-J. Jang, Y.-H. Seo, Music-aided affective interaction between human and service robot, EURASIP J. Audio, Speech, Music Process. 2012 (1) (2012) 1–13.
[269] M. Rashid, S. Abu-Bakar, M. Mokji, Human emotion recognition from videos using spatio-temporal and audio features, Visual Comput. 29 (12) (2013) 1269–1275.
[270] M. Soleymani, M. Pantic, T. Pun, Multimodal emotion recognition in response to videos, Affective Comput. IEEE Trans. 3 (2) (2012) 211–223.
[271] B. Tu, F. Yu, Bimodal emotion recognition based on speech signals and facial expression, in: Foundations of Intelligent Systems, Springer, 2012, pp. 691–696.
[272] T. Baltrusaitis, N. Banda, P. Robinson, Dimensional affect recognition using continuous conditional random fields, in: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, IEEE, 2013, pp. 1–8.
[273] S. Dobrišek, R. Gajšek, F. Mihelič, N. Pavešić, V. Štruc, Towards efficient multi-modal emotion recognition, Int. J. Adv. Robotic Sy. 10 (53) (2013).
[274] M. Glodek, S. Reuter, M. Schels, K. Dietmayer, F. Schwenker, Kalman filter based classifier fusion for affective state recognition, in: Multiple Classifier Systems, Springer, 2013, pp. 85–94.
[275] S. Hommel, A. Rabie, U. Handmann, Attention and emotion based adaption of dialog systems, in: Intelligent Systems: Models and Applications, Springer, 2013, pp. 215–235.
[276] G. Krell, M. Glodek, A. Panning, I. Siegert, B. Michaelis, A. Wendemuth, F. Schwenker, Fusion of fragmentary classifier decisions for affective state recognition, in: Multimodal Pattern Recognition of Social Signals in Human–Computer-Interaction, Springer, 2013, pp. 116–130.
[277] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, G. Rigoll, Lstm-modeling of continuous emotions in an audiovisual affect recognition framework, Image Vision Comput. 31 (2) (2013) 153–163.
[278] L. Chen, S.-Y. Yoon, C.W. Leong, M. Martin, M. Ma, An initial analysis of structured video interviews by using multimodal emotion detection, in: Proceedings of the 2014 workshop on Emotion Representation and Modelling in Human-Computer-Interaction-Systems, ACM, 2014, pp. 1–6.
[279] S. Poria, E. Cambria, N. Howard, G.-B. Huang, A. Hussain, Fusing audio, visual and textual clues for sentiment analysis from multimodal content, Neurocomputing 174 (2016) 50–59.
[280] M. Song, J. Bu, C. Chen, N. Li, Audio-visual based emotion recognition-a new approach, in: Computer Vision and Pattern Recognition, 2, IEEE, 2004, pp. II–1020.
[281] Z. Zeng, Y. Hu, M. Liu, Y. Fu, T.S. Huang, Training combination strategy of multi-stream fused hidden markov model for audio-visual affect recognition, in: Proceedings of the 14th annual ACM international conference on Multimedia, ACM, 2006, pp. 65–68.
[282] S. Petridis, M. Pantic, Audiovisual discrimination between laughter and speech, in: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, IEEE, 2008, pp. 5117–5120.
[283] P.K. Atrey, M.A. Hossain, A. El Saddik, M.S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, Multimedia Syst. 16 (6) (2010) 345–379.

[284] A. Corradini, M. Mehta, N.O. Bernsen, J. Martin, S. Abrilian, Multimodal input fusion in human-computer interaction, NATO Sci. Ser.s Sub Ser. III Comput.Syst. Sci. 198 (2005) 223.
[285] G. Iyengar, H.J. Nock, C. Neti, Audio-visual synchrony for detection of monologues in video archives, in: Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on, 1, IEEE, 2003, pp. I–329.
[286] W. Adams, G. Iyengar, C.-Y. Lin, M.R. Naphade, C. Neti, H.J. Nock, J.R. Smith, Semantic indexing of multimedia content using visual, audio, and text cues, EURASIP J. Adv. Signal Process. 2003 (2) (2003) 1–16.
[287] A.V. Nefian, L. Liang, X. Pi, X. Liu, K. Murphy, Dynamic bayesian networks for audio-visual speech recognition, EURASIP J. Adv. Signal Process. 2002 (11) (2002) 1–15.
[288] K. Nickel, T. Gehrig, R. Stiefelhagen, J. McDonough, A joint particle filter for audio-visual speaker tracking, in: Proceedings of the 7th international conference on Multimodal interfaces, ACM, 2005, pp. 61–68.
[289] I. Potamitis, H. Chen, G. Tremoulis, Tracking of multiple moving speakers with multiple microphone arrays, Speech Audio Process. IEEE Trans. 12 (5) (2004) 520–529.
[290] H. Gunes, M. Pantic, Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners, in: International conference on intelligent virtual agents, 2010, pp. 371–377.
[291] M.F. Valstar, T. Almaev, J.M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, J.F. Cohn, Fera 2015-second facial expression recognition and analysis challenge, in: Automatic Face and Gesture Recognition, 6, 2015, pp. 1–8.
[292] M.A. Nicolaou, H. Gunes, M. Pantic, Automatic segmentation of spontaneous data using dimensional labels from multiple coders, in: Proceedings of LREC Int'l Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, 2010, pp. 43–48.
[293] K.-h. Chang, D. Fisher, J. Canny, Ammon: a speech analysis library for analyzing affect, stress, and mental health on mobile phones, Proc. PhoneSense 2011 (2011).
[294] S. Zhang, L. Li, Z. Zhao, Audio-visual emotion recognition based on facial expression and affective speech, in: Multimedia and Signal Processing, Springer, 2012, pp. 46–52.
[295] F. Eyben, M. Wöllmer, M.F. Valstar, H. Gunes, B. Schuller, M. Pantic, String-based audiovisual fusion of behavioural events for the assessment of dimensional affect, in: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, IEEE, 2011, pp. 322–329.
[296] T. Rahman, C. Busso, A personalized emotion recognition system using an unsupervised feature adaptation scheme, in: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2012, pp. 5117–5120.
[297] Q. Jin, C. Li, S. Chen, H. Wu, Speech emotion recognition with acoustic and lexical features, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 4749–4753.
[298] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, R. Prasad, Ensemble of svm trees for multimodal emotion recognition, in: Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia–Pacific, IEEE, 2012, pp. 1–4.
[299] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, et al., Simsensei kiosk: a virtual human interviewer for healthcare decision support, in: Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, 2014, pp. 1061–1068.
[300] M.E. Hoque, R.W. Picard, Acted vs. natural frustration and delight: many people smile in natural frustration, in: Automatic Face & Gesture Recognition and Workshops, IEEE, 2011, pp. 354–359.