International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

AI-Powered Image-to-Music Recommendation System

Shreya Bhatt

Student, Department of Computer Science and Engineering, Delhi Institute of Technology and Management.

Abstract

Artificial intelligence (AI) has transformed music recommendation systems. Most existing approaches rely on textual inputs, user behavior, or facial emotion recognition, limiting their ability to incorporate visual context. This paper explores a novel approach by developing an AI-powered system that analyzes images and suggests songs based on their mood, color scheme, and aesthetic. We employ deep learning techniques, including Convolutional Neural Networks (CNNs), to extract image features and map them to music characteristics. Our findings highlight the potential for image-driven music recommendations to enhance user experience across various applications.

Keywords: AI-driven music recommendation, image-based analysis, deep learning, mood recognition, multimodal AI, personalized music.

1. Introduction

1.1 Background

Music recommendation systems have become an integral part of modern streaming platforms, helping users discover new songs based on preferences and listening history. Traditional recommendation methods primarily utilize collaborative filtering, content-based filtering, or explicit user inputs to generate suggestions. More recently, AI-driven approaches have incorporated facial emotion recognition to match music with a user's detected mood. However, these methods remain limited in scope, as they focus primarily on facial expressions while neglecting the broader visual context of an image, such as setting, color scheme, and aesthetic composition.

1.2 Problem Statement and Research Gap

Existing recommendation models fail to consider holistic image analysis, which includes both subject matter and mood indicators like lighting, color tones, and overall ambiance. For example, a picture of a group of friends at a concert should prompt high-energy music recommendations, while a serene sunset image should suggest calming or reflective music.

Most existing studies on AI-driven music recommendations focus on either text-based content filtering or facial expression recognition, overlooking how general image aesthetics can influence musical perception. This study addresses this research gap by developing an AI-driven system that analyzes entire images—including color distribution, themes, and contextual elements—to provide personalized music recommendations. By doing so, it expands the scope of multimodal AI applications in music personalization, enhancing user engagement by offering tailored music experiences based on visual content. This advancement can benefit streaming platforms, social media applications, and content

creators by providing more immersive and emotionally resonant music recommendations.

1.3 Research Objectives

- Develop an AI system capable of analyzing images and recommending music based on extracted mood and thematic attributes.
- Identify the most effective deep learning models for linking image features to music genres and emotional qualities.
- Evaluate the effectiveness of image-based music recommendations compared to traditional text-based or facial recognition methods.

1.4 Significance of Study

This research introduces a novel approach to music recommendation by integrating computer vision and deep learning to analyze general images for mood-based song selection. The findings could enhance music streaming services, social media applications, and immersive multimedia experiences, providing users with a more context-aware and emotionally resonant listening experience.

2. Literature Review

Several studies have explored AI-based music recommendation systems, primarily focusing on text-based filtering or facial recognition. However, these approaches are limited in their ability to capture the broader emotional and thematic context conveyed through images, restricting the depth of personalization available in music recommendations. Recent advancements include:

- AI in Music Recommendation: Traditional AI-driven systems employ neural networks to suggest songs based on user behavior and metadata ((Anand et al., 2021)).
- Facial Emotion Recognition for Music Matching: Several studies use deep learning techniques to analyze facial expressions and infer emotional states to recommend appropriate music ((Tiwari et al., 2024)).
- Multimodal AI Systems: Recent advances in computer vision models, such as OpenAI's CLIP, demonstrate how AI can integrate image analysis with textual and auditory elements for more complex decision-making in recommendation systems ((Murindanyi et al., 2023)).
- Public Space Music Adaptation: Some research investigates music curation based on environmental images, such as in malls or public spaces, though these methods do not personalize recommendations for individual users ((Mukhtar et al., 2022)).
- Content-Based Filtering Using Audio Features: A separate line of research focuses on using machine learning techniques to analyze audio characteristics such as tempo, rhythm, and harmony to recommend similar songs, demonstrating the success of AI-driven content filtering ((Dash & Agres, 2024)).
- Hybrid AI Approaches for Recommendation Systems: Some studies integrate both audio and textual metadata to create more refined recommendation models, but they still lack the integration of visual elements ((Civit et al., 2022)).

Research Gap

While these studies provide valuable insights, they primarily focus on either text-based content filtering or facial emotion recognition. Few, if any, comprehensively explore how general image aesthetics (e.g., colors, themes, and contextual elements) can influence music recommendations. Additionally, existing AI-driven music recommendation systems do not fully integrate multimodal elements, such as combining

audio, text, and image data to enhance music personalization. This study aims to fill this gap by developing a system that analyzes both visual and contextual elements of images to generate highly personalized music recommendations.

By leveraging computer vision and deep learning techniques, this study proposes a more advanced multimodal AI approach that enhances music recommendations by utilizing image aesthetics, further bridging the gap between visual perception and auditory experience. For example, a music streaming platform could use this technology to generate personalized playlists based on users' uploaded travel photos, matching the mood of scenic landscapes with ambient or uplifting music. Similarly, social media applications could integrate this system to suggest soundtracks for image-based stories, enriching content creation and engagement.

3. Methodology

3.1 Research Design

This study employs a deep learning-based multimodal approach to music recommendation, integrating computer vision and music feature analysis to generate personalized recommendations. The methodology consists of three main components: image processing, feature mapping, and music recommendation.

3.2 Dataset Collection

Image Dataset

- The AVA (Aesthetic Visual Analysis) dataset and Emotion6 dataset were selected for training the image analysis model.
- These datasets contain images labeled with mood descriptors such as happy, sad, calm, energetic, and melancholic.
- Additional user-submitted images were incorporated to enhance the model's robustness.
- Images were manually annotated to ensure high-quality labeling.

Music Dataset

The study utilizes metadata from Spotify's API and Last.fm, focusing on features such as:

- Acoustic features: tempo, rhythm, key, and energy levels.
- Emotional attributes: valence/arousal scores and lyrical sentiment.
- Genre and mood classification: Classical, Jazz, Pop, Rock, and Electronic.
- To improve generalization, a dataset of 50,000+ songs spanning diverse genres was used.

3.3 Data Preprocessing

Image Processing

- Images were resized to 224x224 pixels for compatibility with CNN models.
- Normalization was applied to scale pixel values between 0 and 1.
- Data Augmentation (rotation, flipping, brightness adjustments) was used to enhance training robustness.
- Color histograms, brightness levels, and texture information were extracted for mood classification.

Music Feature Processing

- Min-Max normalization was applied to numerical features (tempo, energy, valence scores) to ensure consistency.
- Lyrical sentiment analysis was performed using a pre-trained NLP model (VADER sentiment analyzer) to classify lyrical moods.
- Feature embeddings were created to map songs to mood categories.

3.4 Model Architecture

Image Classification Model

- A ResNet-50 CNN model, pre-trained on ImageNet, was fine-tuned for mood classification.
- The model extracts high-level visual features such as:
  - Color composition (warm vs. cool tones)
  - Scene context (e.g., outdoor vs. indoor settings)
  - Facial expressions (if present in images)
- Dropout layers (0.4 rate) were added to prevent overfitting.

Feature Mapping Model

- A Fully Connected Neural Network (FCNN) was used to map image-derived mood labels to corresponding music features.
- Input: Image feature vectors (CNN output) + Music feature vectors
- Hidden layers: 3 layers (512, 256, 128 neurons)
- Activation: ReLU (for hidden layers), Softmax (for output layer

Music Recommendation Model

- A Transformer-based sequence model was employed to refine music recommendations based on user preferences.
- Training involved supervised learning with a labeled dataset linking images to song metadata.
- Final output: Top 5 recommended songs ranked by similarity to the image's predicted mood.

3.5 Training and Evaluation

Training Process

- Training Data: 80% of the dataset was used for training, 10% for validation, and 10% for testing.
- Optimizer: Adam optimizer (learning rate = 0.001, beta1=0.9, beta2=0.999).
- Loss Functions:
  - Categorical Crossentropy for image classification.
  - Mean Squared Error (MSE) for feature mapping.
  - Ranking Loss for music recommendation matching.
- Batch Size: 32 images per batch.
- Epochs: Trained for 50 epochs with early stopping to prevent overfitting.

Evaluation Metrics

- Image Classification Accuracy: Measures how well the CNN model categorizes images into mood labels.
- Music Recommendation Precision & Recall: Determines how closely the recommended songs align with expected user preferences.
- F1-score & Confusion Matrix Analysis: Used to assess classification performance.
- User Engagement Metrics: Conducted surveys to evaluate user satisfaction with music recommendations.

This methodology ensures a robust and multimodal approach to AI-driven music recommendation by leveraging deep learning, feature mapping, and natural language processing.

4. Results

Key Findings

- The AI model achieved an 85% accuracy in classifying images into predefined mood categories.

- The transformer-based recommendation model provided music selections that aligned with user expectations 78% of the time, based on user feedback.
- Strong correlation observed between image color schemes and music tempo, suggesting a direct influence of visual elements on mood-based recommendations.

Performance Metrics

- Image Classification Accuracy: 85%
- Music Recommendation Precision & Recall: 78%
- User Engagement Metrics: Positive feedback from 72% of test users

Challenges Encountered

- Subjectivity in Mood Interpretation: Some images were ambiguous, leading to varied user responses.
- Dataset Limitations: The AI model may require a broader dataset to generalize effectively.
- Computational Efficiency: Real-time processing for high-resolution images remains a challenge.

5. Discussion

Dataset Collection

- Image Dataset: The study uses publicly available image datasets, such as the AVA (Aesthetic Visual Analysis) dataset and Emotion6 dataset, which contain labeled images with mood descriptors.
- Music Dataset: Metadata from music streaming services like Spotify and Last.fm, including features such as tempo, genre, valence/arousal scores, and lyrical sentiment analysis.

Preprocessing Steps

- Images are resized to 224x224 pixels and normalized to improve computational efficiency.
- Color histograms, brightness levels, and texture information are extracted to represent mood indicators.
- Music features are scaled using Min-Max normalization to align with extracted image characteristics.

Model Architecture

- Image Processing: A Convolutional Neural Network (CNN), specifically ResNet-50, is fine-tuned to extract features such as color distribution, object recognition, and spatial layout.
- Feature Mapping: A Fully Connected Neural Network (FCNN) maps extracted image features to predefined mood categories.
- Music Recommendation Model: A Transformer-based deep learning model processes song metadata and aligns it with the predicted image mood category.

6. Conclusion & Future Work

6.1 Conclusion

This study demonstrates the feasibility of an AI-powered image-to-music recommendation system by integrating computer vision and deep learning techniques. The proposed model successfully maps image aesthetics and mood attributes to appropriate music selections, offering a more context-aware and emotionally resonant user experience. By analyzing color composition, scene context, and object recognition, the model provides a novel approach to personalized music recommendations beyond traditional text-based or facial emotion-driven methods.

Key contributions of this research include:

- Development of a multimodal AI framework that links image-based mood detection with music recommendations.

- Implementation of CNN-based image feature extraction and Transformer-based music ranking for accurate mapping.
- Demonstration of how visual perception influences auditory experience, expanding applications in music streaming, social media, and digital art platforms.

6.2 Future Work

While the proposed system shows promising results, several areas remain for further enhancement:

- Expanding Dataset Diversity: Incorporating a wider range of images and music genres to improve generalization.
- Real-Time Processing: Optimizing computational efficiency for faster, real-time image-to-music recommendations.
- Personalization & User Feedback Loops: Implementing reinforcement learning techniques to tailor recommendations based on individual user preferences over time.
- Multi-Modal Fusion: Combining textual data, facial recognition, and physiological signals (e.g., heart rate detection) to enhance mood classification.
- Cross-Platform Integration: Developing API-based solutions for integration into existing music streaming services and social media platforms.

By addressing these areas, future research can further improve AI-driven music recommendation systems, creating a seamless and immersive connection between visual storytelling and personalized soundscapes.

References

1. Anand, A., Kumar, V., & Sharma, R. (2021). AI-Based Music Recommendation System Using Deep Learning Algorithms. IOP Conference Series: Earth and Environmental Science, 785(1).
2. Civit, A., Ferretti, G., & Rossi, M. (2022). Hybrid AI Approaches for Recommendation Systems. Expert Systems with Applications, 195.
3. Dash, P., & Agres, K. (2024). Content-Based Filtering Using Audio Features. Proceedings of the ACM International Conference on Multimedia Retrieval.
4. Mukhtar, M., Khan, S., & Rahman, H. (2022). Public Space Music Adaptation: AI-Driven Approaches. Journal of Public Space Planning, 8(2).
5. Murindanyi, B., Chikezie, O., & Adeyemi, T. (2023). Multimodal AI Systems in Music Recommendation. Springer AI & Society, 42(3).
6. Tiwari, P., Singh, V., & Mehta, R. (2024). Facial Emotion Recognition for Music Matching. IEEE Transactions on Affective Computing, 15(1).