JOURNAL OF CRITICAL REVIEWS
ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

An Intelligent Framework for CNN-Based Multimodal Sentiment Analysis

I Lakshmi Manikyamba
Associate Professor
Department of CSE
JNTUH College of Engineering Hyderabad, Telangana
lakshmi.isit@gmail.com

Abstract

There are several social media analytics activities that rely heavily on sentiment analysis of user-generated textual information from the internet. Adding pictures and videos to one's social media posts has become common practice among today's internet users as a means of conveying one's point of view and experiences. Large-scale text and visual data sentiment analysis helps extract user feelings about brands or subjects. An intelligent multi-modal sentiment analysis framework is urgently needed to effectively mine information from several modalities in order to keep up with the explosion of massive multimodal data. Previous research mostly dealt on either textual or visual material. By using a "convolutional neural network" ("CNN") to extract features from several modalities, this study aims to provide a novel framework for multi-modal sentiment analysis. To showcase the sophistication of our models, we conduct "multi-model sentiment analysis" on two publicly accessible “datasets,” “including text,” “audio,” and “visual input.”

Keywords: “Sentiment analysis,” “multimodal data,” “CNN,” “SVM”

I. Introduction

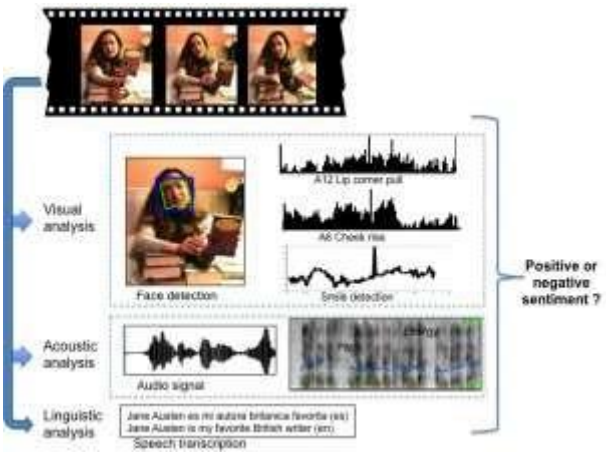
An extremely significant part is played by perspectives and points of view in decision making. Today's internet users have access to a “variety of platforms,” “including social networking sites,” “blogs,” “forums,” and “e-commerce websites,” to voice their opinions and thoughts. As a result, individuals are shifting their focus away from a single text mode and toward the modality of huge social media data as a universal platform. More and more users of micro blogging platforms, such as Twitter, are opting to post multimodal tweets in

which text is accompanied by an image, posing new issues for social media analytics tools that must deal with enormous amounts of social media data at scale. When compared to typical text-based sentiment analysis, the accuracy of an analysis that incorporates user sentiments from two or three input modalities is significantly improved. The primary benefit of video analysis is the multi-modal facts it delivers, including both verbal and visual information. In addition to the textual data, the visual data, which includes things like voice inflections and facial expressions, may reveal people's genuine thoughts and feelings on any given issue. Analyzing emotions through several channels is the goal of “multi-modal sentiment analysis.” Figure 1 demonstrates the contribution that each modality makes in determining emotional state.

There will be a lot of new studies dedicated to something called "multi-modal sentiment analysis," and methods built on deep neural networks have shown to be more accurate. Integration of disparate data types, including text, images, video, and audio, is a significant difficulty in "multi-modal sentiment analysis" projects. The use of deep neural networks, however, may allow for the creation of a novel multimode sentiment analysis model. Multi-modal sentiment analysis is an emerging field of study due to the enormous data quantities generated by social media.

As a result, the majority of the first studies in this field only investigated one kind of sentiment analysis technique: text. To compete with traditional machine learning-based methods, deep neural networks have become more popular in recent years, especially when it comes to extracting textual representation.

In the meanwhile, CNNs have also found use in picture sentiment analysis on account of their impressive track record in the field of image classification. In an effort to remedy these problems, this research presents a comprehensive framework for doing large-scale “multi-modal sentiment analysis” using convolutional neural networks.



"Figure I: Extraction of multimodal features"

II. Applications

Many researchers have found success using sentiment analysis to learn about the dynamics of a situation. A limited number of contexts benefit from its use, including

1. Businesses and organizations

The reaction of consumers is a major factor in shaping corporate strategy. Firms' strategic movements are motivated by public opinion and perspectives since companies aim to serve the requirements and desires of users. Because of how closely the globe is linked together by technology, events may have far-reaching consequences; a problem or setback in one section of the country might have repercussions in another. Since this is the case, it is therefore imperative that goods and services be steered in accordance with the general public's perspective. Successful businesses nowadays spend substantial resources exploring customer opinion.

2. Analyzing and settling on certain items

Sentiment analysis has helped to simplify the process of comparing and selecting among various items. A product's feature specs may be evaluated using this method. Likewise, it is now much less of a hassle to evaluate competing items. Making choices is an important element of being human. It ranges from deciding what to purchase and where to eat to picking out a bank's insurance policy and making investments. In the past, users would make decisions and choose among available alternatives based on the consensus of the community.

3. Ads placements

Show an ad when the user likes or evaluates the product, and show an ad from a participant when the user doesn't. Analyzing reviews written by customers allows us to determine if they are favourable or negative.

4. Methods for making suggestions

The website you're now using probably includes a built-in recommendation system to help you out, whether you're searching for books, online media, entertainment, music, the film business, or any other type of art. Systems like this learn from our habits, preferences, and social networks to generate educated recommendations.

5. Creating new and useful items

Sentiment analysis improves the evaluation of a product's usability and human-friendliness when it's subject to harsh competition and available to criticism through public evaluations

and comments. It fosters an atmosphere where superior and novel items may flourish. Using the quantifiable relationship between good and negative tweets, they may learn how satisfied customers are with the product.

6. Estimating the level of client contentment

They may calculate the ratio of happy to angry tweets regarding the product.

7. Locating Opponents and Supporters

It helps with customer service by picking up on complaints or issues with the merchandise. People who are already satisfied with our brand or services may be recruited to assist spread the word about what we provide.

III. Inferencing with "Convolutional Neural Networks (CNN)"

The "home" of the artificial neural network known as a "Convolutional Neural Network" (CNN), which is often used in computer vision, is in deep, feed-forward machine learning. NLP now use this technique to categorize texts. Multiple CNN designs exist. Modular convolutional and pooling layers make up the majority of the architecture. We use text and image-based sentiment analysis to train this network as a machine learning algorithm and then extract features from it. We combined image recognition and NLP to implement the cutting-edge technique of "multi-modal sentiment analysis" using deep neural networks such as CNN. Good validation accuracy with great consistency was attained by the suggested method using CNN.

IV. Related work

An increasing amount of work is spent using sentiment analysis to the textual data on Twitter during the last decade. Examples include the ground-breaking work done by "Pak" and "Paroubek" [1], who showed that emoticons could be used to collect a tagged dataset for sentiment analysis. The amount of abstraction needed to understand the meaning sent by a picture makes visual sentiment analysis quite different from text analysis [2]. SentiBank, developed by the first system to extract intermediate-level language[4] features from photographs for the purpose of sentiment analysis. Using these linguistic traits, classifiers may make educated guesses about which of the feelings represented by wheel of emotions a picture most likely evokes [5]. To classify tweets and photos into positive or negative categories, "convolutional neural networks,"[6] which were inspired by the rise of deep learning techniques. However, studies on picture annotation [7], [8]" shown that integrating text characteristics into image annotation significantly improved overall performance. Despite the growing body of research on audio-visual fusion for emotion detection, multimodal emotion or sentiment analysis, which additionally considers textual signals in

addition to those given by the visual and auditory modalities, has gained very little attention. Emotion and sentiment were extracted by combining data from many modalities by [9],[10] [11 algorithms for detecting emotions used both audio and textual cues. All of the approaches rely heavily on combining several features. Audio and textual information were fused at the decision level in a study [12] A unified neural network model using CNNs to extract text and picture hints was suggested [13]. Bear in mind that our studies tend to include textual information with the audio and video modalities, while the vast majority of previous work on audio-visual emotion analysis has focused only on integrating these modalities.

In order to do tri-modal sentiment analysis and emotion identification, the current study suggests using a "Convolutional Neural Network" (CNN) based structure to extract features from audio/visual and textual modalities. The level of sophistication of this model is much superior than that of competing approaches. We also investigate how our method performs in terms of speaker independence, model generalizability and modality performance, all of which are topics that are seldom covered by other writers.

V. Method

I. Textual Features

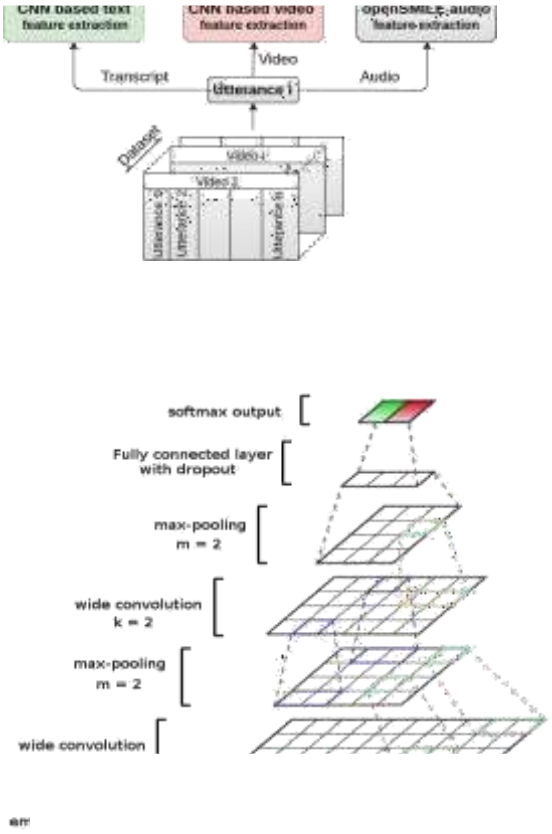
The most popular classification methods benefit from the "bag of words" paradigm since it does not care about the sequence of the words being classified. Support vector machines ("SVM"), maximum entropy, and naive bays are all examples of these methods. If the input consists of a series of words, this might make emotion mining less effective. Because of this, the underlying emotion may be thrown off. Yet, several reports have demonstrated that this barrier might be broken via the use of deep learning for sentiment analysis. In order to do sentiment analysis, this network design employs a deep neural network that utilizes representations at the word, character, and sentence levels. As a rule, we use convolutional neural networks for text feature extraction (CNN). We used Convolutional Neural Networks (CNNs) as a learnable feature extractor and Support Vector Machines (SVMs) as a classifier, such that the learned CNN variants could be fed into SVM. The idea underlying convolution, shown in Figure2, is to multiply each k-gram in the sentence s(t) by a vector of k weights wk (the kernel vector) to generate a new sequence of alternatives “c(t) = (c1(t)”; “c2(t);::: ; cL(t))”:

$$\text{“}cj = wk^{T}xi{:}i+k\text{”} \text{ --- (1)}$$

The maximum value is often kept after applying maximum pooling to the feature map.

“c^(t)=maxfc(t)g” because there are several reasons why this particular kernel vector and

window sizes are so highly prized. Each word "xi(t)" in the vocabulary is "embedded" as a d-dimensional vector using a look-up table constructed from the input. Vectors representing the words of a phrase are strung together to form the picture.



"Fig 2: Extraction of Text Features Using a CNN"

After that, word vectors, rather than individual words, are subjected to the convolution kernels. Similar look-up tables may be created for non-word properties if they are judged valuable. Using these qualities, we have trained the CNN's deeper layers to accurately represent bigger word groups inside phrases. Fh1 denotes the feature learned by a hidden neuron h in layer l. In order to learn several features at once, we employ the same "CNN" layer. A layer's performance is fine-tuned based on the qualities it has learned from the layer below it.

$$"F^{1} = X^{nh}{}_{h\,=\,1wkh\,F1}" \quad \text{--- (2)}$$

Where (*) stands for a convolution, (wk) is the weight kernel for neuron (h), and (nh) is the

number of neurons that are really hidden. Word order is preserved via the CNN sentence model's use of progressively larger convolution kernels to represent an expanding vocabulary and, ultimately, the whole sentence. Word embedding was used to create a representation for each word in a phrase.

We used the freely available word2vec vectors, which were educated on a corpus of one hundred billion Google News articles. A constant bag-of-words framework was used to train 300-dimensional vectors. Non-pre-trained words were given a random starting point. Every single one of those sentences might only be fifty words long. We used a CNN with two convolution layers. All models employed a second convolution layer with 200 feature maps and a kernel size of 3 or 4. The first convolution layer had 50 feature maps. Interleaved with the convolution layers were two-dimensional pooling layers. In the fusion process, our feature vector is the activation value of the network's 500-dimensional fully-connected layers.

II. Sentiment Analysis from Sound Files

Emotional content and acoustic signal analysis is a well developed discipline. As the MelFrequency Cepstral Coefficients (MFCC) have shown to be useful in the speech recognition field, researchers have begun to investigate their potential applications in the realm of musical modeling. At present, MFCCs are widely employed in audio processing and other MIR applications, such as generic classification. The Open-EAR programme is used to extract audio data including pitch, intensity, and loudness, and the SVM classifier is used to identify emotion. To identify the feeling expressed in a video clip, we first utilize the Open EAR package to automatically extract audio characteristics from the audio track, and then we train a "Hidden Markov Models" ("HMM") classifier. Tools like OpenEAR/OpenSMILE may be used to extract features from an input, however in our research we focus on only the stressed and conventional parts of the input, extracting characteristics like MFCC, prosody, and relative prosody. We can construct a sentiment detection system that analyzes ongoing audio streams to ascertain the speaker's emotional state by using Maximum Entropy modeling and Part of Speech tagging. Automatic Speech Recognition is used to transcribe audio files (Automatic Speech Recognition). This method demonstrates the feasibility of automated sentiment detection in purely spontaneous audio. Emotional context may frequently be gleaned from the speaker's vocal tract, level of excitement, and prosody.

III. Visual study of emotional expression

Basic assessment tasks in "visible sentiment evaluation" include modeling, investigative paintings, and capitalizing on the emotions communicated via body language, facial

expressions, and other visual media. Our new video frames have to take this into account, since the underlying data may easily warrant extremely large dimensions. To drastically cut down on the quantity of video content used for instructional purposes. A CNN architecture was then applied to the frames in order to classify the movie as a succession of still images. To capture this temporal dependence, a multi-level convolutional neural network was fed a concatenated picture of all photographs taken between timestamps t and t+1. Consequently, all video frames had to be scaled down to half resolution as part of the pre-processing. In order to get temporal convolution possibilities, each possible combination of successive video frames was reincarnated into a single frame. By using zeros as padding, we were able to get all the images down to exactly 250 by 500 pixels. Our first convolution layer made use of "100 kernels" of size 1020, while our second employed "100 kernels" of size 2030. Afterwards, a logistic layer with 300 completely connected neurons and a softmax layer were included. Pooling layers of dimension 2 2 had been added to the structure, which contained the convolution layers. The characteristics used to categorize videos are the levels of activity of neurons in the logistic layer.

VI. Fusion

The term "multimodal sentiment analysis" describes an approach whereby many input modalities are used to enhance the analysis's efficacy. Multimodal sentiment analysis takes in both written and visual/auditory data.

The same goal may be reached using a variety of “fusion techniques,” “including data fusion,” “feature fusion,” and “decision fusion.” The process of fusing several decisions into one is ubiquitous. Everything about the proposed system's structure is shown forth in Figure-3. Our next meeting will focus on the fusion that occurred after we combined many features into a single joint feature vector and sent it to a support vector machine (SVM) to reach a final determination.

VII. Studies and Tests

I. Datasets

Every day, people all around the world use video sharing websites to express their thoughts and provide their recommendations. There has been a recent uptick in academic and commercial interest in the study of emotion and subjectivity in these opinion films. As it pertains to analyzing the tone of written material, sentiment analysis is effective. As far as multimedia material is concerned, this is a mostly unexplored research area.

Inadequate datasets, methods, baselines, and statistical assessment of the interplay between

many modalities provide the greatest challenges to education in this line of inquiry.

II. MOSI

Being the first opinion-stage annotated corpus of sentiment and subjectivity evaluation in internet videos, we provide the "Multimodal Opinion-Stage Sentiment Intensity dataset" ("MOSI") in this study. Video critiques of movies, books, and other products total 2199 in this Zadeh et al.-built database. “Subjectivity labels,” “sentiment intensity labels,” “consistent-with-frame” and “consistent-with-opinion annotated visual features,” and “consistent-with-milliseconds annotated audio features” are only some of the labels that have been applied to the dataset.

III. MOUD

In this research, we use the "Perez-Rosas et al. " Multimodal Sentiment Analysis Datasets - MOUD. They compiled one hundred videos from YouTube that offered critiques and suggestions for various products. All the words said in a film were dissected, and then each word was assigned to one of many emotions (positive, negative and neutral). There are six separate statements in each film, and each one goes on for five seconds on average. There are 498 statements in the data collection that have been rated as either excellent, bad, or neutral. To ensure our experiment was consistent with prior studies, we did not include a neutral label. In an attempt to solve the generalizability difficulties, this experiment trains the model on MOSI before testing it on MOUD. For each phrase in the dataset, this approach creates a new feature vector by combining the characteristics collected from all the multimodal streams. This vector is then used to produce an emotion judgment. Table-1 below summarizes a wide range of one-, two-, and three-way comparisons. Using an SVM classifier built in the Weka toolkit, we perform 10 fold cross validations on all 412 statements.

“Table 1: Analysis of variance with respect to speaker”

Modality	MOSI	MOUD
Unimodal		
Text	75.16	48.4
Video		
Audio		
Bimodal		
Text+Audio		
Text+Video		
Video+Audio		
Multimodal		
Text+Audio+Video		

IV. Experimental Design Ignoring the Speaker

Most studies of multimodal sentiment analysis use datasets with overlapping speakers during the training and testing iterations. Everyone, of course, has his or her own special manner of articulating feelings and thoughts. Sentiment analysis characteristics that may be used regardless of the speaker's identity plays a crucial role across all channels. For practical use, the model must be able to withstand slight variations across individuals. For this reason, we conducted objective tests to simulate hypothetical circumstances. This time around, we made sure that the datasets used for training and for testing were kept completely separate out of respect for the speakers. In the course of the checkout process, our models were presented with utterances from speakers they had never seen before, and they had to identify the underlying emotions and attitudes. We employ internet video review data to conduct a speaker-independent experiment below.

V. Evaluation information from online videos

First, we compiled a library of YouTube videos utilizing a wide range of keywords that may be used in a review or suggestion of a product. From the resulting pool of films, we chose the ones that best backed up the following recommendations. The speaker should be in the front of the camera, with no more than a small amount of facial occlusion at any one time; no background music or animation should be present; and the recording should be played back in real time. One hundred films that still adhered to the criteria outlined above were randomly picked to make up the final video package. One hundred voices are included in the sample, with ages ranging from around “20 to 60 years old.” The first thing that was done to the videos was to remove the opening titles and advertising. We deliberately chose a 30-second opinion portion from each video to ensure that just one issue was covered in each review, since reviewers often changed subjects while expressing their views. Since there are so many voices in this dataset (100 in all), we ran a speaker-independent test with ten rounds, utilizing a different voice for each round. All improvements in performance were measured using the same macro F score and the same SVM model. In Table 2, we see the outcomes of the experiment conducted without the involvement of the speakers. It is clear that audio consistently outperforms visual across all datasets. When it comes to identifying emotions and analysing sentiment, the text mode is crucial. Integrating many modalities has a greater effect on emotion recognition than on sentiment analysis.

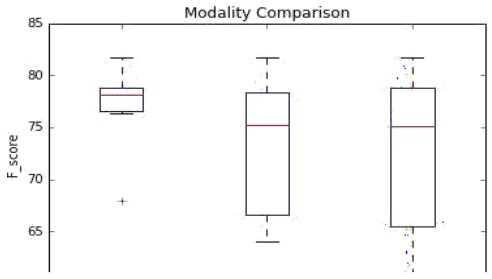
"Table 2: Relative independence of the speakers: a cross-validation result"

"Modality"	"Macro F_Score"
"Baseline"	55.93
"Unimodal"	
"Text"	50.89
"Video"	41.60
"Audio"	45.60
"Bimodal"	
"Text+Audio"	51.70
"Text+video"	52.12
"Video+Audio"	46.35
"Multimodal"	
"Text+Audio+video"	52.44

When compared to unimodal controls, bimodal and trimodal designs have consistently outperformed their single-modal counterparts in trials. When comparing the two modalities, audio consistently outperforms visual across all datasets. Table 2 compares the text modality's unimodal performance to that of the other two modalities and the state of the art.

VI. What Each Modality Contributes

We have performed a qualitative manual examination of the performance of all modalities on the datasets to get a better understanding of the roles played by each modality in the classification process as a whole. The middle F score is shown by the red line.



"Figure 3: Effects of the Methods"

High polarity terms are favored fig 3 in text classifications, with positive polarity being accurately detected and the bi and multi-model characteristics aiding in right categorization. Even when the reviewers' faces weren't easily seen, the textual format was useful. However,

in certain cases, linguistic signals led to a misclassification of the text modality. Positive connotations like "likes to see" or "responsible" are present. However, the person's angry tone and expression help to identify this as a negative statement.

VIII. Conclusion

Researchers in this area have long struggled with the difficulty of identifying covert emotions like sarcasm and irony. Since the book does not explicitly discuss these feelings. Research in this area should move toward integrating multi-modal inputs with an emphasis on uncovering latent emotional states. Further research on "multi-modal fusion approaches," such "decision-level" and "Meta-level" fusion, is something we want to do in the future. The fundamental benefit of a "convolutional neural network" and support vector machine is that they may be used in many different domains with just minor adjustments to the tuning. With the goal of improving the effectiveness of "multimodal feature fusion," we want to scale up our model testing to a social media big data setting and investigate other fusion techniques.

References

[1]. Perez-Rosas, V., Mihalcea, R., Morency, Utterance-level multimodal senti-ment analysis. In: ACL (1). (2013)

[2 Wollmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., Morency, L.P.: Youtube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intelligent Systems 28 (2013)

[3]. Poria, S., Cambria, E., Gelbukh, A.: Deep convolutional neural network textual fea-tures and multiple kernel learning for utterance-level multimodal sentiment anal-ysis. In: Proceedings of EMNLP. (2015)

[4]. M. Wang, D. Cao, L. Li, S. Li, and R. Ji, “Microblog Sentiment Analysis Based on Cross-media Bag-of-words Model.,” ICIMCS, pp. 76–80, 2014.

[5]. Cao D, Ji R, Lin D, et al. A cross-media public sentiment analysis system for microblog[J]. Multimedia Systems, 2014: 1-8.

[6]. Fuhai Chen, Yue Gao, Donglin Cao, and Rongrong Ji, “Multimodal hypergraph learning for microblog sentiment prediction,” presented at the 2015 IEEE International Conference on Multimedia and Expo (ICME), 2015, pp. 1–6.

[7]. S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, “Fusing audio, visual and textual clues for sentiment analysis from multimodal content,” Neuro computing, vol. 174, pp. 50–59, Jan. 2016.

[8]. M. Katsurai and S. Satoh, "Image sentiment analysis using latent correlations among visual, textual, and sentiment views," presented at the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 2837–2841.

[9]. Basant Agarwal, Soujanya Poria, Namita Mittal, Alexander Gelbukh, and Amir Hussain. 2015. Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. Cognitive Computation 7, 4 (2015), 487–499.

[10]. Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, 1–10.

[11]. Metallinou, A., Lee, S., Narayanan, S.: Audio-visual emotion recognition using gaussian mixture models for face and voice. In: Tenth IEEE International Sympo-sium on ISM 2008, IEEE (2008)

[12]. Eyben, F., W•ollmer, M., Graves, A., Schuller, B., Douglas-Cowie, E., Cowie, R.: On-line emotion recognition in a 3-d activation-valence-time continuum us-ing acoustic and linguistic cues. Journal on Multimodal User Interfaces 3 (2010)

[13]. Wu, C.H., Liang, W.B.: Emotion recognition of aective speech based on multiple classiers using acoustic-prosodic information and semantic labels. IEEE Transactions on A ective Computing 2 (2011)

[14]. Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. In CVPR, 2010

[15]. Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, J.Z. Wang, Jia Li, and Jiebo Luo. Aesthetics and emotions in images. 28:94–115, 2011

[16]. Damian Borth, Tao Chen, Rong-Rong Ji, and Shih-Fu Chang. Sentibank: Large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In ACM, 2013

[17] .Robert Plutchik. The nature of emotions. 89:344, 2001.

[18]. L. Kaushik, A. Sangwan, and J. H. Hansen. Automatic audio sentiment extraction using keyword spotting. In *Proc. INTERSPEECH*, pages 2709–2713, September 2015.

[19]. L. Kaushik, A. Sangwan, and J. H. L. Hansen. Sentiment extraction from natural audio streams. In *Proc. ICASSP*, pages 8485–8489, 2013.