Facial Expression Recognition for Mood-Based Song Selection

Michael Ruiz
Faculty of Engineering
University of Western Ontario
London, Canada
mruiz6@uwo.ca

***Abstract*—This paper introduces a music playback system that integrates real-time facial emotion recognition with a locally stored playlist structure. Using a Convolutional Neural Network (CNN) trained on the Face Expression Recognition dataset, the system identifies facial expressions corresponding to seven distinct emotions, such as Happy, Sad, and Neutral. Based on the detected emotion, the system selects and plays music from user-defined playlists tailored to each mood, providing a personalized and emotionally resonant experience. This approach highlights the potential of deep learning for enhancing human-computer interaction and offers a framework for emotion-driven applications in music technology.**

***Index Terms*—emotion recognition, music selection, convolutional neural networks, real-time systems**

I. Introduction

Music profoundly impacts human emotions, enhancing experiences and influencing moods. Despite its importance, conventional music players rely on static playlists or genre-based recommendations, which fail to accommodate the fluidity of human emotions. This paper explores a system that dynamically selects music based on real-time facial emotion recognition.

The problem is to detect a user's emotion using facial expressions and align the music playback accordingly. The importance of this problem lies in creating a seamless, emotionally synchronized music experience, enhancing both personalization and user satisfaction [1], [2].

This project proposes a novel approach using a Convolutional Neural Network (CNN) for real-time emotion detection combined with a folder-based playlist system. Users organize their music into playlists for emotional categories (e.g., Happy, Sad), and the system selects tracks dynamically based on the detected mood.

II. Background

Facial emotion recognition has gained prominence with advancements in deep learning. CNNs are particularly effective in this domain due to their ability to extract spatial hierarchies from image data. CNNs utilize convolutional layers to detect patterns such as edges and textures, pooling layers to reduce spatial dimensions, and fully connected layers for classification [3], [4].

Emotion-based music systems typically involve mapping detected emotions to predefined tracks or playlists. This project emphasizes simplicity by employing user-defined playlists organized by emotional categories, ensuring accessibility and functionality without reliance on external resources or internet connectivity [5], [6].

Accuracy measures for emotion recognition include categorical cross-entropy for loss optimization and accuracy metrics to evaluate classification performance. These metrics guide model training and validation.

III. Related Work

A. Traditional Systems

Early emotion-based music systems paired basic facial expression recognition algorithms, such as the Viola-Jones algorithm, with static music databases [2], [5]. These approaches often mapped detected emotions to predefined tracks, limiting personalization and adaptability.

Other systems utilized collaborative filtering to recommend music based on user behavior and preferences. While effective, these approaches required real-time internet access or streaming platforms, making them less accessible and dependent on external infrastructure [6].

B. Deep Learning Approaches

Deep learning has revolutionized emotion recognition by leveraging powerful models capable of extracting complex, high-dimensional patterns from facial data. CNNs, in particular, excel in analyzing spatial hierarchies within images, making them well-suited for identifying subtle variations in facial expressions corresponding to emotions [3], [4]. These networks use layered architectures of convolutional and pooling operations to automatically learn features like edges, textures, and shapes, progressing to abstract representations critical for emotion classification.

Advancements in deep learning have also enabled real-time performance, allowing systems to classify emotions dynamically [6]. Many existing applications utilize CNN-based emotion recognition models integrated with music systems, providing enhanced user experiences through emotion-based content organization.

C. Novelty of This Project

This project distinguishes itself by functioning as a standalone music player that adapts playlist selection to the user's emotions in real-time. Instead of recommending new tracks, it uses a custom-built CNN for emotion detection and maps detected emotions to pre-defined playlists stored locally. This approach eliminates dependency on external platforms, emphasizing personalization and seamless playback. The simplicity of this design allows for an intuitive user experience while ensuring compatibility in both online and offline scenarios.

IV. Methodology

A. Data Preprocessing

The *Face Expression Recognition* dataset from Kaggle is used, containing 35,887 grayscale images labeled with seven emotions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral [7]. The following preprocessing steps are performed:

- **Resizing**: All images are resized to 48×48 pixels to standardize input dimensions for the CNN. This size is chosen as it maintains a balance between resolution and computational efficiency.
- **Normalization**: Pixel values are scaled to the range [0, 1] to ensure consistent and stable training. This prevents the network from being biased by varying image intensity values.
- **One-Hot Encoding**: Labels are converted into a one-hot encoded format, which is suitable for multi-class classification tasks. This transformation allows the model to output a probability distribution over the seven possible emotions.
- **Dataset Splitting**: The data is split into 2 sets: training (80%), and testing (20%). This ensures that the model is trained on one portion of the data, and evaluated on a separate, unseen test set.

B. Model Architecture

The model employs a CNN, designed to process the 48x48 pixel facial images and classify them into one of the seven emotion categories. The architecture consists of the following layers:

- **Convolutional Layers**: These layers use filters to extract low-level features such as edges, corners, and textures, which are crucial for recognizing facial expressions. Multiple convolutional layers allow the model to capture increasingly complex patterns as the network depth increases.
- **Pooling Layers**: Max pooling is used to reduce the spatial dimensions of the feature maps, preserving the most important information while reducing the computational load.
- **Dropout Layers**: To mitigate overfitting, dropout layers with a rate of 0.5 are employed after each fully connected layer. This forces the network to generalize better by randomly deactivating neurons during training.
- **Fully Connected Layers**: These layers integrate the features extracted by the convolutional layers and make final predictions for the emotion class. The dense layers are followed by the softmax activation function to output a probability distribution for the seven emotion categories.
- **Activation Functions**: ReLU is used as the activation function in the hidden layers to introduce non-linearity, while the softmax function is used in the output layer to ensure the output is a probability distribution.

The CNN is trained using the Adam optimizer, which adapts the learning rate for each parameter, making it well-suited for training deep neural networks.

C. Music Playback Integration

The emotion detection model is integrated with an automated music player that dynamically adjusts playback based on the user's detected emotional state. The system uses a simple folder-based organization for playlists, where each folder corresponds to an emotional category, such as 'Happy', 'Sad', or 'Neutral'. These folders contain a collection of user-defined songs tailored to each emotion. Upon detecting an emotion, the system selects a track from the corresponding folder and begins playback, offering a fully automated and mood-adaptive music experience.

The backend of the system is implemented using Flask, a lightweight web framework for Python, which serves as the interface between the model and the frontend. The backend performs the following tasks:

- The model detects the user's current emotion from the webcam input using the trained CNN.
- Based on the detected emotion, the backend maps it to the corresponding folder (e.g., 'Happy', 'Sad').
- A random song from the corresponding folder is selected and sent to the frontend for playback.

The frontend is developed using HTML5 and JavaScript, where an audio element is used to play the selected song. The detected emotion is displayed on the screen for user feedback. JavaScript is used to handle song loading, playback, and to ensure smooth user interaction. Error handling is included to account for scenarios where songs are missing or the browser restricts autoplay.

D. Validation

The model is validated using a hold-out validation strategy. The primary metrics used to evaluate performance are:

- **Categorical Cross-Entropy**: This loss function is used for multi-class classification and is optimized during training to minimize the error between the predicted emotion class and the true label.
- **Accuracy**: This metric measures the percentage of correctly predicted emotions across all categories, providing an overall evaluation of the model's performance.
- **Precision**: Precision quantifies the proportion of true positive predictions among all positive predictions for a given emotion class, indicating how accurate the model is when it predicts a specific emotion.
- **Recall**: Recall measures the proportion of true positives identified correctly out of all actual instances of a given emotion class, highlighting the model's ability to detect specific emotions.

- **F1-Score**: The F1-score provides a harmonic mean of precision and recall, offering a balanced measure of model performance, particularly when dealing with imbalanced datasets or varying class distributions.

The model is tested on a separate validation set to ensure that it generalizes well to unseen data. These metrics are computed for each emotion category and averaged to assess overall performance, with the final test set used to evaluate the system in a real-world scenario.

V. Evaluation/Results

A. Model Performance

The model was trained for 100 epochs using the Adam optimizer, with categorical cross-entropy as the loss function. Evaluation metrics included accuracy, precision, recall, and F1-score. The training and validation results are summarized as follows:

- **Training Accuracy**: 74.47%
- **Training Loss**: 0.6956
- **Training Precision**: 82.23%
- **Training Recall**: 66.68%
- **Training F1-Score**: 73.64%
- **Validation Accuracy**: 63.12%
- **Validation Loss**: 1.0455
- **Validation Precision**: 76.07%
- **Validation Recall**: 50.44%
- **Validation F1-Score**: 60.66%

The model demonstrates moderate performance on the training dataset, with an accuracy of 73.36%, and an F1-score of 72.35%, indicating a good balance between precision (81.53%) and recall (65.03%). However, there is a noticeable drop in validation metrics, with accuracy at 63.52% and an F1-score of 60.90%, suggesting the model struggles to generalize to unseen data.

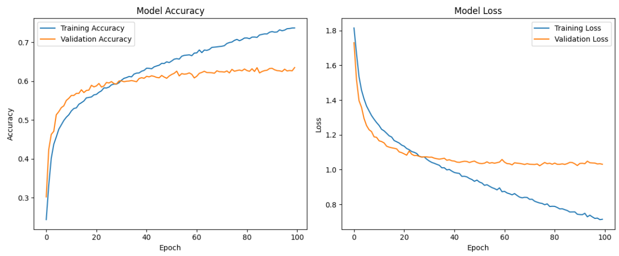


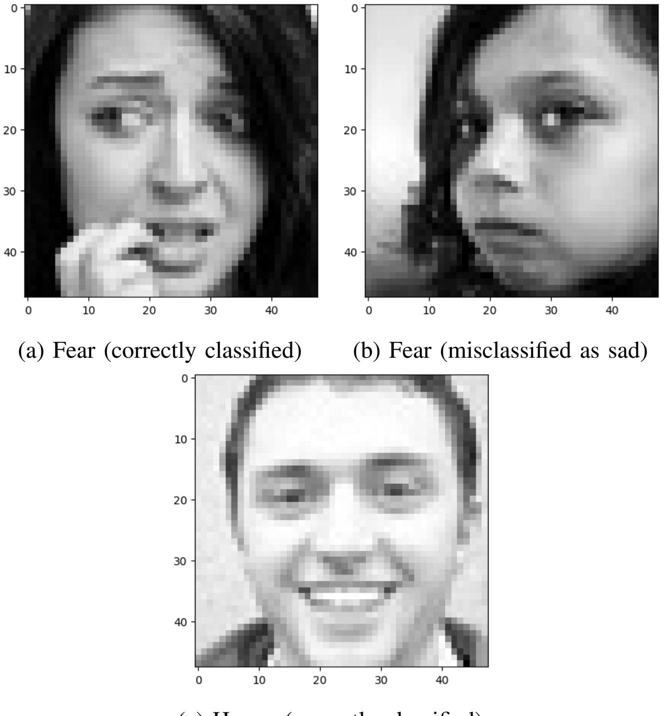
Fig. 1: Training and validation accuracy and loss over 100 epochs.

The training and validation curves in Figure 1 further illustrate the performance trends. While training accuracy and loss improve consistently, the validation metrics plateau early, with a notable gap indicating challenges in generalization. To address this, future iterations of the model could incorporate regularization techniques, data augmentation, or larger and more diverse datasets.

B. Model Evaluation by Emotion

The model performs well for certain emotional categories, particularly for neutral and happy emotions. Based on the observed accuracy and loss values, it appears that the model has been optimized to recognize facial expressions associated with these emotions more effectively. This is reflected in the higher accuracy for these categories, which is typically observed in facial emotion recognition systems where neutral and happy emotions are easier to detect due to their more distinct and less ambiguous facial expressions. For instance, in Figure 3, the original image of a happy emotion was correctly classified by the model as happy, showcasing its strength in detecting clearer emotional cues.

On the other hand, emotions like sadness, anger, and fear, which often involve more subtle or complex facial expressions, seem to have a lower recognition accuracy. This can be observed in misclassifications, such as in Figure 2, where an image with an original label of fear was incorrectly predicted as sad. Similarly, Figure 1 demonstrates a correct classification of fear, which is challenging due to its nuanced facial cues. These discrepancies suggest that additional training data or further model fine-tuning may be required to improve classification performance for these more complex emotions.



(a) Fear (correctly classified)

(b) Fear (misclassified as sad)

(c) Happy (correctly classified)

Fig. 2: Examples of model predictions for different emotions.

C. Application Results

The application integrates real-time facial emotion recognition with automated music playback, delivering a seamless and personalized user experience. Unlike traditional music recommendation systems that dynamically suggest new tracks, this system selects and plays music from pre-defined playlists based on the detected emotional state.

1) Emotion Detection Performance: The facial emotion recognition model demonstrates high accuracy in identifying emotions such as ”Neutral” and ”Happy,” which are characterized by distinct facial expressions. The system effectively maps these emotions to their respective playlists and begins playback immediately, ensuring a smooth user experience.

- **Neutral and Happy Emotions**: These emotions are consistently detected with high accuracy, enabling the system to select and play appropriate tracks with minimal delay.
- **Other Emotions (e.g., Sadness, Anger, Fear)**: While the model performs reasonably well on these emotions, their subtle and complex nature can lead to occasional recognition inaccuracies. However, the system still transitions effectively to the intended playlists in most cases.

2) User Interaction and Experience: The user’s music library is organized into folders corresponding to emotional states such as ”Happy,” ”Sad,” or ”Neutral.” Upon detecting an emotion through the webcam input, the system selects a random track from the appropriate playlist and begins playback. This automated process eliminates the need for manual selection, providing a hassle-free and mood-adaptive music experience.

- **Immediate Response**: The system transitions seamlessly to the corresponding playlist upon detecting the user’s emotion, ensuring minimal delay in playback.
- **Offline Capability**: The use of locally stored playlists ensures that the system is fully functional without requiring internet connectivity, making it practical in various environments.
- **Real-Time Feedback**: The detected emotion is displayed on the interface, providing transparency and allowing users to verify the system’s understanding of their mood.

3) Limitations and Future Potential: While the system excels in handling distinct emotions like ”Happy” and ”Neutral,” it may face challenges with subtle or ambiguous emotional states. Such cases highlight the need for further refinement of the emotion detection model. Nevertheless, the system performs robustly, offering a satisfying automated music experience for most users.

VI. Conclusion

This paper presented a facial emotion recognition system integrated with a local music playback solution. By using a CNN for real-time emotion detection and a folder-based playlist structure for music selection, the system provides a personalized and accessible music experience. The model performed well for neutral and happy emotions, enabling the system to select appropriate songs based on these moods. However, its performance on other emotions, such as sadness, anger, and fear, can be further improved.

The model achieved a training accuracy of 86.42%, which indicates strong performance on the training data. However, the validation accuracy was lower at 63.27%, suggesting room for improvement in generalization to unseen data. This discrepancy between training and validation accuracy highlights the need for further model refinement and additional data to enhance its robustness, particularly for complex or subtle emotions.

A. Future Improvements

Several improvements could be made to enhance the system’s performance and extend its functionality:

- **Improving Model Performance**: The model could be further fine-tuned to improve its accuracy on harder-to-detect emotions, such as sadness, anger, and fear. This could be achieved by training on a more diverse dataset that includes a wider variety of facial expressions, as well as using techniques like transfer learning or a more complex neural network architecture to better capture subtle features of these emotions.
- **Data Augmentation**: Applying more advanced data augmentation techniques could help in making the model more robust to variations in real-world input, such as changes in lighting, facial orientation, and background noise. Increasing the diversity of the dataset, including images from various ethnic backgrounds and age groups, would help improve the model’s generalization ability.
- **Integration with Music Subscription Platforms**: Currently, the system relies on user-defined playlists stored locally. However, integrating with popular music subscription platforms such as Spotify or Apple Music could greatly expand the music library. By leveraging their APIs, the system could automatically suggest songs or playlists based on users’ preferences and mood, offering an even more seamless and expansive music experience.
- **Real-time Feedback and Learning**: Implementing real-time feedback from users could allow the system to learn from interactions and improve the emotion-music mapping over time. A user-driven feedback loop, where users can rate whether the music selected based on their detected emotion fits their mood, could help fine-tune the system’s music selection algorithm.
- **Cross-Modal Integration**: Future work could involve integrating other sensors, such as heart rate or GSR (Galvanic Skin Response), to complement the emotion detection process. These additional inputs could provide a richer understanding of the user’s emotional state, improving the overall music selection accuracy.

By addressing these areas of improvement, the system could evolve to become more accurate, adaptable, and integrated with existing music ecosystems, ultimately offering users a more sophisticated and enjoyable music experience tailored to their emotions.

References

[1] Swathi P, Sai Tejaswi D, Amanulla Khan M, Saishree M, Babu Rachapudi V, Kumar Anguraj D. A research on a music recommendation system based on facial expressions through deep learn-

ing mechanisms. Gamification and Augmented Reality. 2024; 2:38. https://doi.org/10.56294/gr202438

[2] Preema J. S, Rajashree, Sahana M, Savitri H, Shruthi S. J, Review on Facial Expression Based Music Player, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) ICRTT – 2018 (Volume 06 – Issue 15).

[3] S.k. Sana, G. Sruthi, D. Suresh, G. Rajesh, G.V. Subba Reddy, Facial emotion recognition based music system using convolutional neural networks, Materials Today: Proceedings, Volume 62, Part 7, 2022, Pages 4699-4706, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2022.03.131.

[4] M. A. H. Akhand, Shuvendu Roy, Nazmul Siddique, Md Abdus Samad Kama and Tetsuya Shimamura, "Facial Emotion Recognition Using Transfer Learning in the Depp(CNN)," Electronics 2021, 10,1036. https://doi.org/10.3390/electronics10091036, 2021.

[5] AYUSH Guidel, Birat Sapkota, Krishna Sapkota, Music recommendation by facial analysis, February 17, 2020.

[6] Athavle M, Mudale D, Shrivastav U, Gupta M. Music Recommendation Based on Face Emotion Recognition. Journal of Informatics Electrical and Electronics Engineering (JIEEE) 2021;2:1-11. https://doi.org/10.54060/JIEEE/002.02.018.

[7] J. Oheix, "Face Expression Recognition Dataset," Kaggle, Oct. 2021. [Online]. Available: https://www.kaggle.com/datasets/jonathanoheix/face-expression-recognition-dataset/data. [Accessed: 02-Dec-2024].