Multimodal Emotion Recognition System Based on Deep Learning and Data Fusion

## Meiyi Dong

**City University of Hong Kong, School of Electrical Engineering, Hong Kong, China**

[dongmeiyi@158xin.com.cn](mailto:dongmeiyi@158xin.com.cn)

**Abstract**:

With the development of the internet, multimedia has become one of the major information formats that we use in daily life. Multimedia information is combined not just with one data type but with various data types, also known as multimodal data. The diversity of categories makes the information richer, but it also brings difficulties and challenges. This research aims to build a solid and accurate multimodal emotion recognition system using deep learning and data fusion techniques. It would leverage multiple data sources such as physiological signals, speech tones, body movements, and facial expressions to enhance the accuracy and robustness of the emotion recognition model. By improving the algorithm and the data fusion strategy, this system would be able to accurately detect and respond to users' emotional states in real time. The research involves several key components: multimodal data collection, multimodal data fusion, feature extraction and representation, real-time system implementation, model training and validation, model explanation and evaluation, and application testing in diverse scenarios. This study seeks to advance emotion recognition algorithms to improve user experience in human-computer interaction (HCI), with vast potential applications in healthcare, virtual reality, customer service, and general HCI contexts. Additionally, the study will address ethical issues and privacy concerns to ensure responsible use of technology.

**Keywords**: Deep Learning, Multimodal Emotion Recognition, Facial Expression Analysis, Body Movement Analysis, Voice

Tone Analysis, Physiological Signal Processing, Data Fusion, Human-Computer Interaction

# Introduction

Emotion is a psychological state and an important part of human intelligence. It can be seen as the result of neural activity that happened in the brain. Emotion is crucial for human social activities, deeply involved in daily social interaction, learning, cognition, and creation. Thus, emotion recognition has always been a key branch of Artificial Intelligence. Emotion recognition aims to empower computers to understand people's complex emotions through a series of data analysis to improve the human-computer interaction experience a step further.

The study on emotion recognition started decades ago, Picard proposed “Affective Computing” in 1997 [[1]](#ref1), revealing the possibility that the computer can understand and react to human emotions and make better decisions by applying statistical models such as Hidden Markov Model. Computers could make better decisions based on emotional understanding. However, emotions are inherently complex and abstract, making it difficult to quantify them into data or variables that computers can process. As a result, researchers have struggled to establish a universal standard for emotion recognition experiments, which has led to the exploration of various methods and models.

Previous researchers have proposed different methods to quantify emotions, among which the discrete emotion model is commonly accepted in emotion recognition research. The discrete model describes emotions as the most basic discrete forms of joy, anger, sadness, and happiness. For example, psychologist Ekman divided emotions into six categories: happiness, sadness, surprise, anger, disgust, and fear [[2]](#ref2). This theory provided a framework that allows emotion recognition to be treated as a classification problem.

Early work on classifying emotions focused on a single data source which could be called unimodal analysis. Text data for one, obtained attention earlier than others. By applying NLP methods and Machine learning algorithms to written texts, researchers can build the classifier and determine the keywords that represent emotions better than other words [[3]](#ref3). Another one that has been widely studied is the audio data. The voice could deliver various information. For example, the speech would be faster, higher, and louder when people feel happy or angry and slower, deeper, and quieter when people feel sad or calm [[4]](#ref4). In addition, physiological signals, such as breathing, heart rate, and body temperature, have proven useful for emotion recognition [[5]](#ref5).

However, emotion is a complex state that can change rapidly and approaches that analyze a single data source have limitations. Firstly, unimodal analysis would lose accuracy when it meets complicated and dynamic contexts because it has a great possibility to overfit on partial information. For example, words will have the opposite meaning depending on the situation. “What a day” could mean the speaker had a great day or an exhausting day. Incorporating multimodal signals, such as voice tone, can help disambiguate such meanings. If the voice is fast and loud, we can infer a positive meaning; if it is slow and quiet, we infer a negative one.

Moreover, the unimodal analysis shows low robustness because a large individual bias may exist in certain data sources for different people. For example, the same gesture could mean very different things to people from different cultures. Instead of doing unimodal analysis, recent advancements in multimodal emotion identification systems have indicated the potential to tackle these difficulties. It is reported that the model performance could be improved when we combine multiple data sources such as audio, video, and text, in fact, studies have demonstrated that multimodal models can substantially outperform unimodal ones in emotion recognition tasks [[6]](#ref6).

Furthermore, because the emotion could be mixed and complicated, the traditional machine learning algorithms which need to extract features by hand may not be the best choice to make. Instead, deep neural network-based models would have a bigger possibility to handle the job because the combination of multiple layers of nonlinear transformation could extract the information that hides deeply in the data. Besides, numerous models have been developed for specifically processing certain modalities.

Therefore, multimodal emotion recognition using deep learning came out as a promising topic. Multimodal emotion recognition integrates multiple data sources to conduct a general analysis. It is expected to obtain more comprehensive and accurate emotion recognition results than unimodal analysis because multimodal analysis considers not only the information in single modality but also the interactive information between modalities. Through multimodal emotion recognition, computers would accurately identify the user's emotional state, greatly improving the human-computer interaction experience. This makes multimodal emotion recognition have a wide range of application backgrounds and promising commercial potential.

In the medical field, multimodal emotion recognition can help doctors pay attention to patients' emotional changes and provide patients with more accurate medical solutions and mental health counseling for human emotion is closely related to health. In the commercial field, multimodal emotion recognition can provide customers with fast and humanized services by building intelligent customer service. The AI customer service robots nowadays are mostly focused on single modality such as text. If we can apply multimodal emotion recognition, then robots would be closer to a real person that could react fast and accurately to customers’ needs.

In conclusion, multimodal emotion recognition is an indispensable step in the development of “Affective AI”. Therefore, holding in-depth research on multimodal emotion recognition is significant for promoting the development of artificial intelligence technology and even the whole society.

# Personal Research Background

As a graduate with distinction degree from the City University of Hong Kong, one of the leading institutions for research, I have received rigorous professional training in a range of subjects, including machine learning, deep learning, and data mining. Throughout my studies, I have gained extensive knowledge in the mathematical derivation and programming applications of algorithms. These experiences have equipped me with the skills necessary to apply and optimize deep learning algorithms for emotion recognition tasks.

In addition to my coursework, I have completed several projects in Natural Language Processing (NLP) and Computer Vision, where I developed a strong familiarity with text and image processing techniques—key components in leveraging multimodal data for emotion recognition. For instance, I worked on a project where I applied machine learning algorithms to analyze text data and use computer vision techniques to classify images. These projects have honed my ability to handle diverse data types, which will be instrumental for my future research in multimodal emotion recognition.

Moreover, I successfully completed my master’s research project on deep learning, receiving outstanding marks and positive feedback from my supervisors. In this project, I built a solid theoretical foundation in emerging deep learning models, gained hands-on experience in data processing, feature extraction, model building, and evaluation, and enhanced my ability to conduct thorough literature reviews and academic writing. These skills are crucial for pursuing doctoral research in the field of AI.

Currently, I am working as a Senior Algorithm Engineer at BYD, one of the world’s largest electric vehicle manufacturers, in the AI Lab department. My current project involves road surface recognition using multimodal data, including photographs and tire noise. This experience has further strengthened my skills in processing and analyzing data from various formats, which I believe will greatly benefit my doctoral research in multimodal emotion recognition.

More importantly, these research experiences have deepened my interest in advancing Artificial Intelligence research, as I firmly believe that AI has the potential to bring about significant advancements that will revolutionize our daily lives and pave the way for a brighter future.

# Literature Review

The biggest difference between multimodal analysis and unimodal analysis is that multimodal analysis mines not only the information in one modality but also the interactive information between multiple modalities by applying data fusion techniques. Generally speaking, there are three kinds of data fusion used by previous researchers.

The first one is called Early Fusion. Early fusion methods concatenate multimodal features into a single model input, rather than modeling the input features of each modality separately. This approach relies on a general model to learn the internal information of each modality and the interaction information between modalities without requiring specific model design. Hazarika et al. [[7]](#ref7) generates the final representation of an utterance by concatenating all three multimodal features including textual features, audio features, and visual features with linear combination and nonlinear activation function tanh. Morency et al. [[8]](#ref8) fuse multimodal features by taking average of each modality according to certain time clip and input the consternated feature into Hidden Markov Model. The advantage of the early fusion is that there is no need to design a specific model to make it fit each modality involved. The disadvantage is that it is easy to ignore the important and obscure information in the data if we don’t train models for each modality separately.

Late Fusion, in contrast, trains separate models for each modality and combines their predictions at the decision level. This fusion strategy applies different algorithms for different modalities and uses methods such as voting to make the final decision. For instance, Huang et al. [[9]](#ref9) combined facial expressions with EEG by running a neural network on facial expressions and an SVM on EEG, and then fused the decisions using sum and production rules. Abburi et al. [[10]](#ref10) trained GMM and neural networks on audio features and SVM on text, then averaged the results to make the final decision. While late fusion captures modality-specific information effectively, it lacks the ability to model interactions between modalities, which may limit its performance in complex scenarios.

The third one is called Multimodal Information Fusion. This method combines the logic of early fusion and late fusion. Same as the late fusion, it will train models separately on each modality. Then instead of using models for inference, they would be used for feature extraction to get latent feature representation. Then similar with early fusion, the feature representations of multiple modalities are consternated by certain methods like tanh activation. This is for capturing interactive information between different modalities. The multimodal information fusion takes both the advantages of early fusion and late fusion. It could capture the internal information like the late fusion and interactive information like the early fusion. It is expected that the multimodal information fusion would show better performance than the previous two methods.

For future research, we aim to explore multimodal information fusion as a potential approach for building more effective emotion recognition systems.

# Challenges

In this work, we suggest building a multimodal emotion detection system that can precisely identify human emotions in real time by integrating and analyzing input from several sources. This system seeks to deliver robust and precise emotion recognition, improving human-computer interaction across various applications such as virtual reality, healthcare, and customer service. It does this by utilizing powerful deep learning algorithms and data fusion methodologies. However, there are several challenges lying in the research.

1.Data Concealment: One major challenge is that individuals may attempt to conceal their emotions when they know they are being monitored. Modalities such as facial expressions, speech tone, and body gestures can be faked in these cases. To address this, incorporating harder-to-fake physiological signals, like EEG or body temperature, could improve the system's robustness.

2.Data Integration and Synchronization: A significant challenge lies in integrating and synchronizing data from modalities with different dimensions. For example, RGB images are typically represented as 4-dimensional data (Number\*Height\*Width\*Channels), while audio clips are converted into 2D data like Mel-Frequency Spectrum (MFC), which includes time and frequency components. Ensuring that data from these distinct sources can be effectively synchronized and integrated is a complex task.

3.Real-Time Processing: For real-time emotion recognition, the data processing pipeline must be optimized, and the model must be lightweight to ensure efficiency without significantly compromising accuracy.

4.Additional Challenges: Other challenges include denoising raw data, preventing overfitting (as deep neural networks require large datasets), and deploying the model across different platforms.

1. **Research Questions**

There are some common questions brought by previous research that aimed to be solved in this research:

1. How can multimodal data be effectively integrated to improve the accuracy and robustness of emotion recognition system?
2. What deep learning algorithms are most effective for processing and analyzing multimodal emotion data in real-time?
3. How can real-time emotion recognition system be designed to operate efficiently on various hardware platforms, including mobile and wearable devices?
4. What methods can be employed to ensure the scalability and flexibility of multimodal emotion recognition systems for different applications?

This research will build the emotion recognition system based on seeking the solution for these questions.

## Methodology

**6.1 Literature Review**

To gain a thorough understanding of the state-of-the-art in multimodal emotion recognition, spot any shortcomings, and lay the groundwork for the creation of fresh strategies, the fastest way is to review the literature in-depth on the use of speech tonality, body language, facial expressions, and physiological cues to identify emotions. Examine previous research on real-time processing strategies and data fusion approaches. Determine the shortcomings and difficulties of the current approaches and identify possible areas for development. Grabbing the inspiring idea in the previous research and creating the different methodology based on it. Reviewing the latest papers has always been an efficient way to catch up with the research trend.

**6.2 Design and Development of Algorithms**

To create innovative, real-time, multimodal emotion recognition algorithms that can successfully integrate and process many data streams while addressing the difficulties involved, it is essential to develop algorithms that make use of adaptive mechanisms, online learning, and effective data processing methods. Thanks to the earlier researchers, there are a large amount of deep learning algorithms which are designed specifically for different tasks.

In this research, we will start by applying suitable algorithms for processing different modalities and make the adaptation to fit with our collected data for this is a highly effective method comparing to develop a new algorithm from scratch. For example, D. Snyder et al. proposed X-Vectors for enhancing deep learning models on speaker recognition tasks [[11]](#ref11), especially for processing audio modality. Convolutional Neural Network (CNN) has been proved to be powerful on processing images leveraging the convolution and subsampled tactics. Long Short-Term Memory (LSTM) is a variant of Recurrent Neural Network (RNN) which is good at handling the long text by adding memory cells and gate mechanism [[12]](#ref12). LSTM can learn the context by remembering previous input and pass it to the next layer. Such features are considered valuable for emotion recognition especially when the input is the combination of multi-modalities.

The general models do not always show ideal performance on the practical datasets. Thus, after building the models, we will need to optimize the models by modifying operators to improve the classification ability and inference time. To establish theoretical soundness and codify algorithmic procedures, use mathematical modeling. Use libraries like Torch, Librosa, Apache Flink, and MOA (Massive Online Analysis) to implement algorithms using Python or Java. Writing code that is modular to make optimization, debugging, and testing easier.

**6.3 Gathering and Preparing Data**

To train a functional and robust multimodal emotion recognition model, it is important to compile high-quality, multimodal datasets and perform preprocessing on them to guarantee consistency and accuracy in the identification of emotions. Gathering information from a variety of devices, such as microphones, cameras, motion detectors, and physiological signal monitors. To collect extensive and varied multimodal emotion data, make use of both newly created experiments and the public datasets that are already available.

The Multimodal Corpus of Sentiment Intensity (CMU-MOSI) dataset is a collection of 2199 opinion video clips. Each opinion video is annotated with sentiment in the range [-3,3]. The Multi-Modal Movie Opinion (ICT-MMMO) dataset consists of online social review videos that encompass a strong diversity in how people express opinions. The dataset contains 343 multimodal review videos annotated at the video level for sentiment. The Multimodal Opinion Utterances Dataset (MOUD) Dataset consists of 80 product review videos in Spanish. Videos were found on the YouTube search page using the following keywords (translated to English): my favorite products, non-recommended products, my favorite perfumes, recommended movies), non-recommended movies, recommended books and non-recommended books. Each video consists of multiple segments labeled to display positive, negative or neutral sentiment.

These public datasets have been proved to be useful with numerous experiments and valuable samples to test trained models. But for raw data collected from hardware in real-time, noise and error are unavoidable. To manage missing numbers and eliminate noise, clean up and preprocess the data, a processing pipeline needs to be built for washing data and integrate data from several modalities to guarantee logical analysis. Furthermore, data should be standardized and normalized to enable effective processing and analysis.

**6.4 Investigative Assessment**

To give a comprehensive evaluation of the model, it is critical to verify the efficacy and resilience of the created algorithms and systems, assess their performance in a range of scenarios. Create trials to evaluate the algorithms in various settings and use cases, including virtual reality, healthcare, and customer support. Assess the system, use performance measures including latency, accuracy, precision, recall, F1-score, and precision. Perform statistical analysis and cross-validation to evaluate the validity and dependability of the findings.

# Conclusion

There is still a gap in the research of multimodal emotion recognition. This proposal is to demonstrate the attempt to fulfil the voids with my knowledge and efforts. I firmly believe that conducting such promising research under the supervision and instruction of an expert would be a great help and opportunity to reach a higher level of academic achievement.

# References

[1] Picard, Rosalind W.. “Affective Computing.” (1997).

[2] Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3–4), 169–200. <https://doi.org/10.1080/02699939208411068>

[3] Minu, R I & R.Ezhilarasi,. (2012). Automatic Emotion Recognition and Classification. Procedia Engineering. 38. 10.1016/j.proeng.2012.06.004.

[4] Johnstone, T. (2017, August 15). The effect of emotion on voice production and speech acoustics. <https://doi.org/10.31237/osf.io/qd6hz>

[5] S. Jerritta, M. Murugappan, R. Nagarajan and K. Wan, "Physiological signals based human emotion Recognition: a review," *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, Penang, Malaysia, 2011, pp. 410-415, doi: 10.1109/CSPA.2011.5759912.

[6] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98-125. <https://doi.org/10.1016/j.inffus.2017.02.003>

[7] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. [ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection](https://aclanthology.org/D18-1280). In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2594–2604, Brussels, Belgium. Association for Computational Linguistics.

[8] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: harvesting opinions from the web. In Proceedings of the 13th international conference on multimodal interfaces (ICMI '11). Association for Computing Machinery, New York, NY, USA, 169–176. <https://doi.org/10.1145/2070481.2070509>

[9] Huang, Y., Yang, J., Liao, P., & Pan, J. (2016). Fusion of Facial Expressions and EEG for Multimodal Emotion Recognition. Computational Intelligence and Neuroscience, 2017(1), 2107451. <https://doi.org/10.1155/2017/2107451>

[10] Abburi, Harika & Prasath, R.Rajendra & Shrivastava, Manish & Gangashetty, Suryakanth. (2017). Multimodal Sentiment Analysis Using Deep Neural Networks. 58-65. 10.1007/978-3-319-58130-9\_6.

[11] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 5329-5333, doi: 10.1109/ICASSP.2018.8461375.

[12]  S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computation, 9(8):1735–1780, 1997.