Memetic Comp. (2018) 10:3–13
https://doi.org/10.1007/s12293-017-0228-3

REGULAR RESEARCH PAPER

Ensemble application of ELM and GPU for real-time multimodal sentiment analysis

Ha-Nguyen Tran[1] · Erik Cambria[1]

Received: 29 September 2016 / Accepted: 17 March 2017 / Published online: 24 March 2017
© Springer-Verlag Berlin Heidelberg 2017

Abstract The enormous number of videos posted everyday on multimedia websites such as Facebook and YouTube makes the Internet an infinite source of information. Collecting and processing such information, however, is a very challenging task as it involves dealing with a huge amount of information that is changing at a very high speed. To this end, we leverage on the processing speed of extreme learning machine and graphics processing unit to overcome the limitations of standard learning algorithms and central processing unit (CPU) and, hence, perform real-time multimodal sentiment analysis, i.e., harvesting sentiments from web videos by taking into account audio, visual and textual modalities as sources of the information. For the sentiment classification, we leveraged on sentic memes, i.e., basic units of sentiment whose combination can potentially describe the full range of emotional experiences that are rooted in any of us, including different degrees of polarity. We used both feature and decision level fusion methods to fuse the information extracted from the different modalities. Using the sentiment annotated dataset generated from YouTube video reviews, our proposed multimodal system is shown to achieve an accuracy of 78%. In term of processing speed, our method shows improvements of several orders of magnitude for feature extraction compared to CPU-based counterparts.

Keywords Multimodal sentiment analysis · Opinion mining · Multimodal fusion · GPGPU · Sentic computing

✉ Ha-Nguyen Tran
hntran@ntu.edu.sg

Erik Cambria
cambria@ntu.edu.sg

[1] School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

1 Introduction

Sentiment analysis is the task of automatically classifying the polarity (positive or negative) of a given text at the document, phrase or sentence levels. The field of sentiment classification has recently become an attractive research direction due to a large number of real-world applications which require identifying human opinions for better decision-making. Most of the recent works on sentiment analysis are based on natural language processing techniques in which the detection of emotions is conducted from humanly created textual data and resources including lexicons or large annotated datasets [27]. With the rapid growth of social media websites such as Facebook and YouTube, people are expressing their opinions in various forms which include videos, images, and audios. Compared to the textual data, these resources might provide more valuable information through richer channels such as the tones of speakers and facial expressions. As a result, the necessity of analyzing and understanding on-line generated data from multimodal cues has arisen in recent years [20,21].

Collecting and processing such information, however, are very challenging tasks as they involve in dealing with a huge amount of information that is changing at a very high speed. Recently, general purpose GPUs have become popular computing devices owing to their massively parallel processing architectures which are successfully used as efficient accelerators to leverage the performance of big data analytics [24]. Extreme learning machine (ELM) [10] has also become an efficient technique for data mining tasks. ELMs are feed-forward neural networks (SLFNs) for classification or regression with a single-hidden layer which randomly chooses hidden nodes and analytically determines the output weights of SLFNs. The advantages of ELMs are fast learning speed and ease of implementation [13,29]. Inspired by those

Interval	Pleasantness	Attention	Sensitivity	Aptitude
[G(1), G(2/3))	ecstasy	vigilance	rage	admiration
[G(2/3), G(1/3))	joy	anticipation	anger	trust
[G(1/3), G(0))	serenity	interest	annoyance	acceptance
(G(0), G(–1/3)]	pensiveness	distraction	apprehension	boredom
(G(–1/3), G(–2/3)]	sadness	surprise	fear	disgust
(G(–2/3), G(–1)]	grief	amazement	terror	loathing

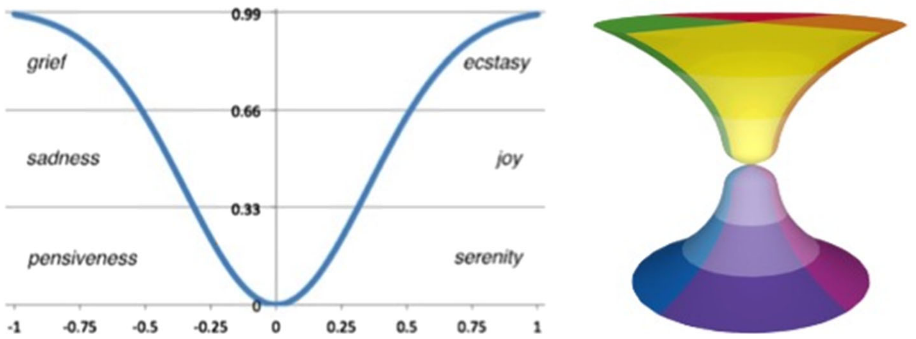

Fig. 1 The Hourglass of emotions: by leveraging on multiple (polarized) activation levels for each affective dimension (sentic memes), the Hourglass model covers cases where up to four emotions can be expressed at the same time and allows for reasoning on them in an algebraic manner

promising techniques, we investigate whether and how GPU and ELM can be leveraged to accelerate the task of real-time multimodal sentiment analysis from audio, visual and textual sources.

In this paper, we propose a multimodal framework for sentiment analysis that leverages on the ensemble application of ELM and GPU to ensure fast processing and on sentic memes to increase the accuracy of the sentiment classification. As defined in the Hourglass model [4], sentic memes are basic units of sentiment that regulate the activation of four different but concomitant affective dimensions (namely Pleasantness, Attention, Sensitivity and Aptitude) to describe the full range of human emotions (Fig. 1). Similarly, sentic memes enable fine-grained polarity classification (as a linear combination of sentic memes). For the real-time processing part, our method takes advantage of the massively parallel processing power of modern GPUs to enhance the performance of feature extraction from different modalities. In addition, powerful ELM classifiers are applied to build the sentiment analysis model based on the extracted features. To highlight the efficiency of our solution, we conducted our experiments on the YouTube dataset and achieved an accuracy of 78% which outperforms all previous systems. In term of processing speed, our method shows improvements of several orders of magnitude for feature extraction compared to CPU-based counterparts.

The rest of the paper is organized as follows: Sect. 2 covers related works on emotion and sentiment recognition from different modalities; Sect. 3 proposes an overview of the multimodal sentiment analysis; next, Sect. 4 explains how visual, audio and textual data are processed using GPUs; Sect. 5 illustrates the methodology adopted for fusing different modalities; Sect. 6 presents experimental results; finally, Sect. 7 concludes the paper and outlines future work.

2 Related works

In this section, we present an overview of state-of-the-art studies in the fields of (1) textual sentiment analysis; and (2) audio–video emotion identification.

2.1 Text-based sentiment analysis

The techniques developed so far for subjectivity and sentiment analysis have mainly focused on textual data. These approaches consist of either rule-based techniques that take advantage of opinion lexicon [27], or statistical methods that assume the availability of a large dataset annotated with emotion labels [18].

For rule-based classifiers, significant works have been conducted to automatically identify positive, negative, or neutral sentiment associated with different levels, from opinion words [25] to more linguistically complex phrases [28]. Several methods [1] go beyond this to further explore how to automatically identify fine-grained emotions (for example, anger, joy, surprise, fear, disgust, and sadness) that are either explicitly or implicitly expressed within the textual data. On the other hand, the data-driven methods make use of large datasets manually annotated for opinions which might cover the domain of product reviews, news articles or newspaper headlines [3]. Many supervised as well as unsupervised classifiers have been constructed to recognize textual emotional content [30]. The SNoW architecture [6] is one of the widely applied frameworks to detect emotion in textual data.

Over the last few years, a large number of researchers have been working on sentiment extraction from texts of various genres, e.g., product reviews, blogs, essays, and more [14,19]. Sentiment extraction from social media allows us to make useful predictions in many aspects, e.g., the popularity of a product release, and the results of an election poll. To accomplish this, there have been several knowledge-based sentiment and emotion lexicons for word- and phrase-level sentiment and emotion analysis. Cambria et al. [5] introduced a novel commonsense knowledge base, SenticNet, which assigns polarity values to 50,000 multi-word expressions for concept-level sentiment analysis.

2.2 Audio–video emotion analysis

Over the last few years, we have witnessed a lot of researches [2,23] in emotion recognition which address the problems of facial expression detection and/or audio affect recognition. Audio affect recognition of speech signal aims to identify the emotional states of humans by analyzing their voices. Many speech-based emotion analyses [11,17] have given high attention to identifying several acoustic features such as intensity of utterance, bandwidth, duration, fundamental frequency (pitch), and Mel frequency coefficients [9]. To accelerate the performance of audio feature extraction, Michalek et al. [15] released an open-source tool which is based on parallel processing on General Purpose GPUs.

There are also studies that analyze the visual cues such as facial expression and body movement. One of the most significant works on facial expressions has been done by Ekman et al. [8]. According to this study, universal facial expressions can be sufficiently employed as clues for detecting emotions. They considered anger, sadness, surprise, fear, disgust, and joy as six basic emotion classes, and claimed that such categories can sufficiently describe most of the emotions expressed by facial expressions. Ekman et al. also introduced a facial expression coding system (FACS) to encode facial expressions by dismantling a facial expression into a collection of action units (AU), where the AUs are defined via some particular facial muscle movements. An AU essentially consists of three basic parts: AU number, FACS name, and muscular basis. Some common examples of approaches that use FACS to understand expressed facial expressions are active appearance model [22] and active shape model [12]. By employing the AUs as features (like k-nearest-neighbors, Bayesian networks, hidden Markov models (HMM), and artificial neural networks (ANN) [23]), many research works have successfully managed to infer emotions from facial expressions.

In addition to the above studies which only focused on individual audio or video modalities, there is a growing body of works that include both video and audio emotion recognition [2,23]. The features used by those methods are mainly low level features, such as tracking points for collection visual data, or extracting audio features at pitch level.

3 GPU–ELM multimodal sentiment analysis

Figure 2 shows an overview of our real-time GPU–ELM multimodal sentiment analysis approach, named *GESentic*. At the beginning, our system receives the multimodal data from users. The data can be created by the users expressing their opinions in front of the camera. After that, our system splits the obtained video into small segments. A transcriber is used to extract the text transcription of the audio.

In the vector stream generation section, the multimodal data consisting of visual, audio and text content are processed on the GPU. In the facial expression analyzer, we employ a GPU-based active shape model (ASM) method to track 66 facial points. In the algorithm, we first apply the sketch based facial expression recognition technique [23] to speed up the convergence of ASM searching. Then, we accelerate the performance of the ASM tracking by processing multiple sketched images simultaneously. The relevant achieved points are then used to construct facial features. The important acoustic features such as MFCC, spectral centroid, and pitch are also extracted on the GPU for emotion recognition by employing and extending an open source tool of Michalek et al. [15]. In our system, we apply a block-based approach to process multiple windows concurrently. For textual data, we propose a GPU-based SenticNet engine to select only important features and aspects of human opinions. In the engine, commonsense knowledge bases such as ConceptNet and SenticNet are employed to automatically construct the

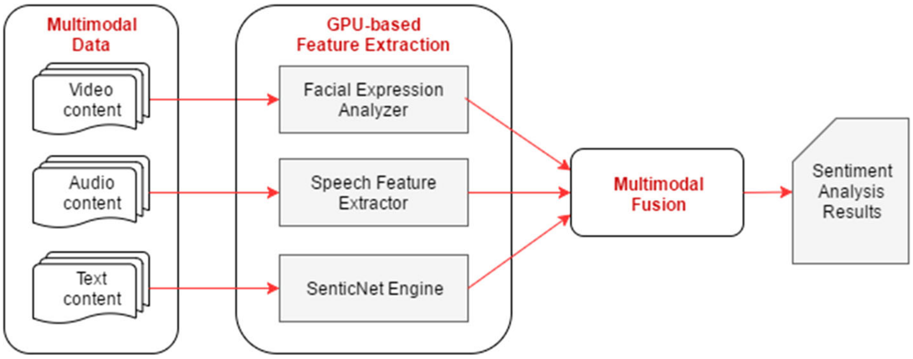

Fig. 2 Overview of GPU-based multimodal sentiment analysis

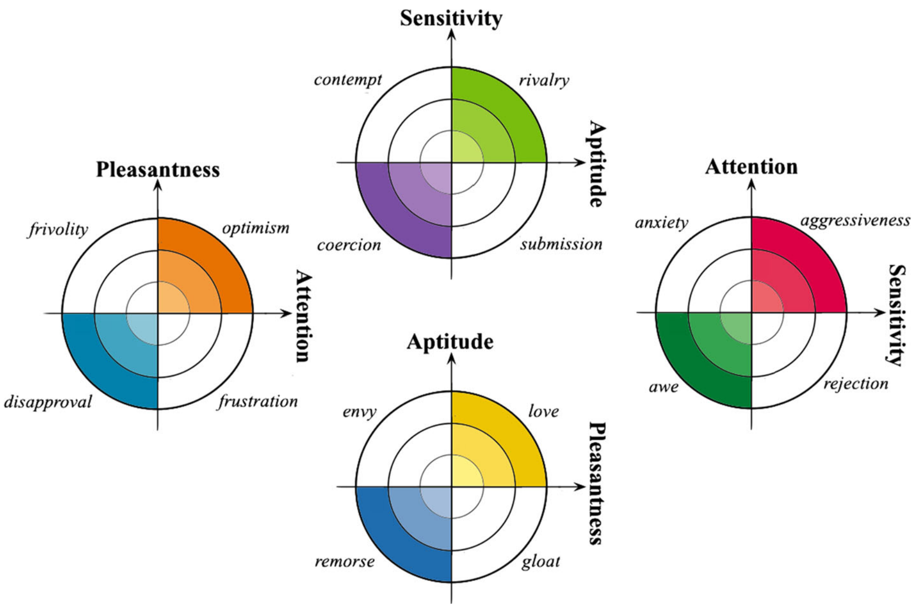

Fig. 3 Possible combinations of emotions generated by the simultaneous pair-wise activation of two affective dimensions (sentic memes) in the Hourglass model

ontology containing domain-specific commonsense knowledge. The obtained ontology is used to track the important review aspects as well as opinion words in the input texts. In particular, each extracted concept (single- and multi-word expressions) is mapped to a 100-dimensional sentic vector [4] specifying its affective valence and its polarity, both defined as a linear combination of sentic memes. An example of how sentic memes enable the generation of basic compound emotions (that is, emotions arising from the concomitant activation of only two affective dimensions) is provided in Fig. 3.

Later, the feature vectors of visual, audio and textual modalities are fused in the multimodal fusion module. The resemble vectors are then used to classify segments of the input video into sentiment classes. To be more formal, let the input signal $S = (s_v, s_a, s_t)$

$$\begin{aligned} R &= \textit{sentiment\_analysis}(S) \\ &= \textit{fusion}(\textit{face}(s_v), \textit{speech}(s_a), \textit{sentic}(s_t)) \end{aligned}$$

4 Feature extraction

In this section, we explain how visual, audio and textual features are automatically extracted from multimodal data using the GPU-based techniques. For each data source, we introduce an efficient module which is specially designed for fast extracting its features on GPUs.

4.1 Visual feature extraction

Humans are known to express emotions in different ways through the face. Facial expressions can provide important clues for identifying emotions, which we use to complement the linguistic and acoustic features. As a result, automating facial expression analysis could bring facial expressions into man–machine interaction as a new modality for identifying subjectivity and sentiment analysis (i.e., classifying emotional clues into positive, negative and neutral). Visual feature extraction is an important step of subjectivity and sentiment analysis which automatically extracts from the video sequences.

The annotated YouTube dataset provided by Morency et al. [16] is used as our visual training dataset. Each video was split into small segments, each of which had the length of several seconds. Each segmented part is labeled as positive, neutral and negative sentiments by using either 1, 0 and −1 respectively.

In order to extract the visual features, we first convert all videos in the dataset to image frames. After that, we extract facial feature points from each image frame. Table 1 lists some relevant facial feature points out of the 66 feature points that we use to construct facial features. These features are defined as the distances between feature points; see example in Table 2. To detect feature points from the images, we use active shape model (ASM) [7] which is a popular statistical model for facial feature detection. However, the performance of ASM searching might be very poor when the input image is noisy. In addition, the tracking result is very sensitive to the initial location and size. Noise elimination and smoothing algorithms are very important to improve the tracking accuracy and efficiency. These operations are, however, in general computationally expensive [26]. To overcome this issue, we introduce a GPU-based approach to enhance the performance of the ASM method for facial feature extraction. At the beginning, images and models are transferred into the global memory of the GPU. After that, we perform

Table 1 Relevant facial feature points

Features	Description
0	Left eye
1	Right eye
23	Outer corner of the left eye
24	Inner corner of the left eye
38	Bottom of the left eye
35	Top of the left eye
29	Left iris corner of the left eye
30	Right iris corner of the left eye
12	Outer corner of the left eyebrow
16	Left eyebrow middle
13	Inner corner of the left eyebrow
25	Inner corner of the right eye
26	Outer corner of the right eye
41	Bottom of the right eye
40	Top of the right eye
33	Left iris corner of the right eye
34	Right iris corner of the right eye
14	Outer corner of the right eyebrow
17	Right eyebrow middle
15	Inner corner of the right eyebrow
54	Mouth top
55	Mouth bottom

Table 2 List of facial features

Facial features
Eye separation
Width of the left eye
Height of the left eye
Width of the left iris eye
Width of the right eye
Height of the right eye
Width of the right iris eye
Width of the left eyebrow
Width of the right eyebrow
Mouth width

GPU-based active shape model in two main steps: (1) sketch generation and (2) ASM model matching. The overview of the ASM-based feature points detection on the GPU is illustrated in Fig. 4.

The sketch generation implemented on the GPU follows two consecutive operations, namely edge detection and tone mapping, which are similar to the GASM algorithm used in [23]. A thread-based approach is applied to detect and sharpen the edge of the contour of the face components. In this stage, we convert the color of each pixel to a luminance value, and then calculate the square diagonal differences which are summed up afterward. After reverting the results, we multiply them by a large number to make the values visible. To achieve the best performance, we use the shared memory of GPUs to store data points during the computation. However, the result obtained from the edge detection and enhancement still remains noisy, which makes the latter ASM tracking process slow and unreliable. The next step will further eliminate the high-frequency noise by changing the tonal range.

In the ASM model matching, the sketch images (rather than the original images) are used as data inputs. Since the edges in the sketch images are much stronger than the original ones, the model matching process converges more quickly. To execute parallel ASM searching in many images synchronously in the GPU [12], we first need to transform the scale of the images according to the size of the detected face. The approach to accelerate the time-consuming image scaling is to make use of the texture memory of the GPU and bilinear interpolation of texture. The GPU-based parallel ASM searching algorithm is similar to the iterative CPU-based method. First, a region of the image around each point is sampled and calculated to find the best match. Then, the model parameters are updated to fit the new search shape. Finally, the shapes are conformed by the parameters. After each iterative computation, the appropriate shapes are transferred to CPU for displaying and further processing. Unlike the traditional ASM, however, the searching process is calculated with GPU on multiple images concurrently.

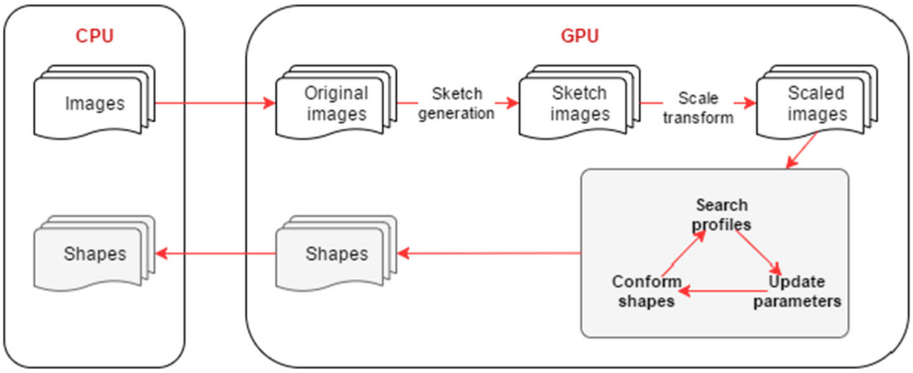

Fig. 4 Overview of GPU-based feature extraction

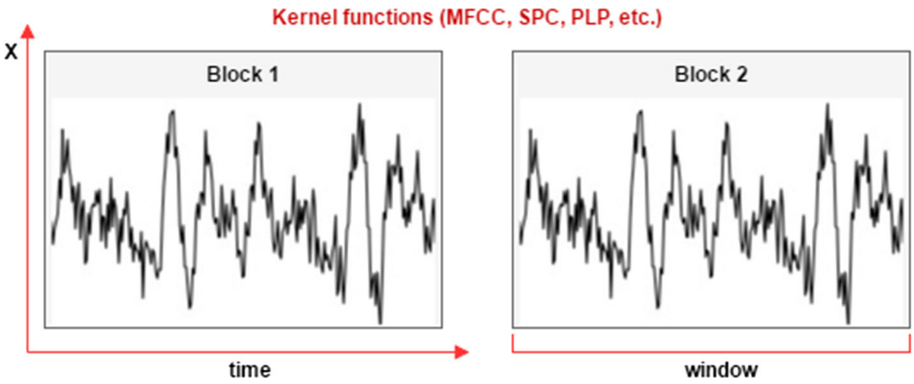

Fig. 5 Block-based audio feature extraction

4.2 Acoustic feature extraction

Besides visual information, the audio data also reflects the speaker's emotion. Thus, in the second module our method automatically extracts the audio features from each annotated segments of the videos. For accelerating the audio feature extraction processing, we optimize and extend the GPU-based implementation of Michalek et al. [15]. In the approach, all windows are stored into the device memory and then copied to the fast shared on-chip memory at the beginning of each kernel. Due to the data independence between windows, we process the windows concurrently on different thread blocks, as can be seen in Fig. 5. The audio features include several statistic measures, e.g., max and min value, standard deviation, variance etc. of some key feature groups. In order to extract those features in parallel, we apply some CUDA core primitives such as reduction and scan[1] to efficiently compute the results in the GPU. Some of the important audio features are described as follows:

- *Mel frequency cepstral coefficients (MFCCs)* MFCCs are commonly used as features in speech recognition systems. The advantage of the features is to represent the amplitude of the spectrum in a compact form. The MFCCs are computed by using short time Fourier transform (STFT). The first step is the computation of the frequency domain representation of the input signal. In the next processing step, the mel-frequency spectrum is then computed by filtering and executing the logarithm through the perceptually motivated mel-frequency scale filtering.
- *Spectral centroid (SPC)* the spectral centroid is commonly associated with the measure of the brightness of a sound. This measure is obtained by evaluating the center of gravity using the Fourier transforms frequency and magnitude information. The individual centroid of a spectral frame is defined as the average frequency weighted by amplitudes, divided by the sum of the amplitudes:

$$C_i = \frac{\sum_{i=0}^{n} n M_i[n]}{\sum_{i=0}^{n} M_i[n]}$$

Here $M_i[n]$ denotes the magnitude of the Fourier transform at frequency bin n and frame i.
- *Spectral flux* spectral flux is a measure of how quickly the power spectrum of a signal is changing. It is calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame. In particular, it is calculated as the Euclidean distance between the two normalized spectra. The spectral flux can be used to determine the timbre of an audio signal.
- *Beat histogram (BH)* the beat histogram shows the distribution of various beat periodicities of the signal. BH is calculated as the auto-correlation of the root mean square by taking the FFT of the result.
- *Beat sum (BS)* this feature is a good measure for identifying the important roles of regular beats in a signal. Beat Sum is calculated by finding the sum of all values in the beat histogram.
- *Strongest beat* this feature is defined as the strongest beat in a signal, in the beats per min, which is found by finding the strongest bin in the beat histogram.
- *Pause duration* given the start and end time of the utterance, how many audio samples are identified as silence. This audio feature is then normalized by the number of audio samples in this utterance. This feature can be interpreted as the percentage of the time where the speaker was silent.
- *Pitch* the feature is computed by the standard deviation of the pitch level for a spoken segment. This measure represents the variation of voice intonation during the same utterance.
- *Voice quality* voice quality is defined as the characteristic auditory coloring of the voice of the speaker, derived from a variety of laryngeal and supralaryngeal features and running continuously through the speaker's speech.

[1] https://nvlabs.github.io/moderngpu/scan.html.

During the feature extraction, we use a 30 Hz frame-rate with windows of 100 ms. The features are averaged over all the frames to obtain one feature vector for each utterance.

4.3 Textual feature extraction

Sentiment analysis from unstructured natural language text is a challenging problem which has recently received considerable attention from the research community. In this paper, we present a commonsense-based method to identify both opinion targets and emotional words. Inspired by the promising results of GPU-based commonsense reasoning [24] and commonsense-based sentiment analysis [4], we introduce a method to extract important features based on commonsense knowledge bases in parallel on the GPU. Figure 6 represents the flow diagram of the approach.

4.3.1 Domain-specific knowledge graph generation

At first, we represent the commonsense knowledge bases as directed graphs. In addition to the node/edge arrays used in GpSense [24], we create a value array to maintain the string values of concepts and another array whose elements point to the offsets of values in the value array. The advantage of storing the value array is that we can directly search for the concept names on the GPU. Next, we transfer the data to the GPU memory. However, the commonsense knowledge graph still contains many irrelevant concepts which might not be used during the feature extraction process. In the initial step, we aim to filter out the concepts which are not relevant to the product reviews domain. We first take some key concepts as the inputs and then explore the knowledge graph to find the domain-specific concepts. The GPU implementation of this graph exploration process is similar to the Filtering phase in GpSense algorithm. We extract the important concepts from commonsense knowledge bases up to 3–4 levels.

4.3.2 Key concepts extraction

In order to extract important concepts from the given textual data, we check whether the data contains concepts in the generated domain-specific commonsense knowledge graph. To implement the task on the GPU, we detect words in the textual data in following steps: (1) the separate symbols are located in parallel; (2) the offsets of words are maintained in an intermediate array by using the CUDA compaction algorithm. After that, our system finds the segmented words in the commonsense knowledge graph.

In the next step, we identify the opinion words associated with the domain-specific features. We then build the sentiment lexicon with the polarity values based on the publicly available resources, namely, SenticNet and EmoSenticNet. After identifying the polarities of the opinion words, we may need to change their values if any negative words such as “not” are discovered nearby. Thus, if we find “not” within two words from a positive opinion word w, we must reverse its polarity value from positive to negative.

5 Multimodal fusion

In the previous sections, we introduce three GPU-accelerated modules for fast extracting important features from textual, audio and visual modalities. The results obtained from those feature extractors are individual feature vectors. This section discusses the multimodal fusion, the most important part of any multimodal sentiment analysis engine, to integrate these features. In particular, we employ the two most common fusion techniques proposed in recent years, namely feature-level and decision-level fusion.

5.1 Feature-level fusion

In this approach, the feature sets originating from multiple input sources are consolidated into a single feature set. In the other word, we concatenate three individual feature vectors into a single long vector of multimodal features. Despite its simplicity, the method has been shown to produce significantly high accuracy. Later, the resulting feature vector is taken as an input of classifiers for labeling video segments with sentiment classes. In our experiments, we use tenfold cross validation to assess the precision.

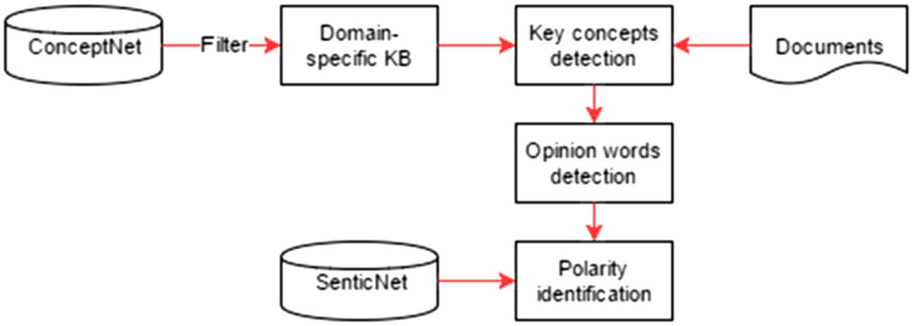

Fig. 6 Commonsense-based text feature extraction

5.2 Decision-level fusion

Unlike the feature concatenation approach in the feature-level fusion, the decision-level method takes visual, audio and textual feature vectors as the separate inputs for different classifiers. The results generated from each classifier are the probability scores associated with sentiment classes (i.e., positive, negative, and neutral). To produce the final sentiment label, we employ a weighted voting strategy in which each classifier is assigned a weight, and the label receiving the greatest total weighted vote is selected. The approach is formalized in the following formula:

$$l' = \arg\max_{i} \left(w_1 s_i^a + w_2 s_i^v + w_3 s_i^t \right), \quad i = 1, 2, 3, \ldots, C$$

where w_1, w_2 and w_3 are the weights assigned for the employed classifiers of three modalities. In our experiments, we set the values of w_1, w_2 and w_3 to 0.33. C is the number of sentiment classes. s_i^a, s_i^v and s_i^t denote the probability scores for the audio, visual and textual classifiers respectively.

6 Experiments and discussion

In this section, we discuss experiment results of our GPU–ELM multimodal sentiment analysis system on YouTube dataset using both decision-level and feature-level fusions. Then, we describe the performance of GPU-based feature extraction as well as analyze the importance of extracted features used in our classification processes.

6.1 Datasets

6.1.1 YouTube dataset

Our multimodal sentiment analyzer employs the video dataset extracted from the YouTube website by Morency et al. [16] as the main dataset in our experiments for both building the multimodal sentiment analysis system and evaluating the performance. The YouTube dataset consists of 47 videos of various topics such as product review, or politics opinions. In the dataset, there are 20 videos spoken by females and 27 videos from male speakers randomly chosen from YouTube, with the age range from 14 to 60 year-olds. Each video has the length of 2–5 min and is formatted in MP4 with a standard size of 360 × 480. The text transcriptions are also provided for YouTube videos. Each video is split into small segments, each of which is labeled by a sentiment class. Because of the annotation scheme of this dataset, the textual data is also available for our experiments.

Table 3 Sample of SenticNet data

Concept	Polarity
A_lot	+0.258
A_lot_of_food	+0.858
Abandon	−0.566
Abase	−0.153
Abash	−0.174
Abashed	−0.174
Abashment	−0.186
Abhor	−0.391
Abhorrence	−0.391

Table 4 Results of feature-level fusion

	Precision	Recall
Textual modality	0.62	0.59
Audio modality	0.65	0.67
Video modality	0.68	0.68
Visual and textual modalities	0.72	0.72
Visual and audio modalities	0.73	0.73
Audio and textual modalities	0.71	0.71
Visual, audio and textual modalities	0.78	0.77

6.1.2 SenticNet datasets

As for prior polarity lexicons of concepts, we employ the latest version of the SenticNet commonsense knowledge base, i.e., SenticNet 4.0 [5], which consists of more than 50,000 natural language concepts. Each concept is associated with a polarity score which is a floating number between −1.0 and +1.0, where −1.0 is extreme negativity and +1.0 is extreme positivity. Table 3 shows several concepts with their polarity scores in the SenticNet knowledge base.

6.2 Feature-level and decision-level fusions

In our experiments, we employ different supervised classifiers such as Naïve Bayes, support vector machine (SVM), artificial neural networks (ANN), and extreme machine learning (ELM) to measure the accuracy of our GPU-based multimodal sentiment analysis system on the YouTube dataset. The best accuracy, however, is obtained by applying ELM classifiers.

Tables 4 and 5 show the classification performances of different models: text-only, visual-only, audio-only, text–visual, visual–audio, audio–text, and trimodal integration. Results on feature-level fusion are presented in Table 4 while Table 5 illustrates the experiment results of decision-level fusion. Clearly, the accuracy improves when we use audio, visual and textual modalities together in the experiment. These improvements are observed for both precision and recall

Table 5 Results of decision-level fusion

	Precision	Recall
Textual modality	0.59	0.58
Audio modality	0.62	0.65
Video modality	0.67	0.66
Visual and textual modalities	0.68	0.68
Visual and audio modalities	0.71	0.70
Audio and textual modalities	0.66	0.66
Visual, audio and textual modalities	0.75	0.73

Table 6 Comparison with Morency's method

	Decision-level	Feature-level	Morency's method
Textual modality	0.59	0.62	0.43
Audio modality	0.62	0.65	0.41
Video modality	0.67	0.68	0.45
Trimodal integration	0.75	**0.78**	0.54

Bold value indicates the best result on trimodal integration among methods

in comparison with text-only, visual-only, and audio-only modalities.

Compared to the experiment results reported in the work of Morency et al. [16], our method outperforms the approach in term of precision, as can be seen in Table 6.

6.3 Feature extraction analysis

In this subsection, we discuss the performance evaluation of GPU-based feature extraction in comparison with the CPU-based methods. Then, we also analyze the importance roles of features used for sentiment analysis. The best accuracy is obtained when all features are used together. The runtime of the CPU-based algorithms is measured using an Intel Core i7-870 2.93 GHz CPU with 8 GB of memory. Our GPU algorithms are tested using CUDA Toolkit 6.0 running on the NVIDIA Tesla C2050 GPU with 3 GB global memory and 48 KB shared memory per stream multiprocessor.

For the visual feature extraction, our method is 42 times faster than CPU-based ASM algorithm when we run our experiments on 32 images concurrently (Fig. 7). By using the sketch image generation technique, the convergence time of the ASM algorithm is reduced by approximately 50%.

For audio feature extraction task, the overall speedup of GPU-based acceleration is around 12 times in comparison with the CPU-based method, as can be shown in Fig. 8. Among extracted acoustic features, Mel Frequency Cepstral Coefficients (MFCCs) and spectral centroid (SPC) play a lower important roles on the overall accuracy of our sentiment analysis method. However, if we exclude those features from the feature extraction, the accuracy in the audio based sentiment analysis task will decrease accordingly. We also perform experiments on other audio features such as *root mean square*, *time domain zero crossing*, *compactness*. However, we cannot achieve a higher accuracy using these features.

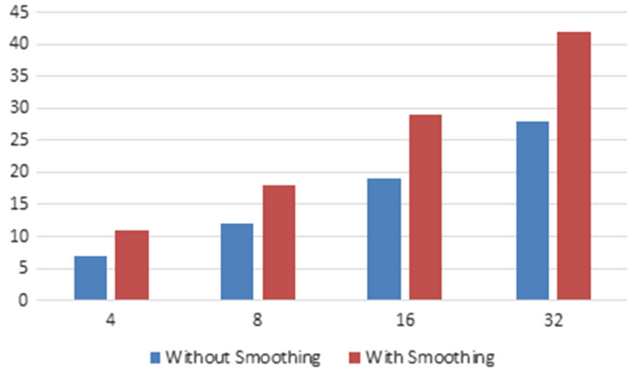

Fig. 7 Facial feature extraction speedup

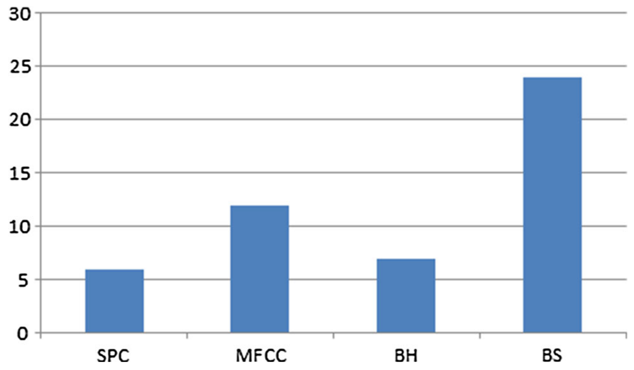

Fig. 8 Acoustic feature extraction speedup

In the case of text based sentiment analysis, we find that concept-gram features play a major role compared to the SenticNet based feature. In particular, SenticNet based features mainly help to detect associated sentiment in a text using an unsupervised method. We aim to develop a multimodal sentiment analysis system where sentiment from the text will be extracted in an unsupervised way using SenticNet as a knowledge base.

6.4 Performance comparison of different classifiers

In this section, we discuss the performance comparison of different classifiers in terms of both accuracy and training time.

6.4.1 Accuracy

On the same training and test sets, we perform the classification experiments using support vector machine (SVM),

Table 7 Comparison of classifiers

Classifiers	Recall (%)	Training time
SVM	77.03	2.7 min
ELM	77.10	25 s
ANN	57.81	2.9 min

artificial neural networks (ANN), and extreme machine learning (ELM). The results in Table 7 show that ELM outperforms ANN by 25% in in term of recall. However, there is only a slight difference in accuracy obtained by ELM and SVM.

6.4.2 Training time

In term of training time, ELM outperformed SVM and ANN by a large margin. As our goal is to develop a real-time multimodal sentiment analysis engine, we prefer ELM as a classifier because it helps to provide the best performance in terms of both accuracy and training time.

7 Conclusion

In this paper, we have developed an ensemble application of ELM and GPU for real-time multimodal sentiment analysis that leverages on the power of sentic memes (basic inputs of sentiments that can generate most human emotions). This work includes sets of relevant features for text and audio–visual data, as well as a simple technique for fusing the features extracted from different modalities. Our method employs various GPU-friendly techniques to enhance the performance of the feature extraction process from different modalities. In addition, powerful ELM classifiers are applied to build the sentiment analysis model based on the extracted features. In particular, our textual sentiment analysis module has been enriched by sentic-computing-based features, which have offered significant improvement in the performance of our textual sentiment analysis system. Visual features also play key role to outperform the state of the art.

As discussed in the literature, gaze and smile based facial expression features are usually found to be very useful for sentiment classification task. Our future research aims to incorporate gaze and smile features, in facial expressions based sentiment classification. We will also focus on the use of audio modality for the multimodal sentiment analysis task. Furthermore, we will make an effort to include a culture and language independent multimodal sentiment classification framework. We will also try to employ other unsupervised, semi-supervised learning algorithms for multimodal sentiment classification.

References

1. Alm CO, Roth D, Sproat R (2005) 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, pp 579–586
2. Busso C, Narayanan SS (2007) Interrelation between speech and facial gestures in emotional utterances: a single subject study. IEEE Trans Audio Speech Lang Process 15(8):2331–2347
3. Cambria E (2016) Affective computing and sentiment analysis. IEEE Intell Syst 31(2):102–107
4. Cambria E, Hussain A (2015) Sentic computing: a common-sense-based framework for concept-level sentiment analysis. Springer, Cham
5. Cambria E, Poria S, Bajpai R, Schuller B (2016) SenticNet 4: a semantic resource for sentiment analysis based on conceptual primitives. In: the 26th international conference on computational linguistics (COLING), pp 2666–2677
6. Chaumartin F-R (2007) Upar7: a knowledge-based system for headline sentiment tagging. In: Proceedings of the 4th international workshop on semantic evaluations. Association for Computational Linguistics, pp 422–425
7. Cootes TF, Taylor CJ, Cooper DH, Graham J (1995) Active shape models-their training and application. Comput Vis Image Underst 61(1):38–59
8. Ekman P, Keltner D (1970) Universal facial expressions of emotion. Calif Ment Health Res Dig 8(4):151–158
9. El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit 44(3):572–587
10. Huang G-B, Cambria E, Toh K-A, Widrow B, Xu Z (2015) New trends of learning in computational intelligence. IEEE Comput Intell Mag 10(2):16–17
11. Johnstone T (1996) Emotional speech elicited using computer games. In: Proceedings, 4th international conference on spoken language, 1996. ICSLP 96. vol 3. IEEE, pp 1985–1988
12. Li J, Lu Y, Pu B, Xie Y, Qin J, Pang W-M, Heng P-A (2009) Accelerating active shape model using GPU for facial extraction in video. In: IEEE international conference on intelligent computing and intelligent systems, 2009. ICIS 2009, vol 4. IEEE, pp 522–526
13. Li X, Mao W, Jiang W, Yao Y (2016) Extreme learning machine via free sparse transfer representation optimization. Memet Comput 8(2):85–95
14. Majumder N, Poria S, Gelbukh A, Cambria E (2017) Deep learning based document modeling for personality detection from text. IEEE Intell Syst 32(2):42–49
15. Michálek J, Vaněk J (2014) An open-source GPU-accelerated feature extraction tool. In: 2014 12th international conference on signal processing (ICSP). IEEE, pp 450–454
16. Morency L-P, Mihalcea R, Doshi P (2011) Towards multimodal sentiment analysis: harvesting opinions from the web. In: Proceedings of the 13th international conference on multimodal interfaces. ACM, pp 169–176
17. Murray IR, Arnott JL (1993) Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. J Acoust Soc Am 93(2):1097–1108
18. Oneto L, Bisio F, Cambria E, Anguita D (2016) Statistical learning theory and ELM for big social data analysis. IEEE Comput Intell Mag 11(3):45–55
19. Poria S, Cambria E, Gelbukh A (2016) Aspect extraction for opinion mining with a deep convolutional neural network. Knowl-Based Syst 108:42–49

20. Poria S, Chaturvedi I, Cambria E, Hussain A (2016) Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: ICDM. Barcelona, pp 439–448
21. Poria S, Cambria E, Bajpai R, Hussain A (2017) A review of affective computing: from unimodal analysis to multimodal fusion. Inf Fus 37:98–125
22. Sattar A, Seguier R (2010) Hmoam: hybrid multi-objective genetic optimization for facial analysis by appearance model. Memet Comput 2(1):25–46
23. Song M, You M, Li N, Chen C (2008) A robust multimodal approach for emotion recognition. Neurocomputing 71(10):1913–1920
24. Tran H-N, Cambria E, Hussain A (2016) Towards GPU-based common-sense reasoning: using fast subgraph matching. Cognit Comput 8(6):1074–1086
25. Turney PD (2002) 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, pp 417–424
26. Várkonyi-Kóczy AR (2010) New advances in digital image processing. Memet Comput 2(4):283–304
27. Wiebe J, Riloff E (2005) Creating subjective and objective sentence classifiers from unannotated texts. In: International conference on intelligent text processing and computational linguistics. Springer, pp 486–497
28. Wilson T, Wiebe J, Hoffmann P (2005) 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, pp 347–354
29. Xiao C, Dong Z, Xu Y, Meng K, Zhou X, Zhang X (2016) Rational and self-adaptive evolutionary extreme learning machine for electricity price forecast. Memet Comput 8(3):223–233
30. Yang C, Lin KH-Y, Chen H-H (2007) 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, pp 133–136