

Multimodal Sentiment Analysis: Sentiment Analysis Using Audiovisual Format

Sumit K Yadav

Dept. of Computer Science
IGDTUW, Kashmere Gate, Delhi
Email ID: sumitarya007@gmail.com

Mayank Bhushan

Dept. of Computer Science
G. L. Bajaj ITM, Greater Noida
Email
ID: mayankbhushan2006@gmail.com

Swati Gupta

Dept. of Information Tech.
IGDTUW, Kashmere Gate, Delhi
Email
ID: gupta.swati1992@gmail.com

Abstract: *Multimodal sentiment analysis is the analysis of emotions, attitude, and opinion from audiovisual format. A company can improve the quality of its product and services by analyzing the reviews about the product [5]. Sentiment analysis is widely used in managing customer relations. There are many textual reviews from which we cannot extract emotions by traditional sentiment analysis techniques. Some sentences in the textual reviews may derive deep emotions but do not contain any keyword to detect those emotions, so we used audiovisual reviews in order to detect emotions from the facial expressions of the customer.*

In this paper we take audiovisual input and extract emotions from video and audio in parallel from audiovisual input, finally classify the overall review as positive, negative or neutral based on the emotions detected.

Keywords: *Analyzing reviews, facial expression, audiovisual, audio features.*

I. INTRODUCTION

The traditional Sentiment analysis can be done in following ways [7]:

A. Keyword Spotting

This approach classifies the text on basis of presence of keywords like Happy, Sad, Afraid, and Bored [2]. This is a very naïve approach having drawback in following areas:

- a) It can't reliably classify negated sentences.
E.g.:- "It was a **happy** moment" and "It wasn't a **happy** moment" both the sentences on the basis of keyword HAPPY will be classified as POSITIVE".
- b) Sometimes meaning of a sentence conveys the emotion rather than any keyword in it.
E.g.:- "My husband decided to file for divorce and he wants to take custody of my children away from me".
[7]

B. Concept Based Approach

Sentiment analysis is done on the basis of web ontology [6]. The system grasp the conceptual and affective information associated with natural language opinions [2]. In this approach,

keywords are not used blindly but it relies on implicit meanings associated with natural language [4]. This approach heavily relies on knowledge base it uses.

Textual reviews may involve ambiguous words, for e.g.: - *bomb* [7], which may lead to incorrect polarities assigned to reviews. In such cases we can have videos which contain reviews about products and on the basis of those reviews we can assign the polarity to the product. The audiovisual format provides an opportunity to mine opinions and sentiment [7]. The smiles, gazes, pauses, and voice pitch are identified as relevant features. Many new areas such as facial expression, voice intensity, pauses, pitch etc. are used in opinion mining from audiovisual formats.

Both audio and video signals are analyzed in parallel with the same frame rate i.e., 30 Hz with the sliding window of 50 milliseconds. Emotions extracted from video like happiness, sadness, boredom involve geometry based approach. The emotions extracted from audio depend on the features like pitch intensity, pauses, loudness etc. End of the statement is detected by detecting any pause in audio frames. The polarity is assigned to that statement spoken by the user on the basis of both audio and video features. The overall polarity assigned to product is the polarity assigned to maximum number of statements in the video i.e., positive, negative or neutral. By using this multimodal sentiment analysis approach we will design an intelligent opinion mining system for identifying, understanding and feeling emotions.

II. RELATED WORK

Various features such as the location of the eyes, eyebrows and mouth are used for the analysis of facial expressions. There are three types of expressions [1] namely:

A. Micro expressions

These are the expressions with the duration of 1/50 to 1/25 seconds. These expressions are generated when a person feels less confidence.

B. Macro expressions

These are the expressions with the duration of 0.5 to 4 seconds. These expressions are generated when a person is confident.




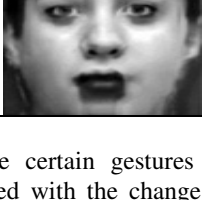
C. Subtle expressions

These expressions are generated when the intensity of emotion is not strong.

In order to detect facial expressions the distance between several points like ends of eyebrows, tip of nose, ends of lips, and tip of cheeks is calculated. The change in facial expressions is detected by detecting the changes in the distance between these points. Voice signals are used in order to increase the accuracy of emotion recognition called multimodal emotion recognition.

The expressions can be classified as [1]:

TABLE I: DIFFERENT EXPRESSIONS [1], [3]

Emotions	View	Description
Happiness		Showing interest, engagement with activity [3].
Sadness		No commitment to the activity, boredom.
Confusion		Difficulty in understanding things.
Neutral		No effects in facial expressions.

Some people have certain gestures on their faces so this technique is worked with the change expressions rather than static measurements. Postures with facial expressions can also be combined in order to improve accuracy.

Inference process of emotion detection consists of following steps [1]:

1. Extracting feature points.
2. Calculating distance.
3. Change indicators.
4. Estimate probabilities(p_a) of defined emotion based changes detected.

A threshold value is determined in order to classify the changes in distance as positive, negative or neutral. With reference to the baseline input the distance is measured and classification is done as follows:

A. Positive

If there is much change in the distance between feature points.

B. Negative

If there is less change in the distance between feature points.

C. Neutral

If there is very large change in the distance between feature points.

The audio features are extracted at the same frame rate i.e., 30 Hz with the sliding window of 50 milliseconds. The audio feature like pitch, pause, intensity of voice and loudness are used to classify emotions [8].

A. Pause

It is the number of audio frames counted as silent. A threshold value is used to identify the samples with or without speech.

B. Pitch

It measures the variation in voice during entire video.

C. Intensity

It is the measure of sound power of spoken words.

D. Loudness

Strength of voice is measured as loudness.

Classification of emotions is done by using both audio and video features in combination. Positive videos are characterized by increased number of smiles and an increased number of pauses, whereas Negative videos are characterized by higher voice intensity and sadness over face [8].

III. PROPOSED WORK

We studied several statements that cannot be analyzed using traditional sentimental analysis techniques either due to ambiguity or multiple meaning of words [7]. Some statements convey deep emotions but don't contain words to express them. The list of studied statements is:

1. She thought she ditched me but I am free now.
2. I met my best friend by an accident.
3. I avoided an accident.
4. I was sad because I had no shoes until I met a man who had no feet.
5. I didn't break up, she broke up, and I just kissed someone else.
6. My husband decided to file for divorce and he wants to take custody of our children from me.

In order to overcome these ambiguities in sentiment classification we decided to switch to multimodal sentiment analysis. Multimodal sentiment analysis involve more than one signal i.e., voice and video signals in order to enhance the accuracy in classifying emotions.

Trusted videos can be collected from the internet so that we can analyze them to finally assign the polarity to the product. We can also apply some filtering techniques to collect trusted and genuine videos over the big data. This filter should be distributive in nature i.e., it should search and filter the videos over several machines in parallel so that it can efficiently deal with such a huge dataset.

We will follow the given steps in order to perform multimodal sentiment analysis:

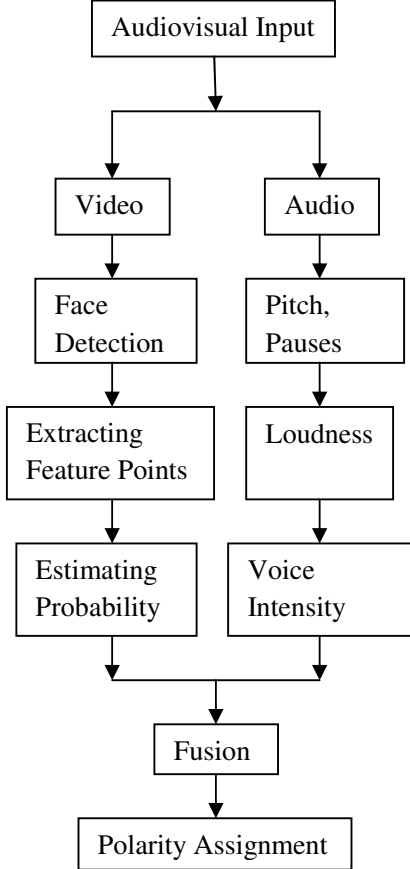


Fig. 1: Processing flowchart of multimodal sentiment analysis.

An audiovisual review is given as input to the system and the output from the system is the polarity assigned the audiovisual review. Both audio and video signals are analyzed in parallel with the same frame rate i.e., 30 Hz with the sliding window of 50 milliseconds [8]. The analysis of video input involve following steps:

A. Face Detection

There are many softwares like Face++ (<http://www.faceplusplus.com/>) [10], FaceRect (<https://www.mashape.com/apicloud/facerec>) [11] in order to automatically detect face. The normalization of size of face is done.

B. Extracting Feature Points

FaceSDK (<https://www.luxand.com/facesdk/>) [12] tool detects 66 facial feature points from the face image. But we used only 7-8 facial feature points for analyzing expressions.

C. Calculating distance

The distance between various feature points is calculated in the baseline input image (d_i). Now the distance between the points in the video image (d_v) is calculated and compared with the baseline image's distance. A threshold value (μ) is used to classify this change in distance [1]. Based on this difference between distances w.r.t. to threshold, the change indicator (c_i) is calculated as follows [1]:

$$c_i = \begin{cases} 1 & \text{if } d_i - d_v < \mu \\ -1 & \text{if } d_i - d_v > \mu \\ 0 & \text{if } d_i - d_v \text{ is very large} \end{cases} \quad \dots \text{Eqn. No. 1}$$

This classification can be defined as:

TABLE II. : CLASSIFICATION OF EMOTIONS

Class	Description
Positive	If there is much change in the distance between feature points
Negative	If there is less change in the distance between feature points
Neutral	If there is very large change in the distance between feature points

D. Estimating Probabilities

The probability (p_a) of classifying emotions into various classes like positive (happiness), negative (sadness) and neutral (no change in emotions) depends on the value of (c_i) i.e., change estimated in feature points.

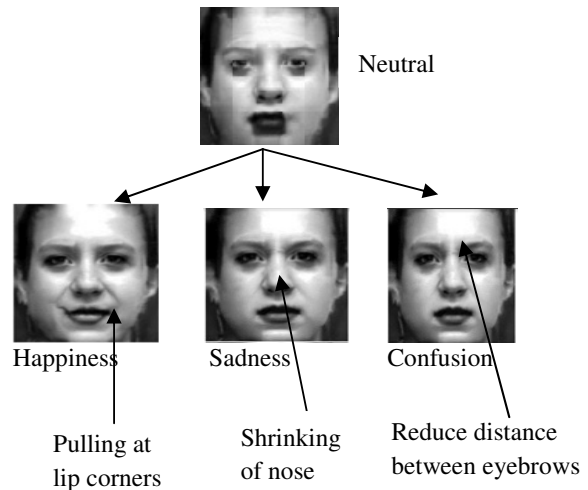


Fig. 2: Change in distance between feature points w.r.t. to baseline image.

Analysis of audio input involves the analysis of four features that are pitch, pause, voice intensity and loudness. These four features are analyzed as follows:

1. The pitch and voice intensity are automatically computed by using open source software PRAAT [9].
2. Pause is the number of audio frames counted as silent. A threshold value is used to identify the samples with or without speech.
3. Loudness in voice is also measured using PARAAT [9]. Loudness in voice is controlled by the amplitude of sound waves.

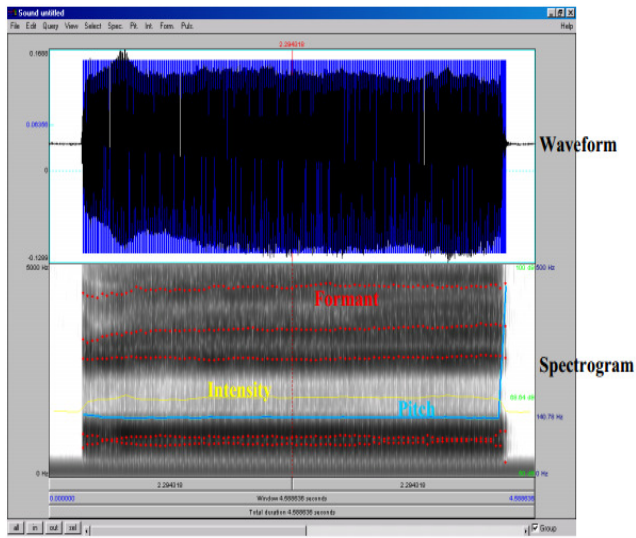


Fig. 3: Snapshot of PRAAT.

Various features measured with PRAAT:

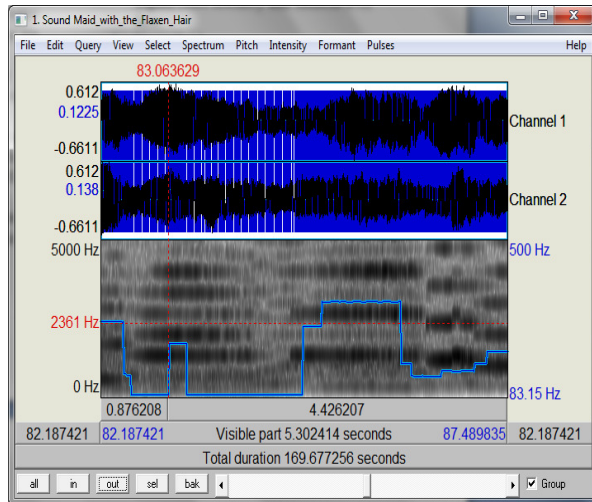


Fig. 4: PRAAT showing maximum pitch in the audio.

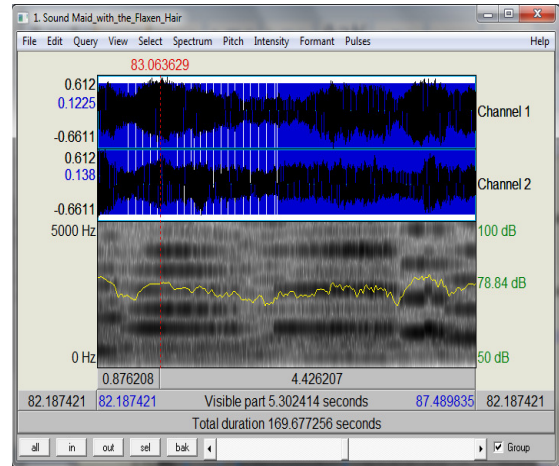


Fig. 5: PRAAT showing maximum intensity in audio

IV. FUSION OF BOTH AUDIO AND VIDEO FEATURES

A statement is detected by detecting a pause in the audio frames; the polarity to statement is assigned on the basis of both extracted audio and video features.

1. Positive videos are characterized by increased number of smiles and an increased number of pauses.
2. Negative videos are characterized by higher voice intensity and sadness over face.
3. Neutral videos are monotonous in nature and involve no change in audio and video features.

The overall polarity assigned to product is the polarity assigned to maximum number of statements in the audiovisual review i.e., positive, negative or neutral. A count is performed in order to detect polarity occurred maximum number of times.

V. CONCLUSION

By using this multimodal sentiment analysis approach we will design an intelligent opinion mining system for identifying, understanding and feeling emotions. This system involves more than one signal in order to analyze emotions and hence overcome the drawbacks of traditional sentiment analysis system. The ambiguities present in text based reviews leads to confusion. Those confusions are removed by the analyzing person's feeling while delivering that statement. This system involve more than one signal i.e., voice and video signals in order to enhance the accuracy in classifying emotions.

VI. FUTURE WORK

Future work involves exploration of our proposed approach for larger videos, including more feature points in order to identify other known emotions. Automatic identification of several other emotions is challenging. Fusion process can also be improved by compiling several new facial emotions with voice features. This technology can also be used in video summarization in future. Sometime user doesn't have time to

watch entire video, video summarization generates highlights of few minutes from the entire video of hours or more. The highlights should convey as much as valuable information they can. In this way user can save their time by just watching highlights rather than whole video.

REFERENCES

- [1] P. Lago and C. J. Guarín, “An Affective Inference Model based on Facial Expression Analysis”, IEEE Latin America Trans, Vol. 12, No. 3, May 2014.
- [2] Ms.K.Mouthami, Ms.K.Nirmala Devi, Dr.V.Murali Bhaskaran, “Sentiment Analysis and Classification Based On Textual Reviews”, International Conference on Information Communication and Embaded System (ICICES 2013), Page(s): 271 – 276.
- [3] Songfan Yang and Bir Bhanu, “Understanding Discrete Facial Expressions in Video Using an Emotion Avatar Image”, IEEE Trans. Systems, Man, and Cybernetics—Part B: Cybernetics, Vol. 42, No. 4, August 2012.
- [4] Lizhen Liu, Xinhui Nie and Hanshi Wang, “Toward a Fuzzy Domain Sentiment Ontology Tree for Sentiment Analysis”, 5th International Congress on Image and Signal Processing (CISP 2012) IEEE, Page(s): 1620 – 1624.
- [5] Lei Zhang and Bing Liu. “Extracting Resource Terms for Sentiment Analysis”, Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP-2011), pp.1171-1179, November 8-13, 2011.
- [6] B. Lu et al., “Multi-Aspect Sentiment Analysis with Topic Models,” Proc. Sentiment Elicitation from Natural Text for Information Retrieval and Extraction, IEEE CS, 2011, pp. 81–88.
- [7] Cambria, E.; Schuller, B. ; Yunqing Xia ; Havasi, C. “New Avenues in Opinion Mining and Sentiment Analysis”, Intelligent System, IEEE, Vol. 28, 2013, Page(s): 15-21.
- [8] Rosas, V.P. ; Mihalcea, R. ; Morency, L. “Multimodal Sentiment Analysis of Spanish Online Videos”, Intelligent System, IEEE, Vol. 28, 2013, Page(s): 38-45.
- [9] <http://www.fon.hum.uva.nl/praat/>, Last accessed on Date: 05-11-14
- [10] <http://www.faceplusplus.com/>, Last accessed on Date: 15-10-14
- [11] <https://www.mashape.com/apicloud/facerec>, Last accessed on Date: 12-10-14
- [12] <https://www.luxand.com/facesdk/>, Date: 01-10-14