Survey of various approaches of emotion detection via multimodal approach

Prof. Deepa Abin[1], Ms. Samiksha Saini[2], Mr. Rohan Rao[3], Mr. Vinit Vaichole[4], Mr. Anand Rane[5]

[1]Professor, Dept. of Computer Engineering, PCCOE, maharashtra,India
[2,3,4,5] Student, Dept. of Computer Engineering, PCCOE, maharashtra,India

---

Abstract - *Emotion detection of users is a challenging and exciting field where user's data is analysed to recognise emotions such as happy, sad, angry etc. This data could be in one or multiple formats such as audio, video ,text ,still images et al. Relevant features are extracted and fused together to give a label. Fusing data from two or more sources(modalities) is another challenge, feature level or decision level fusion is employed. This paper inspects and studies the various approaches to multimodal extraction of emotions.*

1.INTRODUCTION

In today's day and age smart intelligent systems relying on machine learning have become ubiquitous. Everything from search engines, recommendation system on online shopping websites to personal assistants, fit bits and of course smart phones, various day to day objects are being empowered by a machine learning algorithm that lends it the ability to make decisions based on past experiences and decisions to accomplish its goal. All this though doesn't consider the mood of the user into account. Creating computer applications that are empathetic to us ,i.e understand ,analyses and respond to human emotions definitely would be better at providing solutions. Therefore in this paper we look into detecting the emotion of a user using their tweets and facial expressions and combine it with the answers of a beck decision inventory questionnaire. Scientists at UC Berkeley have identified 27 distinct human emotions .These emotions are associated with a wealth of information about the human mind. Equipping computers to recognize emotions will have benefits in various fields. As aides to psychologists in diagnosing depression ,in designing better products that connect well with customers' needs, to develop smart tutoring systems[2] that teach in relevance to a student's learning ability, an autonomous car system which can recognize tired river and switch to autopilot , a personal assistant that can understand tone of user etc. Emotion detection and recognition market is projected to be worth $22.65 Billion by 2020, according to the market research firm Markets and Markets.

A multimodal approach that is any combination of text, audio, visual, body posture, hand gestures, facial expressions etc would give a comprehensive insight about the user's mind. A single modality may give only one sided info or may miss out on an inherent parameter. A lot of literature exists that combines audio and video data [3][6][7].In this paper we propose a system combining text(tweets) and video features[9] to predict user's emotion as this combination gave better result than any other[1].Since twitter is the most popular micro blogging site which is regularly and frequently used to express sentiments it was our ideal choice as source for text, video data is taken in real-time as user answers the beck 9 questionnaire.

The next point of focus is extracting suitable features and combining them to predict the emotion label. There are two popular approaches to combining feature set which have been compared 1) decision level 2) feature level[3][5][10].

The remainder of the paper describes the following sections 2. Related Literature 3. Text based analysis 4. Video based Analysis 5. Proposed System 6.Conclusion

2. LITERATURE SURVEY

There is vast literature that explores a multi-modal approach for detecting the emotion of a user.They all share the common theme of acknowledging the difficulty in understanding the subjectivity of human emotions and accuracy of of understanding the context of interaction at the time in which the emotion was expressed. All papers seem to address the common shortcoming of limited dataset , and that is why multimodal approach is the best as they outperform systems where only a single modality is chosen.Since a single modality may miss out on an inherent parameter for depression detection or may not give full information.

In a paper[1] authored by Rahul Gupta, Nikolaos Malandrakis, and Bo Xiao.proposed system combined features extracted from text,video and audio to predict depression. All features were linearly combined and classified using a support vector regressor (SVR).A multi stage feature selection process was adopted, a brute force strategy is applied to extract a subset of feature groups. Then a best-rst forward search strategy on the combination of features obtained.

This paper also outlined the results of various combinations of features, video, audio and text.

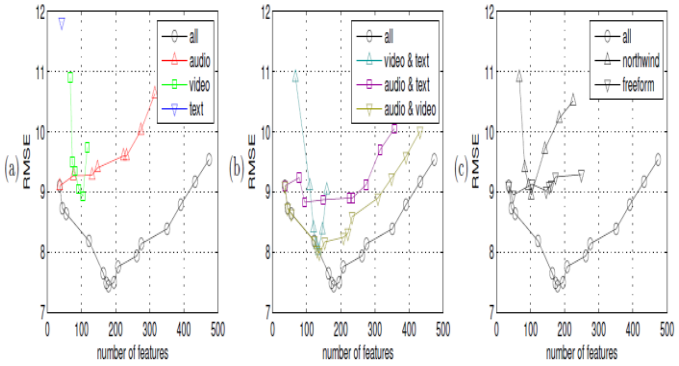

Fig 1: Comparison of audio,video and text

As evident from the experiments a combination of video and text gave the best result with feature level fusion.

Ramon , Mara, Gilberto propose a tutoring system[2] that extracted the current emotion of student using video and text from chatbox. The results from both these modalities along with other parameters like time to solve question and error rate were combined with fuzzy logic to determine next level

Another work by Carlos, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann*, Shrikanth Narayanan an audio video bimodal system classified data into 4 categories of anger, happiness, sadness and anger. A comparison between feature level fusion and decision level feature showed both had almost similar accuracy of 89%.And both superseded in perform of single individual modality trained classifiers.

Table -1 : Confusion matrix for feature level integration of bimodal system

	Anger	Sadness	Happiness	Neutral
Anger	0.95	0.00	0.03	0.03
Sadness	0.00	0.79	0.03	0.18
Happiness	0.02	0.00	0.91	0.08
Neutral	0.01	0.05	0.02	0.92

Table -2: Confusion matrix for decision level integration of bimodal system

	Anger	Sadness	Happiness	Neutral
Anger	0.84	0.08	0.00	0.08
Sadness	0.00	0.90	0.00	0.10
Happiness	0.00	0.00	0.98	0.02
Neutral	0.00	0.02	0.14	0.84

As evident from the confusion matrix a feature based bimodal system performed slightly better than decision based system.

The authors of another paper[6] Simina Emerich1, Eugen Lupu1, Anca Apatean1 first identified the various suitable feature to classify data into 6 emotions namely, sadness, happiness, anger, disgust, fearand neutral. It used a SVM classifier employed with a RBF- kernel was used as it gave better results than other classifiers such as k-means and naive bayes.It followed two approaches for feature integration one was feature level fusion wherein a single feature vector was formed and normalized using z-score transformation and a match score method. The feature level fusion gave better results in correctly identifying emotions with 93% which was 1% more than match score.

Emotion Recognition System / Technique		Speech Information	Facial Expressions	Feature Level Fusion
Classifier	10-fold cross validation			
	SVM (RBF)	87.7%	90.3%	93%
	SVM (POLY)	85.2%	88.8%	90.2%
	Naïve Bayes	67%	68.14	68.7%
	K-NN (k=3)	73.7%	84.4%	86.6%
	80% training 20% testing			
	SVM (RBF)	83.3%	86.7%	91.1%
	SVM (POLY)	83.3%	85.2%	88.8%
	Naïve Bayes	66.6%	70.3%	70.7%
	K-NN (k=3)	72.2%	83.3%	85.2%

Liyanage C. De Silva, Pei Chi Ng, of Singapore used statistical techniques and Hidden Markov Models (HMM)[8] in the recognition of emotions. The method aims to classify 6 basic emotions(angry, dislike, fear, happy, sad and surprise) from both facial expressions (video) and emotional speech (audio). A bimodal system with accuracy of 72% was made using a rule based classifier, however about 10% of data could not be sampled as the data extracted was inconclusive.

3 Text based Analysis

In the paper by Ramon , Mara, Gilberto that proposes a tutor system, an ASEM algorithm was utilised to recognise emotions from text input by student ..It showed a success rate of 80%. A student input a text line (input dialogue).The text is normalized: the accented words, numbers, and special characters are removed and uppercase letters are converted to lowercase. Non emotion words like he, she, the, etc. are removed by using the corpus StopWords. Semantic words are sought in corpuses Semantic and improper words. If semantic words are not found in the corpuses the new words are added to corpus NewWords. If the semantic word is found in the corpuses, the word features (PFA and emotion) are extracted.The emotion in classified according to the features of the word. The emotion with greatest intensity is produced.

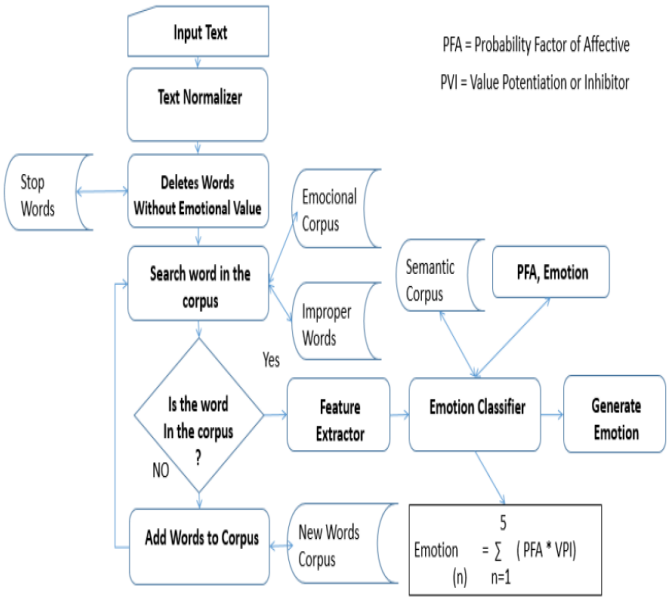

Fig -2 : Text analysis

In another paper [7] the authors extracted emotions from youtube dataset, for text the extracted the transcripts and followed the sentic computing paradigm developed by Cambria and his collaborators, which considers the text as expressing both semantics and sentics [5]. They conducted concept level sentiment analysis, concept extraction from the text is the fundamental step of the experiment. Below, we first describe the concept extraction algorithm [39] from text and then we describe the feature extraction methods based on the extracted concepts for concept level sentiment Analysis.

The text was represented using bag of concepts feature. For each text we extracted concepts using the concept extraction algorithm. Later, the concepts were searched in the EmoSenticSpace and if any concept was found then the corresponding 100 dimensional vector was extracted from the EmoSenticSpace. After that individual concept vectors were aggregated into one document vector through coordinate-wise summation.The polarity scores of each concept extracted from the text were obtained from SenticNet and summed to produce one scalar feature.

4 Video based Analysis

Emotion detection is based on different expressions of face and these expressions are generated by variations in facial features. A video consist of varied facial frames that are beneficial in the matter of detecting emotions since a video is capable of capturing multiple facial frames and we could use appropriate methods to find the outputs. A broad study in emotion detection and analysis shows algorithms and techniques to capture facial images from a video that have been tested and concluded to be more accurate than still images. In a paper, authors [] have used Support Vector Machine to detect emotions from facial images and used PCA to extract the features and reduce the dimensions into 2-dimensional vector space. SVMs are memory efficient and effective in higher dimension spaces. Also OpenCV contains cascade classifiers in which Viola & Jones face detection algorithm is used. By using these classifiers the face region is detected from the image. It classifies the images into positive and negative images respectively. An image with a face is positive and without face is negative. After classifying the training and testing images PCA is applied on training set and classification into emotions Happy, Sad and Neutral is performed.

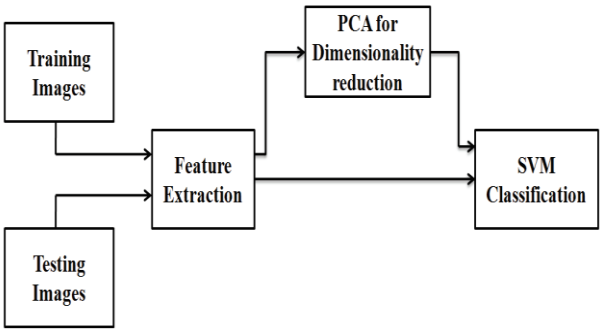

Fig -3 : Classification of images

Generally a emotion detection from video contains three stages - face detection and tracking, facial feature extraction and finally the classification stage. Some of the papers referred stated different methods to process images from the videos. Images are converted into 2-dimensional or 4-dimensional vector space. And commonly face detection algorithms studied were Viola Jones Algorithm and Gabor filters are also applied on eyes and mouth regions to extract relevant features.

5 PROPOSED SYSTEM

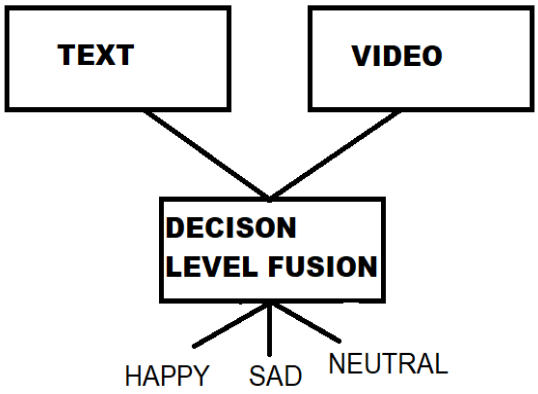

Fig -4 : System Architecture

In this system two modalities text and video are chosen as evident by work in paper [1] this combination gives best accuracy.The source for text is user's twitter account and

source for video is real time input while user answers a beck inventory depression based questionnaire.

5. CONCLUSION

Thus a number of papers detailing the study of extracting features from various modalities , combining features from distinct sources and labelling user's current emotion are are studied. From the papers it can be inferred that combination of two or more modalities gives better result than an individual modality. From the different bimodal systems a combination of video and text gives best results.

6. REFERENCES

[1]Rahul Gupta, Nikolaos Malandrakis, and Bo Xiao. Multimodal prediction of affective dimensions and depression in human-computer interactions.

[2]Ramon Zatarain-Cabada, Mara Lucia Barrn-Estrada, Jorge Garca-Lizrraga, Gilberto Muoz-Sandoval, Java Tutoring System with Facial and Text Emotion Recognition Instituto Tecnolgico de Culiacn, Culiacn Sinaloa, Mexico, Research in Computing Science 106 (2015)

[3]Carlos Busso, Zhigang Deng *, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann*, Shrikanth Narayanan ,Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information

[4]Rajesh K M , Naveenkumar M , A Robust Method for Face Recognition and Face Emotion Detection System using Support Vector Machines Dept. of Telecommunication Siddaganga Institute of Technology (SIT),Tumkur, India, ICEECCOT (2016)

[5]Louis Philippe Morency, Rada Mihalcea, Payal Doshi Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web, ACM 9781450306416/ 11/11

[6]Simina Emerich1, Eugen Lupu1, Anca Apatean1, Emotions recognition by speech and facial expression analysis, 17th European Signal Processing Conference (EUSIPCO 2009)

[7]Soujanya Poria Amir Hussain Erik Cambria Beyond Text based sentiment analysis: Towards multimodal systems.

[8]Liyanage C. De Silva, Pei Chi Ng, Bimodal Emotion Recognition, The National University of Singapore Department of Electrical Engineering, (IEEE 2009)

[9]Lexiao Tian, Dequan Zheng, Conghui Zhu, Image Classification Based on the Combination of Text Features and Visual Features(2013)

[10]Pooya Khorrami, Tom Le Paine, Kevin Brady, Charlie Dagli, Thomas S. Huang1, HOW DEEP NEURAL NETWORKS CAN IMPROVE EMOTION RECOGNITION ON VIDEO DATA, IEEE(2016)