IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, November 2024

Survey on Multimodal Emotion Detection

Kurhe Prajakta[1], Sanap Ashwini[2], Gade Gayatri[3], Batwal Ankita[4], Deokar Varsha[5]
Assistant Professor, Department of Computer Engineering[1]
Students, Department of Computer Engineering[2,3,4,5]
SND College of Engineering Research Center, Yeola, India
Savitribai Phule Pune University, Yeola, India
kurheprajakta.coe@gmail.com[1], sanapashwini2003@gmail.com[2], gadegaytri86@gmail.com[3],
ankitabatwal99@gmail.com[4], deokarvarsha@gmail.com[5]

Abstract: *As we know, many people suffer from unhappiness in today's world, so there is a requirement for a system that recommends music based on human emotions such as anger, sadness, happiness, etc. In this paper, we propose a system named Multimodal Emotion Detection which helps people by suggesting music and movies based on their emotions. We implemented this system using the FER-2013 dataset, which contains 35,887 images of different emotions such as sadness, happiness, and anger. We used Python's TensorFlow framework and Haar Cascade algorithms to recommend movies and music based on human facial emotions. The main concernin existing recommendation systems is manual sorting. To avoid this, we propose this model, which automatically plays music and movies without requiring much browsing time*

Keywords: Image processing, Facial recognition, Movie Recommendation, Song Recommendation.2 [1-2]

I. INTRODUCTION

Music and movies are well-known for their ability to alter human emotions, enhancing mood and reducing depression. The primary objective of this project is to develop a model that can detect emotions based on facial expressions. To achieve this, a database of songs and movies is involved, which recommends content based on the detected expressions. The project also aims to create an easy-to-use interface so that users can easily access the website. The interface is built using HTML, CSS, and Flask. The key focus is to evaluate and test the accuracy and effectiveness of the model in detecting emotions and providing recommendations. Since the abundance of available music and movie options can create confusion, this model helps users seamlessly receive personalized recommendations.

The development of an intuitive interface is critical for providing real-time recom- mendations. A Convolutional Neural Network (CNN) is employed to detect and analyze emotions. It consists of an input layer, convolutional layers, dense layers, and an output layer. The CNN extracts features from images to determine specific facial expressions.

To accurately detect emotions, faces must be identified within an image before the model can classify expressions. The ultimate goal of this project is to build a framework that helps users discover an ideal choice of music or movie based on their emotional state. This system aims to establish correlations and similarities between different songs and movies, constructing a recommendation system that suggests new content accordingly.

II. BACKGROUND AND KEY CONCEPT

Multimodal Emotion Detection

Multimodal emotion detection refers to the process of identifying and interpreting human emotions by combining multiple types of data inputs, typically visual (facial expressions), audio (speech, voice tone), and textual (spoken or written words). The goal of multimodal emotion detection is to create a more accurate and holistic understanding of a user's emotional state by integrating these different sources of information.

Emotion detection has gained increasing relevance with the rise of human-computer interaction (HCI) applications, where systems need to understand and respond to users' emotions to enhance user experience. This capability is particularly significant in fields such as personalized content recommendation, healthcare, education, and virtual assistants.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, November 2024

Key Elements in Multimodal Emotion Detection

- **Facial Expression Analysis:** One of the most commonly used methods for detect- ing emotions. Facial expressions are analyzed through computer vision techniques, such as Convolutional Neural Networks (CNNs), which process facial landmarks (e.g., eyes, mouth, eyebrows) to classify emotions like happiness, sadness, anger, or surprise.
- **Speech and Voice Analysis:** Audio features such as tone, pitch, and intensity are analyzed to identify the emotional state of a speaker. This modality is often paired with facial expression analysis to provide a more comprehensive understanding of emotions.
- **Textual Data Analysis:** In some cases, text from social media, reviews, or chat- bots is analyzed using Natural Language Processing (NLP) techniques to detect sentiment and emotions.

Facial Expression-Based Emotion Detection

Facial expressions are a primary channel through which humans express their emotions. Advanced computer vision techniques are used to capture and analyze these expressions to detect underlying emotions. The process typically involves the following steps: 1. Face Detection: Identifying the presence and location of faces in images or videos us- ing algorithms like the Viola-Jones detector or modern deep learning-based techniques such as Haar cascades. 2. Feature Extraction: Detecting facial landmarks (such as eyes, mouth, and eyebrows) or using deep learning models (like CNNs) to extract key fea- tures from the face. 3. Emotion Classification: Assigning a category (e.g., happiness, anger, fear) to the detected emotion using machine learning classifiers or neural networks. Popular models include CNNs, Recurrent Neural Networks (RNNs), and Support Vector Machines (SVMs). Facial expression analysis is considered reliable for real-time emotion detection, particularly in environments where visual data can be easily captured, such as smartphones, webcams, or in-vehicle cameras.

Recommendation Systems

Recommendation systems use algorithms and data-driven techniques to suggest person- alized content (e.g., movies, music) to users. In traditional recommendation systems, user preferences are derived from historical data, such as past interactions (e.g., movies watched, songs liked) or demographic information. However, integrating emotion detec- tion with recommendation systems enhances personalization by adapting recommenda- tions to users' emotional states in real-time.

Key Approaches in Recommendation Systems

- **Collaborative Filtering**: Makes recommendations by identifying users with sim- ilar preferences and suggesting content that those users have enjoyed.
- **Content-Based Filtering**: Recommends content similar to items a user has pre- viously liked, based on features of the items (e.g., genre, director, actors).
- **Hybrid Systems**: Combine both collaborative and content-based approaches to provide more accurate recommendations.

In the context of emotion-based recommendations, the system dynamically adapts suggestions based on the user's detected emotional state. For example, if the system detects that a user is feeling sad, it might recommend uplifting music or feel-good movies to improve their mood.

III. LITERATURE REVIEW

Multimodal emotion detection and its application in personalized recommendation sys- tems, such as movie and music recommendations, have gained significant attention in recent years. This section reviews key approaches and technologies used in emotion detection through facial expressions and the integration of these techniques into recommendation systems.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, November 2024

Emotion Detection through Facial Expressions

Facial expression analysis has been a foundational technique in emotion detection due to its non-invasive and widely applicable nature. Early methods relied on handcrafted features and traditional machine learning models, while recent advancements have shifted towards deep learning-based approaches.

Traditional Methods

Before the advent of deep learning, emotion recognition from facial expressions was based on feature extraction techniques, such as:

- **Facial Action Coding System (FACS):** This system breaks down facial expres- sions into a set of Action Units (AUs) based on muscle movements, which can be mapped to specific emotions. Early machine learning models, like Support Vector Machines (SVMs), were used to classify emotions based on the extracted AUs.
- **Gabor Filters:** These filters were employed to capture facial texture features, which were then fed into classifiers such as k-Nearest Neighbors (k-NN) or Decision Trees to identify emotions.

Deep Learning Approaches

With the rise of deep learning, Convolutional Neural Networks (CNNs) have become the dominant method for facial emotion recognition. CNNs automatically learn hierarchical features from facial images, eliminating the need for handcrafted feature extraction.

- **Convolutional Neural Networks (CNNs):** CNNs have shown remarkable suc- cess in emotion detection by processing raw facial images and learning emotion- relevant features through layers of convolutions and pooling.
- **Recurrent Neural Networks (RNNs):** When dealing with video data, RNNs, particularly Long Short-Term Memory (LSTM) networks, are employed to cap- ture the temporal dynamics of facial expressions. These networks help model the sequential nature of emotions, which unfold over time rather than being static.
- **Generative Adversarial Networks (GANs):** GANs have been explored in syn- thesizing emotional expressions or enhancing datasets by generating more varied facial expressions. This helps improve the robustness of emotion detection models, particularly in dealing with data scarcity.

Multimodal Fusion for Emotion Detection

Facial expressions are often combined with other modalities like audio and text to improve the robustness of emotion detection. Common multimodal fusion techniques include:

- **Feature-Level Fusion:** Combines raw features from different modalities (e.g., facial landmarks, voice pitch) before feeding them into a classifier.
- **Decision-Level Fusion:** Processes each modality separately and then combines the predictions from each modality to make the final emotion classification. En- semble methods like boosting and bagging are often used here.

This fusion improves the accuracy of emotion detection, as each modality can com- pensate for the shortcomings of the others (e.g., if facial expressions are ambiguous, voice analysis can provide clearer emotional cues).

Emotion-Based Recommendation Systems

Integrating emotion detection into recommendation systems represents a shift from static, historical data-based recommendations to dynamic, real-time personalization. The goal is to enhance user experience by recommending content (music, movies) that matches or modulates the user's emotional state.

Traditional Recommendation Systems

Traditional recommendation systems rely on two main approaches:

- **Collaborative Filtering:** Uses past behaviors and preferences of similar users to recommend content. This approach faces limitations when emotional context is required, as it only looks at historical data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, November 2024

- **Content-Based Filtering:** Recommends items similar to those the user has pre- viously interacted with, based on item features (e.g., genre, mood). However, it doesn't take the user's real-time emotional state into account.

Emotion-Driven Recommendation

Emotion-driven recommendation systems enhance traditional methods by incorporating emotion detection as an input. Techniques include:

- **Emotion-Adaptive Collaborative Filtering:** Modifies collaborative filtering algorithms to factor in real-time emotional data. For instance, if a user feels sad, the system may recommend comforting music or movies that have been popular with other users in a similar emotional state.
- **Emotion-Based Content Filtering:** Uses detected emotions to filter content based on emotional tone. For instance, if a user appears stressed, the system might recommend soothing or uplifting content to improve the user's mood.

Techniques for Emotion-Based Recommendations

- **Hybrid Models:** Combine traditional recommendation algorithms with emotion detection systems to offer more nuanced and personalized recommendations. For example, collaborative filtering can be paired with real-time facial expression anal- ysis to modify suggestions based on the user's detected emotional state.
- **Real-Time Personalization:** Deep learning techniques such as Reinforcement Learning have been explored for real-time personalization. These systems learn from real-time user interactions and emotional feedback to continuously adjust the recommended content, offering a more responsive and personalized experience.

IV. SYSTEM ARCHITECTURE

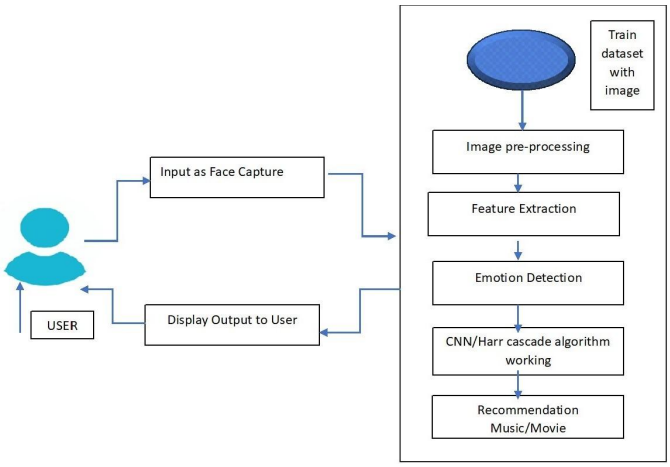

Figure 1: Architecture The entire system is structured into five key steps:

Image Acquisition

The first step in the image processing workflow is to gather user images from a camera source, ensuring they are in the .jpg format. article

Pre-processing

This stage is crucial for removing unnecessary details from the captured images and standardizing them. Images are converted from RGB to grayscale. Key facial areas such as the eyes, nose, and mouth are identified using the Haar Cascade algorithm.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, November 2024

Feature Extraction

In this step, important facial features are extracted and represented as vectors during both the training and testing phases. The main features analyzed include the mouth, eyes, nose, and forehead, as these areas reflect the most expressive emotions. Principal Component Analysis (PCA) is employed to extract and represent these features.

Expression Recognition

A Euclidean distance classifier is used to identify a person's expression. It finds the closest match between test images and training images, assigning an expression label (happy, sad, fear, surprise, anger, disgust, or neutral) based on the smallest distance from the average image.

Music/Movie Recommendation

The final step involves recommending music and movies based on the user's identified emotional state. After facial expression classification using a CNN algorithm, a curated list of songs or movies corresponding to the detected emotion is presented for selection.

These are grouped by emotional categories, allowing users to choose based on their current mood.

V. EVALUATION METHODOLOGY

Datasets

Selecting appropriate datasets is crucial for evaluating the performance of multimodal emotion detection systems. Commonly used datasets in the field include:

- **FER+:** A widely-used dataset containing labeled facial expressions. It improves on the original FER dataset by providing more consistent and accurate annotations.
- **AffectNet:** A large facial expression dataset containing images labeled with emo- tions such as happiness, sadness, anger, surprise, and more.

These datasets provide a diverse set of emotions and multimodal data, making them suitable for testing and training your model.

Evaluation Metrics

To assess the effectiveness of the system, you can use several metrics that measure how well the system predicts emotions. Common metrics include:

- **Accuracy:** The percentage of correctly classified emotions out of all predictions.
- **Precision:** The ratio of true positive predictions to the sum of true positives and false positives. It shows how many of the predicted emotions were correct.
- **Recall (Sensitivity):** The ratio of true positives to the sum of true positives and false negatives. It indicates how many of the actual emotions were correctly identified.
- **F1-Score:** The harmonic mean of precision and recall. It provides a balanced measure, especially useful when the class distribution is uneven.
- **Confusion Matrix:** A table that shows the number of correct and incorrect pre- dictions for each emotion. It provides a deeper understanding of which emotions are harder to classify.

Cross-Validation

To ensure that the model generalizes well to unseen data, you can perform k-fold cross- validation. In this method:

- The dataset is divided into k subsets (folds).
- The model is trained on $k - 1$ folds and tested on the remaining fold.
- This process is repeated k times, and the results are averaged to provide a reliable estimate of the model's performance.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, November 2024

Comparison with Baselines

You should compare your proposed model's performance against baseline models to demonstrate improvements. Common baselines include:

- Single-modality models (e.g., facial expressions or speech only).
- Traditional Machine Learning Methods (e.g., Support Vector Machines, Decision Trees).
- Existing multimodal systems from recent literature.

Ablation Study

An ablation study is conducted to understand the contribution of different components of your system. For example:

- Test the system with only facial expression data.
- Test with only speech data.
- Test with both modalities together to show how combining them improves perfor- mance.

Computational Efficiency

It's important to assess the computational cost of your system, especially for real-time applications. You may include:

- **Inference Time:** The time it takes for the system to detect emotions and recom- mend content.
- **Resource Usage:** Evaluate the CPU/GPU memory usage and the complexity of the model in terms of the number of parameters.

VI. CHALLENGES IN MULTIMODAL EMOTION DETECTION

Data Quality and Diversity

- **Limited Data Availability:** High-quality, labeled datasets for multimodal emo- tion detection are often scarce. The lack of comprehensive datasets that capture a wide range of emotions across diverse demographics can hinder the training of effective models.
- **Data Imbalance:** Emotion datasets may have a disproportionate number of sam- ples for certain emotions, leading to biased models that perform well on frequently represented emotions while struggling with less common ones.

Real-Time Processing

- **Latency:** Achieving low-latency processing for real-time emotion detection and recommendations can be difficult, especially when dealing with simultaneous pro- cessing of video and audio data. Ensuring that the system can provide immediate feedback is crucial for user experience.
- **Resource Constraints:** Real-time applications may have limitations on hard- ware resources (CPU/GPU), necessitating the optimization of models to balance performance and resource consumption effectively.

VII. FUTURE DIRECTION

Expanding Dataset Diversity

Encourage the creation of larger and more diverse datasets that include a variety of emotional expressions across different cultures and demographics. This will enhance the model's ability to generalize.

Addressing Ethical Concerns

Develop privacy-preserving techniques, such as differential privacy and federated learning, to ensure user data is protected during emotion detection processes.

Real-World Application Enhancements

Enhance personalized recommendation systems by considering not only detected emotions but also contextual information (user history and preferences) to tailor suggestions more effectively.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, November 2024

Advancements in Model Interpretability

Invest in research on explainable AI to improve transparency in how emotion detection systems make decisions, fostering user trust and understanding.

VIII. CONCLUSION

This paper discusses the importance of using facial expressions to detect emotions, partic- ularly in creating personalized recommendations for music and movies. While there have been significant advancements in this field, challenges remain, such as the need for diverse and high-quality data and concerns about privacy and bias. Moving forward, research should focus on improving the accuracy of facial expression recognition and making these systems easier for users to understand and trust. By effectively using facial expressions to detect emotions, we can develop technologies that are more engaging and responsive, ultimately enhancing how users interact with digital content.

REFERENCES

[1]. Li, Huihui. Research on facial expression recognition based on cognitive machine learning [D]. Guangzhou: South China University of Technology, 2019.

[2]. Zou, Jianchen, and Deng, Hao. An automatic facial expression recognition method based on convolutional neural network [J]. Journal of North China University of Technology, 2019, 31(5): 51-56.

[3]. Yao, L. S., Xu, G. M., and Zhap, F. Facial Expression Recognition Based on CNN Local Feature Fusion [J]. Laser and Optoelectronics Progress, 2020, 57(03): 032501.

[4]. Chauhan, Shavak, Mangrola, Rajdeep, and Viji, D. Analysis of Intelligent Movie Recommender System from Facial Expression. In: 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021.

[5]. Sudhakaran, Pradeep, Kizhakkevellat Nair, Pranav, and Suraj, Anand. Music Rec- ommendation Using Emotion Recognition. In: 2022 IEEE 2nd Mysore Subsection International Conference, 2022