Proceedings of the 3rd International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/13/20240860

Multimodal sentiment recognition: A comprehensive review of analysis techniques, applications, and challenges

Yuetong Hao

Chongqing Nankai Secondary School, Chongqing, 400030, China

mlewis73798@student.napavalley.edu

Abstract. This article offers a comprehensive review of multimodal emotion recognition. Initially, we provide a succinct overview of human emotions by delving into various emotion models. The focus then shifts to the representation of emotions through unimodal data sources, illustrating how these singular channels capture emotional cues. As we progress, the article underscores the advancements in multimodal fusion techniques that amalgamate multiple data sources for more accurate emotion detection. Three predominant architectures for multimodal fusion are then detailed, each showcasing its unique approach and benefits. Despite the leaps and bounds made in this field, several challenges persist. The latter section of this review highlights these lingering issues in multimodal emotion recognition, hinting at potential areas for further research and development. By weaving together the intricacies of emotion models, unimodal representations, fusion techniques, and existing challenges, this paper seeks to provide readers with a holistic understanding of the current landscape of multimodal emotion recognition.

Keywords: Multimodal, Sentiment Recognition, Multimodal Fusion, Multimodal Fusion Architecture.

1. Introduction

Facial expressions, posture, and vocal intonation encompass the broad spectrum of non-verbal human communication. Facial expressions are not only an innate means of conveying emotions but are also the primary indicator by which we discern others' feelings. Modern tools allow us to study minuscule changes in the face, enabling us to differentiate between genuine smiles and insincere ones. For instance, a genuine smile sees a lift in the cheeks and the muscles around the eyes, accompanied by increased electrical activity in the left hemisphere. In contrast, a feigned smile typically involves only the lips, with less pronounced activity in the left hemisphere. In-depth studies into human emotions began in the late 19th century. In the digital age, as technology advanced, there arose a vision of imbuing machines with the ability to "feel". This has become a leading research direction, focusing on deriving emotional attributes from sensory signals and understanding the interplay between human emotions and these signals. Shifts in mood and cognitive patterns in humans are always associated with fluctuations in various physiological or behavioral traits. These are shaped by numerous factors, including environment, cultural background, and individual personality. To enable machines to interpret emotions, understanding human interaction is paramount. Humans convey emotions through facial expressions, bodily movements, and speech, and detect emotional shifts using sight, sound, and touch [1]. Visually, emotions are primarily discerned from facial expressions and gestures; whereas sound is dominated by

DOI: 10.54254/2753-8818/13/20240860

speech and music. Tactile senses range from the feel of a caress, handshake, perspiration, to the rhythm of a heartbeat.

While the face, posture, and voice can independently convey specific emotions, interpersonal communication often synthesizes information from all these sources. Thus, only through a multi-modal human-machine interface [2] can we achieve the most organic interaction between humans and computers. This interface amalgamates natural language, voice, sign language, facial cues, lip movements, head nods, and body language into a cohesive system, encoding, compressing, integrating, and fusing visual, auditory, and textual multimedia data. Today, multi-modal technology stands at the forefront of human-computer interaction research. The fusion of emotional computing with multi-modal processing can harness a comprehensive emotional palette, significantly enriching the depth of research in emotional computing and paving the way for a more nuanced and harmonious human-computer interaction system [3].

2. Theoretical Frameworks for Emotion Expression

To realize speech emotion recognition, we need to define emotion first. The emotion description model combines emotion representations as a combination of mutually exclusive discrete emotion categories or numerical dimensions. According to the different representation methods, it is divided into discrete emotion model and dimensional model [4].

2.1. Discrete Emotion Models

And these six basic emotions can be combined to drive other compound emotions. Roseman et al. assessed emotions through evaluation factors and gave seventeen basic emotions. Due to the complexity of emotions, it is difficult to simulate them accurately [5]. However, classification models for discrete emotion models are relatively common in practical use scenarios.

2.2. Dimensional Emotion Models

The dimensional emotion models can be constructed in two-dimensional or multi-dimensional spaces to describe continuous emotions. Russell proposed a two-dimensional emotion model based on the two dimensions of and motivation, using a ring structure to describe the emotion [6].

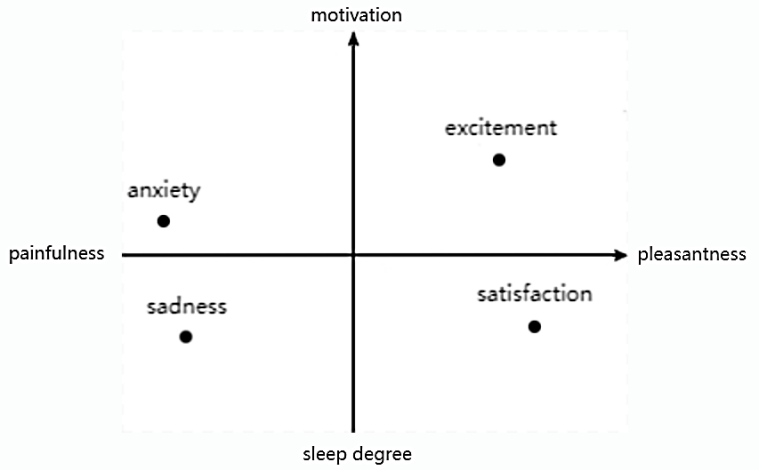

Figure 1. Sentiment distribution (Photo/Picture credit: Original).

In the dimensional emotion model, the PAD model with the highest recognition is based on the three dimensions of pleasure, arousal and dominance [7]. The PAD three-dimensional model can effectively explain the human emotions, simulating the similarity and antagonism of emotions. As shown in Figure 1.

DOI: 10.54254/2753-8818/13/20240860

3. Unimodal Emotion Analysis Methods

Emotion analysis is mainly the analysis of people's emotions through some ways such as facial expression to express their emotions. Different people have different ways expressing emotions: when a person tends to express emotions in language, their audio features may contain more emotional cues; if a person tends to use facial expressions [8], their facial expression features may contain more emotional cues. At present, the mainstream unimodal emotion analysis is mainly based on facial expression, textual, acoustic and image and video.

3.1. Facial Expression Analysis

In daily life, facial expression information is a common way for people to obtain the emotional state with each other. Thus, facial expression information has an important significance in the process of emotion analysis. According to the different feature representation, FER system can be divided in to two categories: static image and dynamic sequence. In the dynamic sequence of FER, facial expression show two characteristics: timing and significance. Significance of facial expressions is often ignored in FER of dynamic sequences [9]. To solve this problem, a facial expression recognition method based on space-time attention network adds corresponding attention modules to the airspace sub-network and time domain sub-network to improve the performance of CNN and RNN when extracting features.

3.2. Textual Analysis

For the textual analysis, it is mainly to convert the text into a language recognized by the machine. There are usually two representations: One-hot Representation and Distribution Representation. The commonly used text feature tool is time word embedding model. The word embedding represents the text by vector, the dimension is defined in advance, and similar words will have similar vector representation. Commonly used word embedding models such as the Word2vec model published by Google mainly rely on two models, Skip-grams or CBOW, to predict nearby words through central words and central words through nearby words, such as Word2vec model for feature extraction of text modes. The GloVe word vector takes into account global information using the co-occurrence matrix, and ELMo word vector can capture the meaning of words and context as the language environment changes. In 2018, Google proposed the Bert pre-training model [10], many scholars use large-scale corpus for pre-training and learn semantic relations and input downstream task word vectors. For example, uses GloVe word vector and Bert model respectively for the representation of text features and compare their performance.

3.3. Acoustic Analysis

Acoustic features cover rich information, and the emotional information transmitted can be obtained through the analysis of acoustic features, which has a significant impact on the emotion recognition of the classifier. The most commonly used acoustic features are Mel Frequency Cepstral Coefficient, energy or amplitude features and Linear Predictor Cepstral Coefficients and so on. Based on Python Librosa tools can be time-frequency processing, extract a variety of speech features, draw sound all kinds of related images such as spectrum, Schuller team in 2015, can develop OpenSmile tool for speech prepossessing and feature extraction, the frame energy, frame strength, autocorrelation function, amplitude spectrum weighted features to extract. For example, using OpenSmile to extract MFCC and LPCC and so on for emotional recognition features of language.

3.4. Image and Video Analysis

Face information is based on the facial features and other constantly changing, contains rich emotional information. Face image and video feature extraction is mainly based on ensemble features and texture features. Set features are represented using a set of vectors according to the position, size and proportion of the facial features. The texture features mainly include SIFT, local binary model, Gabor wavelet coefficient, HOG and so on. For the dynamic image sequence, the optical flow method reflects the gray scale change in the dynamic frame and can reflect the movement based on the face muscle. The Python

DOI: 10.54254/2753-8818/13/20240860

OpenCV and Dlib libraries and often used for face feature key point recognition. The Open-Face tool proposed by Brandon et al. in 2016 can also extract facial features and obtain low-dimensional representations for expression analysis.

4. Multimodal Emotion Analysis Methods

Single emotion analysis has limitations such as recognition rate and poor stability. In the development of emotion analysis, researchers use a variety of modes to conduct emotion analysis to improve its accuracy and stability.

In the existing literature, multimodal-based sentiment analysis requires modal fusion in addition to unimodal feature extraction.

4.1. Data Level Fusion

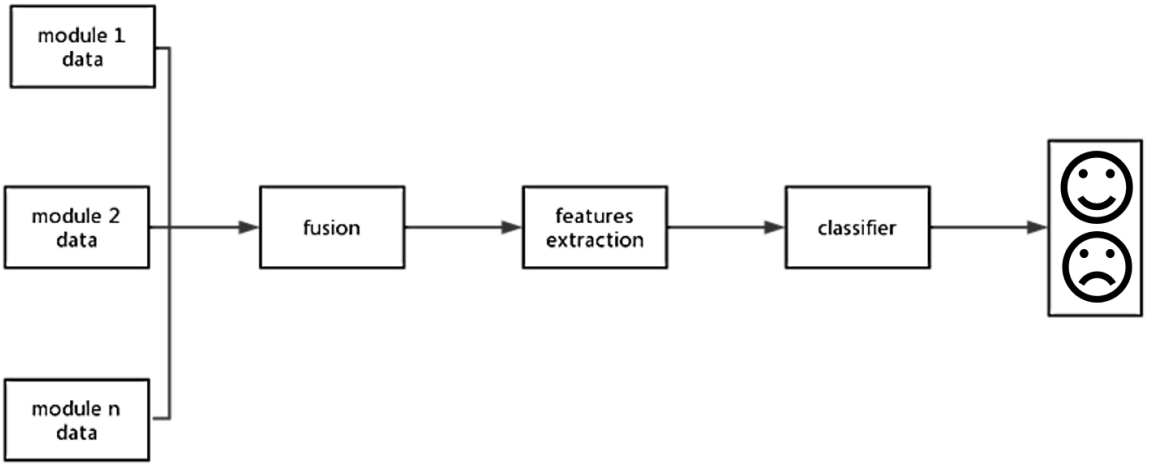

Figure 2. Emotion recognition flowchart (Photo/Picture credit: Original).

The advantage is that it can store data information in each sensor mode properly to avoid loss of information and maintain information integrity. However, the weakness is obvious. The processing process is very complicated because the data is fused in its original state. As shown in Figure 2.

4.2. Feature Level Fusion

Feature level fusion is the shallow fusion in the early stage after feature extraction, which is the direct connection of multiple modes, that's the splicing, addition and weighted sum of shallow layers. Before performing deep learning, feature engineering is often used to extract modal features. Feature fusion aims to integrate multiple features of different modes into a common viewing space. Due to the difference of each mode, a large amount of redundant information is often covered. Dimensional reduction method will be adopted to eliminate redundant information, and principal component analysis is usually adopted.

For example, Emerich et al. combine the length-normalized speech emotion features and facial expression features to construct a feature vector. When the feature dimension reaches a certain scale, the performance of the model will decline. For this, a feature level fusion strategy based on sparse kernel reduced-rank regression (KRRR) was proposed by Yan et al.. OpenSMILE feature extractor and SIFT descriptor extract effective features from speech modality and facial expression modality, respectively, and then fuse the emotional features of the two modalities using the SKRRR fusion method.

When the modal information is targeted at the same content but is not contained to each other, although the feature level fusion method can retain the original information to the maximum extent and can achieve the best recognition effect in theory, it does not take into account the differences between different modal emotional features.

DOI: 10.54254/2753-8818/13/20240860

4.3. Model Level Fusion

Model level fusion is also known as mid-term model fusion. Model-based fusion is to input different modal data into the network and based on the middle layer of the model. The benefit of model fusion is that the fusion location can be chosen, and the interaction between modes can also be achieved. Model-based fusion usually uses tmulti-kernel learning, neural net works, image models, and other methods.

For example, Zheng et al. use the stacked confined Boltzmann machines to expand into a deep confidence network. They first used the manually extracted EEG and eye movement features as input to the two Boltzmann machines separately and learned a shared representation of the two modes from the neural network. The experimental results show that the model level fusion based on the deep neural network can significantly improve the performance. The flexibility to choose fusion locations is the most significant advantage of model level fusion.

Zhang et al. put forward a make full use of deep neural network powerful feature learning ability of mixed deep learning model, audio-visual data by convolutional neural networks (CNN) and 3DCNN audio-visual segment features into the deep confidence network, joint learning a audio-visual features, on emotion recognition task than manual features and deep learning fusion method performed better.

4.4. Decision Level Fusion

Decision level fusion is to score the results of each mode after the training of each single mode in the later stage, that is, to integrate the prediction results of each mode. When some modal data is missing, decision level fusion can also have a good performance, and the data from different modes can be trained with appropriate classifiers, and the errors between different modes will not affect each other. The common fusion mechanisms of decision level fusion include weighting, voting, integrated learning, rule fusion and so on.

Huang et al. used enumeration weights and adaboost, two different decision level fusion strategies to compare the effect of emotion recognition, used the facial expression classifier and EEG classifier as the enhanced sub-classifier, and applied then to two earning tasks (valence and arousal) respectively. The results show that both methods can give the final valence and arousal results, and achieve good results in the public datasets DEAP and MSHNOB-HCI and online applications.

Therefore, the forecast results are somewhat inaccurate. Lu et al. use a fusion strategy called fuzzy integration was employed. The fuzzy integral is an integral of the real function of the fuzzy measure. The experiment found that the eye movement features and EEG have complementary effects on emotion recognition, and the best accuracy of the fuzzy integration fusion strategy was 87.59 percent. Compared with other fusion methods, the accuracy of emotion recognition can significantly improved by the fuzzy integration fusion. In general, the information between multiple modes is not completely independent, and the decision level fusion will lose the correlation between different modes, so the results of identification in the practical application environment may not be better than that of unimodal identification.

4.5. Mixed Fusion

This fusion approach combines the advantages of feature level fusion and decision level fusion, along with increasing model complexity and implementation difficulty. Due to the attention mechanism and GRU show good performance in emotion analysis, CHEN M H et al. proposed a multimodal embedded the LSTM model, with time attention gating model on word level fusion, and can focus on the most important time frame, solved the “at every moment to look for what kind of situation” and “when in the communication” the two key issues. SHENOY A et al. proposed an end-to-end RNN model for the analysis of emotion. This model can capture all modal conversation contexts, the dependence between the listener and the speaker’s emotional states, and the correlations between the available modes. Structurally, two gating loop units, sGRU and cGRU, were used to model the states and emotions for the interlocutors. In addition, an interconnected context network is used to learn context representations, and pairwise attention mechanisms are used to make simple representations of useful information for each modality.

DOI: 10.54254/2753-8818/13/20240860

The system was trained with 8898 images to obtain the CNN(f) model, with the output as a probability vector of seven discrete emotion categories. CNN(v) and CNN(a) were trained from a network of EEG and galvanic skin response (GSR) modes. The weighted units calculate the weighted sum of valence and arousal of CNN(v) and CNN(a) output, respectively, and then send it to the distance calculator to calculate the emotional distance. Finally, send the emotional distance to the decision tree, which is fused with the results obtained by CNN(f) to obtain the emotional state. GUNES H et al. Combined with facial expressions and gestures in videos, they propose a vision-based multimodal emotion analysis framework, which automatically identifies facial expressions and upper body gesture features from the video sequence for feature level fusion, and then uses the analysis results for decision level fusion by the method of product and weighting to obtain the results.

5. Multimodal Fusion Architectures

5.1. Collaborative Architecture

The goal of the collaborative architecture is to find the correlation between the various modes in the collaborative subspace. Multimodal collaborative architecture is to implement various single modes under the action of constraints. Such architectures are widely used in cross-modal learning. Methods based on cross-modal correlations aim to learn a shared subspace that maximizes the correlation of sets of different modal representations. The cross-modal similarity method maintains the inter-modal and inner modal similarity structure under the constraints of the similarity measure, making the cross-modal similarity distance as small as possible and the distance of different semantics as large as possible.

The advantage of collaborative architecture is that each independent architecture is that each independent mode can be run, and this advantage facilitates transfer learning across patterns, with the aim of transferring information between various modalities. But the disadvantage of such architecture is that mode fusion is difficult, and it is difficult to realize transfer learning between multiple (more than two) modes.

5.2. Joint Architecture

Joint mode refers to mapping a multimodal space into a shared semantic subspace to fuse multiple modal features. Each independent mode is mapped to a shared subspace, where it performs well in multimodal classification and regression tasks, such as emotion analysis and speech recognition. The core of the joint architecture is to achieve feature “fusion”, and direct addition is one of the simplest methods. Although the above method is simple to achieve, it is easy to cause semantic loss. The “multiplication” method optimizes this disadvantage and makes the feature semantics fully integrated.

This kind of architecture has relatively high requirements for the semantic integrity of a single mode, and the incomplete data will be solved in the later fusion. Compared with other architectures, the joint architecture has the advantage of simple fusion mode, and its shared subspace has semantic invariance, which is conducive to the transformation of one mode into another mode. The disadvantage is that each individual mode is more difficult to handle and find at early times.

5.3. Codec Architecture

Such architecture are generally used when mapping one mode to another mode is required, and consists of two parts: the decoder and the encoder. The encoder maps the initial mode into the vector, and the decoder generates a new mode based on the previous vector. Codec architecture is widely used in research fields such as video decoding, image annotation and image synthesis.

The advantage of such architectures is that a new one can be generated from the initial mode. The disadvantage is that each encoder and decoder can only uniquely encode one mode.

6. Challenges

The reliability of the features is not exactly the same between the different modules. Most of the current studies showed that the reliability of the text modality is relatively strong and there is a dependence

DOI: 10.54254/2753-8818/13/20240860

between the different modes. It is easy to produce high-dimensional disasters after the mode connection, which increases the computational complexity. Also, it is difficult to use the complementarity between modes when using models. Due to the different sampling rate, noise type, intensity, etc., and the different density of modes at the same time, information in different modes is difficult to align completely. Most multimodal sentiment analysis currently uses text, video, and audio, and datasets of modes such as EEG and physiological signals are missing. Human decision-making is highly irrational in some cases, so most of the studies mostly consider the fusion between different modes. There are usually multiple speakers in a conversation. There are usually interactions in human emotions in real world dialogue scenes, so timing information is particularly important and the same words will have different meanings in different dialogue scenarios.

7. Conclusion

Multimodal emotion recognition has witnessed significant advancements, particularly with the advent of sophisticated computational techniques such as CNN. Leveraging different modalities, such as valence and arousal from distinct CNN outputs, has enabled the calculation of emotional distances which can be integrated to determine emotional states. Innovative approaches, as highlighted by GUNES H et al., involve the fusion of facial expressions and gestures from video sequences, taking the domain of emotion analysis a step further. Three principal fusion architectures underscore the breadth of research in this arena: The Collaborative Architecture facilitates transfer learning across modalities and thrives on finding correlations between diverse modes. Despite its ability to operate independently across modes, it faces challenges in mode fusion, especially with more than two modes. The Joint Architecture maps multiple modes into a shared semantic subspace, proving invaluable in tasks like emotion analysis. Its main strength lies in its simple fusion approach and semantic invariance, although individual modes present challenges in early identification and handling. The Codec Architecture, comprising encoders and decoders, serves the primary purpose of mapping one mode into another. Its widespread applications span fields like video decoding and image synthesis. Its uniqueness in encoding one mode, however, can be a limiting factor.

References

[1] Chandrasekaran, G., Nguyen, T. N., & Hemanth D, J. (2021). Multimodal sentimental analysis for social media applications: A comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(5), e1415.

[2] Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems.

[3] Houssein, E. H., Hammad, A., & Ali, A. A. (2022). Human emotion recognition from EEG-based brain–computer interface using machine learning: a comprehensive review. Neural Computing and Applications, 34(15), 12527-12557.

[4] Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Umar, A. M., Linus, O. U., ... & Kiru, M. U. (2019). Comprehensive review of artificial neural network applications to pattern recognition. IEEE access, 7, 158820-158846.

[5] Salcedo-Sanz, S., Ghamisi, P., Piles, M., Werner, M., Cuadra, L., Moreno-Martínez, A., ... & Camps-Valls, G. (2020). Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources. Information Fusion, 63, 256-272.

[6] Bhuvaneshwari, M., Kanaga, E. G. M., Anitha, J., Raimond, K., & George, S. T. (2021). A comprehensive review on deep learning techniques for a BCI-based communication system. Demystifying big data, machine learning, and deep learning for healthcare analytics, 131-157.

[7] Zhu, X., Xu, H., Zhao, Z., & others. (2021). An Environmental Intrusion Detection Technology Based on WiFi. Wireless Personal Communications, 119(2), 1425-1436.

DOI: 10.54254/2753-8818/13/20240860

[8] Bharati, S., Podder, P., Mondal, M., & Prasath, V. B. (2021). Medical imaging with deep learning for COVID-19 diagnosis: a comprehensive review. arXiv preprint arXiv:2107.09602.
[9] Hooda, R., Joshi, V., & Shah, M. (2021). A comprehensive review of approaches to detect fatigue using machine learning techniques. Chronic Diseases and Translational Medicine.
[10] Zhang, D., Ma, M., & Xia, L. (2022). A comprehensive review on GANs for time-series signals. Neural Computing and Applications, 34(5), 3551-3571.