Visual Sentiment Analysis: An Analysis of Emotions in Video and Audio

Rushali A. Deshmukh, Vaishnavi Amati, Anagha Bhamare, and Aditya Jadhav

Abstract Natural Language Processing (NLP)-based sentiment analysis examines opinions, feelings, and emotions expressed in emails, social media posts, YouTube videos, reviews, business documents, etc. Sentiment analysis on audio and video is a mostly unexplored area of study, in which the speaker's sentiments and emotions are gathered from the audio, and feelings are gathered from the video. The goal of visual sentiment analysis is to understand how visuals affect people's emotions. Despite being a relatively new topic, a wide range of strategies based on diverse data sources and challenges has been developed in recent years, resulting in a substantial body of study. This study examines relevant publications and provides an in-depth analysis. After describing the task and its applications, the subject is broken down into different primary topics. The study also discusses about the general visual sentiment analysis design principles from three perspectives: emotional models, dataset creation, and feature design. The problem is formalized by considering multiple levels of granularity and components that can affect it. To accomplish this, the research study looks at a structured formalization of the task that is often used in performing text analysis and assesses its relevance to perform visual sentiment analysis. The discussion includes new challenges, progress toward sophisticated systems, related practical applications, and a summary of the study's findings. Experimentation was also conducted on the FER-2013 dataset from Kaggle for facial emotion detection.

Keywords Visual sentiment analysis (VSA) · Opinion · Support vector machine (SVM) · Convolutional Neural Network (CNN) · OpenCV

R. A. Deshmukh · V. Amati (✉) · A. Bhamare · A. Jadhav
JSPM's Rajarshi Shahu College of Engineering, Tathawade, Pune 411033, India
e-mail: vaishnaviamati2507@gmail.com

P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*, Lecture Notes in Networks and Systems 789, https://doi.org/10.1007/978-981-99-6586-1_21

1 Introduction

A huge amount of data is produced per second in today's world, and making sense of it is a tedious effort. It's crucial to highlight that sentiment detection using text is still a work in progress, and while product reviews have received a lot of attention, we are concentrating on dual sentiment detection in videos using text analysis.

The classification of an input text as positive, negative, or neutral in terms of polarity is the basic task in sentiment analysis. This analysis can be done at the level of the document, sentence, or feature. Consumer perceptions of goods, commodities, branding, political views, and social activities can be captured using methodologies from this field. Analysis of Twitter users' activities, for example, can aid in predicting reputation of political groups or alliances. Sentiment analysis studies when it comes to micro blogging revealed that Twitter messages accurately consider the political situation.

One of the most delicate academic topics is mental health, because it is heavily influenced by the people mindset and feelings. The use of social media platforms like Facebook, Instagram, Flickr, and other grows daily, with photographs and videos playing an increasingly important role. Nowadays, our emotions may be deduced from our facial expressions. We can learn about each other's moods by observing their facial expressions. Sentiment analysis plays a critical role in making this recognition easier and more efficient. The word "sentiment," which means "emotions," will be assessed using the sentiment analysis system. Our objective is to predict sentiments using video because the majority of prior research has been on text-based sentiment analysis.

Though academics in NLP and pattern extraction presented numerous techniques to handle the issue of sentiment analysis, the social networking setting presents several unique obstacles. Aside from the massive volumes of data present, most verbal exchanges on virtual communities are short and informal. Further, in addition to verbal communications, users increasingly use photographs and videos to represent themselves on even the most popular social media sites. The data provided in such video images is connected not just to semantic contents such as things or activities in the obtained image and also to impact and sentiment signals communicated by the displayed picture. As a result, such data is important in determining the emotional effect even beyond semantic. As a result, photographs and videos are one of the greatest common methods for individuals to show their feelings and share their views on social networking, which has become increasingly important in gathering information regarding folk's thoughts and emotions.

2 Literature Survey

In Paper [1]: in this research, a method for automatically recognizing human emotion is developed utilizing CNN and facial expressions. Author has applied BPNN, CNN, and SURF feature extraction methods. Data was collected from the CASIA webface. The system has an 88% accuracy rate. The prediction was constrained by the small dataset (200 samples).

In Paper [2]: a voice-to-text conversion and management application was developed by the study's author using the Google Cloud Speech API and a collection of user-generated audio text files. They were devoid of any relevant textual data that may have enabled the user to change the mistakenly recognized content. The system utilized Google Cloud Speech API methods. Their organization and study were beneficial and could be used as proof.

In Paper [3]: in order to identify the emotions, this article uses deep learning-based multimodal emotion recognition from speech and facial expression. This study bases its ability to recognize emotions on deep learning. There is a dataset of verbal and facial expressions in the text. The system used the CNN and LSTM algorithms. While researching more efficient feature extraction techniques and multimodal fusion, they did not incorporate modalities like text and gesture into multimodal models. Combining speech and facial expression data has substantially enhanced the evaluation methodologies. They also contrasted their approach with current multimodal systems and found a significant improvement.

In Paper [4]: the author of the research utilized a collection of YouTube video comments to anticipate the sentiment using YouTube videos. Author used NLP to analyze the sentiment of customer reviews. 75.435% of the relevant video access is accurate. As a result, it may be concluded that their technique may correctly predict a favorable conclusion if a YouTube video is examined based on comment language.

In Paper [5]: this work used ML integrated with IOT to introduce sentiment analysis and mood detection on the Android platform, which can recognize the emotion. The North Face, Google Now, Alexa, Akinator, and chatbots were just a few of the tools used by the study's author. Data collection activities involved social media. There is no such possibility in emotion analysis or mood prediction. This study seeks to provide an explanation from the source, the main factor that underlies all of the issues, in order to address the problem, which appears to be challenging and intriguing.

In Paper [6]: in order to predict the sentiment on an image, this study suggests a machine learning-based classification method that uses SVM classifiers. The CNN + SVM algorithms are used in the article. The author utilized the Twitter and Tumbler dataset. The accuracy rate for the task was 99.2%. As a result, they lacked a deep learning method for multimodal sentiment analysis.

In Paper [7]: the sentiment on user-generated video, audio, and text was predicted by the author using the dataset of user-generated audio, video, and text. Python, SVM, the decision tree method, and OpenCV were some of the approaches used.

During testing, the task has a 70% accuracy rate. In conclusion, no certain precision was attained.

In Paper [8]: the work of sentiment analysis and topic recognition in video transcriptions is presented in the publication. Two of the author's key methods were SVM and LSTM. The details were supplied via the MUSE-TOPIC SUBCHALLENGE. Accuracy on the test set was 66.16%, whereas development accuracy was 56.18%. The groups were indicated rather than clearly determined from continuous single, necessitating more investigation.

In Paper [9]: in this study, the datasets from Twitter, Flickr, and Instagram were used. They introduced ME2M, a simple-to-use but effective model for picture sentiment analysis. The ME2M model's usefulness and applicability were shown by the observed results. They lacked any popularity forecast based on image sentiment analysis.

In Paper [10]: according to the study, deep learning might be used for picture sentiment analysis. Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Region Convolutional Neural Networks (RCNN), and Fast RCNN were some of the techniques used by the author. The FERET dataset was used in the study. They highlighted some of the important work that has been done for picture sentiment analysis when combined with deep learning approaches throughout the years in this study. As a result, there was no room for sentiment analysis or mood recognition.

In Paper [11]: improving sequence-to-sequence voice conversion by adding text-supervision, the author used the text-based phonetic information dataset. Machine learning methods such as the Hidden Markov model (HMM) and seq2seq VC model were used. Although the proposed methods considerably improve the seq2seq VC model, model execution is still hindered when there is a lack of training data. As a result, they had to deal with the challenge of performing significantly poorer when there are only a few training sets available.

In Paper [12]: the C3D network, VGG16 network, and ConvLSTM model approach were used to do sentiment recognition for brief annotated GIFs. The verbal emotion scoring rate is derived using the SentiWordNet3.0 model after that. Data included a gif video. Extensive testing encompassing both theoretical and practical assessments have proven the efficacy of the provided GIF video sentiment analysis program.

They were therefore without an effective strategy for dealing with complex parameters that appeared in brief annotated GIFs.

In Paper [13]: the study focuses on sentiment analysis and emotion identification in static images. The author made use of the UMD faces dataset. VGGNet16 and CNN models were used as techniques. The work lacked an efficient system for dynamic visuals. The suggested approach beats prior models and yields more accurate, upbeat results when tested on a model to estimate.

In Paper [14]: this study suggests a brand new facial expression element for sentiment analysis of videos. In order to validate our suggested feature, we employ a machine learning framework. The outcomes of the trial show that the feature is beneficial.

In Paper [15]: an overview of current developments in the field of multimodal sentiment analysis was provided in this survey study. The most prominent feature extraction techniques and datasets in the areas have been sorted into categories and discussed. On the CMU-MOSI and CMU-MOSEI datasets, two frequently used datasets for multimodal sentiment analysis, the efficacy and efficiency of the thirty-five models have been examined.

3 Related Work

A CNN is a deep learning system that can recognize when an image has been processed, assign significance to various aspects within the image, and differentiate between them. Excellent software for image processing and computer vision is OpenCV. It is just a free-source library providing operations including object tracking, finding landmarks and face detection, among others. The machine learning method "support vector machine" could be used to resolve regression and classification problems. It really is, however, generally employed in categorization difficulties. PCA is a method for lowering the dimension of these kind of datasets, boosting accurateness while minimizing data redundancy. To extract the sentiment of each word, each utterance, and eventually, each video, the CNN converts a textual utterance to a logical form: a machine-understandable representation of its meaning. Block diagram of sentiment approach is shown in Fig. 1.

Our aim is to predict the sentiments from video, so we will be using a video to create the data for audio as well as for video we will be capturing certain pictures from the provided video and applying CNN algorithm for image to text conversion, and similarly for the audio, we will be collecting the audio data and applying the ML approach that is the CNN for speech to text conversion. Features will be extracted from the video and audio through OpenCV model like detection of faces in this

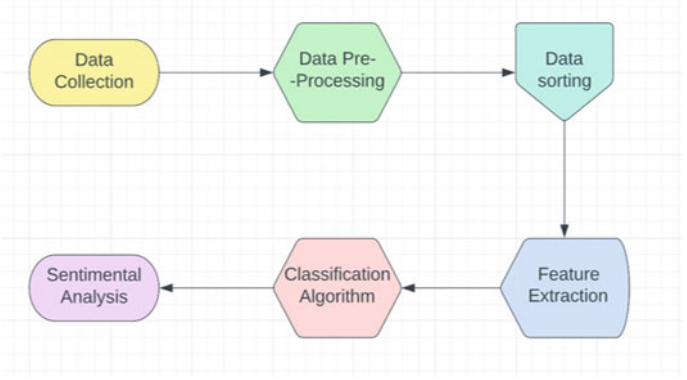

Fig. 1 Block diagram of sentiment approach

case eyes movement of lips and keywords/phrases in case of audio. CNN employs a feature extractor during the training phase. The weights are determined by training specialized neural network types that make up CNN's feature extractor. A neural network called CNN extracts the features of the input images, while a different neural network categorizes the characteristics. The feature extraction network uses the input image as a starting point. The neural network uses the extracted feature signals for classification. The result is subsequently generated by the neural network categorization based on the picture characteristics. The convolution layer stacks and sets of pooling layers are part of the neural network for feature extraction. The convolution layer, as its name suggests, uses the convolution method to modify the picture. For the classification, SVM and PCA techniques will be applied for video as well as for audio which will classify the extracted features such as happy, sad, anger, and surprise.

Expression Intensity: The recognition of an expression is significantly influenced by the expression's intensity. When the expression is less subtle, it is easier to recognize it. It has a significant impact on the model's accuracy.

- Step I: Get the image frame from a data.
- Step II: Image preprocessing (cropping, resizing, rotating, color correction).
- Step III: To use a CNN model, extract the key features.
- Step IV: Categorize your emotions.
- Image and Video Frame Face Detection

The human face is detected and located in the first stage using video from a camera. In real time, the coordinate of a frame is to determine the position of the real face. Face recognition is still a challenging procedure furthermore, it is not assured that all faces in a specific input picture will be retrieved, particularly in uncontrolled conditions with inadequate illumination, varying head positions at long a distance or an obstruction.

II. Image Preparation

After the faces are discovered, the pictures are optimized before being sent to the sentiment classifiers. This action greatly enhances the classification accuracy. Validating the image for varying illumination, thresholding, picture reduction, fixing picture rotation, sizing the picture, and cutting the picture are all important substeps in image preprocessing.

III. AI Model for Emotion Classification

Soon after preprocessing, the required features have been extracted from which was before data containing the discovered faces. There are numerous approaches for detecting various aspects of the face. For example, Action Units (AU), face landmark motions, landmark distances, features of gradients, face texture, and so forth. The most common utilized classifiers in AI emotion identification are SVM or CNN. Finally, the detected human face is assigned a pre-defined class (label) based on facial expression, such as "joyful" or "neutral."

3.1 Facial Expression Recognition of FER-2013

The FER-2013 dataset for facial emotion detection is provided by Kaggle, and this dataset was introduced at the International Conference on Machine Learning (ICML). Few images of the dataset are shown in Figs. 2, 3, 4, 5, 6, 7, and 8.

Each face in this dataset has been categorized on the basis of emotion categories, where the grayscale of every image is 48pixelx48pixel. In Fer-2013 dataset, there are 35,887 number of images with seven distinct expression kinds are identified by

Fig. 2 Angry

Fig. 3 Disgust

Fig. 4 Fear

Fig. 5 Happy

Fig. 6 Neutral

Fig. 7 Sad

Fig. 8 Surprise

seven distinct categorization descriptors. Number of data in the FER-2013 is given in Table 1.

Table 1 Number of data in the FER-2013

Micro-expression (classification)	Validation data		Training data	Dataset total
	Public	Private		
Angry	467	491	3995	4953
Disgust	56	55	436	547
Fear	496	528	4097	5121
Happy	895	875	7215	8989
Neutral	607	626	4965	6198
Sad	653	594	4830	6077
Surprise	415	416	3171	4002
	3589	3589	28,709	35,887

3.2 Micro-Classification of Facial Expression

In social psychology, a micro-expression is a facial expression that is simple to see and recognize as a form of communication. Information is transmitted through facial expressions about emotions, our objectives and goals, and are fundamental to interpersonal communication. Understanding and being able to read facial emotions naturally makes the desired conversation easier. The classification of human facial expressions involves three steps: face recognition, feature extraction, and facial expression classification. The authors of this study used a method that could categorize facial expressions on a large scale and included seven fundamental human expressions (Figs. 9, 10, 11, 12, 13, 14, and 15).

Human Face is detected as following:

1. Eyebrows pulled down (shows anger)
2. Eyebrows pulled up and together (shows fear)
3. Upper lip pulled up (shows disgust)
4. Eyes neutral (shows neutral)

Fig. 9 Features of joyful expressions

Fig. 10 Anger expression characteristics

Fig. 11 A sad expression's defining features

Fig. 12 Typical fear expression

Fig. 13 Disgust expression

Fig. 14 Typical surprise expression

Fig. 15 Neutral face

5. Cheeks raised (shows happy)
6. Lip corners pulled down (shows sad)
7. Mouth hangs open (shows surprise)

(1) Happy

A smile is a facial expression that can convey enjoyment or like for something. The happy expression is characterized by an upward movement of the cheek muscles and the sides or edges of the lips to form a smile.

(2) Anger

When expectations and reality diverge, angry facial expressions result. The expression is visible in the way the eyes are focused when staring, the way the lips are contracting, and the way the inner eyebrows on both sides are merging and bending down.

(3) Sadness

Based on the traits of a sad facial expression, a sad face will arise when there is disappointment or a sensation of missing something which includes a loss of focus in the eye, a downward pull of the lips, and a drooping of the upper eyelid.

(4) Fear

Fear is a type of expression that manifests when a person finds themselves unable to handle a situation or in a frightening environment. The two eyebrows that raise simultaneously, the tightened eyelids, and the horizontally wide lips all indicate anxiety on a person's face.

(5) Disgust

A person who displays facial disgust after witnessing something unusual or after listening to information that is unimportant. A person's face will show signs of distaste when the upper lip raises and wrinkles appear at the nasal bridge.

(6) Surprise

When someone receives a sudden, unexpected, or significant event or communication and is unaware of it previously, they will express surprise. A surprised expression is depicted by the lifted brows, wide-open eyes, and reflexive widening of the mouth.

(7) Neutral

A person who is perceived as snobbish and lacking in regard for others frequently underestimates others by their facial expression.

4 Results Analysis

The system was tested in this work at various stages of the design recognition of facial micro-expression. The outcomes demonstrated that the face expression detection system could use the CNN architectural model in an ideal and timely manner. In Table 2 according to the evidence, data training can be carried out most effectively when utilizing a separate convolution layer, and trained model's face expression can be accurately predicted for anger 0.40%, disgust 0.24%, fear 0.35%, happy 0.66%, neutral 0.40%, sad 0.37%, surprise 0.68% of the time. Analysis of the system's results after implementation is absolutely necessary.

Table 2 Result of facial expression testing

Class	Accuracy	Sample
Angry	0.40	600
Disgust	0.24	66
Fear	0.35	615
Happy	0.66	1083
Neutral	0.40	745
Sad	0.37	725
Surprise	0.68	476

Table 3 Confusion matrix

Class	Angry	241	5	74	62	66	113	39
	Disgust	10	16	6	13	3	11	7
	Fear	91	0	215	43	59	119	88
	Happy	79	2	40	720	78	85	79
	Neutral	74	4	86	86	301	114	80
	Sad	98	3	118	84	112	266	44
	Surprise	32	0	53	33	17	18	323
		Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Prediction								

4.1 Prediction Test of Facial Expression

For all seven expressions, experiment is carried out for ten times, and the system is successful to recognize the expression.

The outcomes of expressing anger and fear incorrectly came about once each, whereas expressing disgust incorrectly came around twice. Table 3 of the report displays the findings. Table 3 displays which expressions are straightforward to anticipate and which ones are more challenging.

5 Future Scope

Companies may learn via sentiment research how consumers feel about a brand, whether it's favorable, negative, or neutral. One of the most crucial methods for retaining clients' attention and engagement is brand monitoring, which includes sentiment research. Anyone can use sentiment analysis to assemble and evaluate massive volumes of text data, such as news, social media, views, and suggestions, to predict the outcome of an election. It considers how both candidates are seen by the

general population. The availability of huge and stable datasets makes a significant contribution in this regard. Indeed, we brought out some difficulties with the available datasets in this research. Modern social media platforms allow for the collection of large volumes of photographs as well as a range of linked data. These can be used to specify both input and "ground truth" properties. To avoid the association of noisy data with the photos, these textual data must be adequately filtered and processed, as previously described. Systems with larger purposes could be designed to address new difficulties or to focus on new emergent tasks. For example, idea programs can help people bridge the gap between real and virtual communication. Emoji's have been growing in popularity for years, mainly to the proliferation of social media platforms, and they are now an essential element of how people communicate online. They're commonly used to convey user reactions to messages, photos, or breaking news. As a result, investigating novel communication routes may help to improve present state-of-the-art performance. This can also be utilized in cybercrime to study criminals' expressions in order to determine the true reason for the malpractice committed by the crooks.

6 Conclusion

The study's purpose was to create a system that is flexible, cost-effective, adaptable, and, most importantly, portable. It's a trustworthy method for ensuring the accuracy of social product reviews. Machine learning is where our proposed sentimental analysis system fits in. Our main goal was to achieve high-accuracy sentimental video detection. This analyzing feature can also help us analyze video reviews. Many social media platforms now demand audio and video surveillance, including Facebook, Twitter, and YouTube. Using our technology, we can analyze consumption and detect opinion in a certain product. Because of the rapid expansion of social media, multimedia data has become a crucial transporter of human thoughts and opinions. The study of social networks has risen to prominence as a possible research area. We looked at the most common methodologies for textual sentiment analysis on social media based on a superficial assessment. The most common multimodal sentiment analysis approaches, as well as visual sentiment analysis were examined.

The goal of this work was to provide a thorough examination of the visual sentiment analysis topic, related challenges, and region techniques. Significant meaning with real enterprise software that would benefited from sentiment analysis on image and video research has indeed been explored.

Acknowledgements We would like to thank our guide (Dr. Rushali Deshmukh) who gave us this opportunity to work on this project. We got to learn a lot from this project about various machine learning techniques used in sentiment analysis. It gives us tremendous pleasure to extend our sincere appreciation to Dr. R. K. Jain, principal, JSPM's RSCOE, Tathawade, Pune, for providing necessary infrastructure and creating good environment. At last, we would like to extend our heartfelt thanks to our teachers because without their help this project would not have been successful.

References

1. Madupu RK, Chiranjeevi K, Vasanthi Y, Sonti H, Basha CZ (2020) Automatic human emotion recognition system using facial expressions with convolution neural network. In: 2020 4th international conference on electronics, communication and aerospace technology (ICECA. IEEE), pp 1179–1183
2. Choi J, Gill H, Ou S, Song Y, Lee J (2018) Design of voice to text conversion and management program based on Google Cloud Speech API. In: 2018 international conference on computational science and computational intelligence (CSCI). IEEE, pp 1452–1453
3. Cai L, Dong J, Wei M (2020) Multi-modal emotion recognition from speech and facial expression based on deep learning. In: 2020 Chinese automation congress (CAC). IEEE, pp 5726–5729
4. Bhuiyan H, Ara J, Bardhan R, Islam MR (2017) Retrieving YouTube video by sentiment analysis on user comment. In: 2017 IEEE international conference on signal and image processing applications (ICSIPA. IEEE), pp 474–478
5. Kushawaha D, De D, Mohindru V, Gupta AK (2020) Sentiment analysis and mood detection on an Android platform using machine learning integrated with Internet of Things. In: Proceedings of ICRIC 2019: recent innovations in computing. Springer International Publishing, pp 223–238
6. Das P, Ghosh A, Majumdar R (2020) Determining attention mechanism for visual sentiment analysis of an image using SVM classifier in deep learning based architecture. In: 2020 8th international conference on reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). IEEE, pp 339–343
7. Rao A, Ahuja A, Kansara S, Patel V (2021) Sentiment analysis on user-generated video, audio and text. In: 2021 international conference on computing, communication, and intelligent systems (ICCCIS). IEEE, pp 24–28
8. Stappen L, Baird A, Cambria E, Schuller BW (2021) Sentiment analysis and topic recognition in video transcriptions. IEEE Intell Syst 36(2):88–95
9. Zhang H, Wu J, Shi H, Jiang Z, Ji D, Yuan T, Li G (2020) Multidimensional extra evidence mining for image sentiment analysis. IEEE Access 8:103619–103634
10. Mittal N, Sharma D, Joshi ML (2018) Image sentiment analysis using deep learning. In: 2018 IEEE/WIC/ACM international conference on web intelligence (WI). IEEE, pp 684–687
11. Zhang J-X, Ling Z-H, Jiang Y, Liu L-J, Liang C, Dai L-R (2019) Improving sequence-to-sequence voice conversion by adding text-supervision. In: ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing (ICASSP, IEEE), pp 6785–6789
12. Liu T, Wan J, Dai X, Liu F, You Q, Luo J (2019) Sentiment recognition for short annotated GIFs using visual-textual fusion. IEEE Trans Multim 22(4):1098–1110
13. Doshi U, Barot V, Gavhane S (2020) Emotion detection and sentiment analysis of static images. In: 2020 international conference on convergence to digital World-Quo Vadis (ICCDW). IEEE, pp 1–5
14. Li H, Xu H (2019) Video-based sentiment analysis with hvnLBP-TOP feature and bi-LSTM. Proc AAAI Conf Artif Intell 33(01):9963–9964
15. Abdu SA, Yousef AH, Salem A (2021) Multimodal video sentiment analysis using deep learning approaches, a survey. Inf Fusion 76:204–226