Emotion2Music: Facial Expression-Based Music Generation

Mingjun Ma, Hansen Wang, Zhikai Zhao
Instructor: Eric Ewing
TA: Yuyang Luo
Github Repo: https://github.com/MingjunMa/CSCI1470_final

Introduction

Music and emotion are deeply intertwined in the human experience. In this project, we explore how artificial intelligence can bridge the gap between human emotional states and auditory expression. We propose a deep learning-based system that recognizes a user's emotional state from facial expressions and recommends music that aligns with the detected emotions. Our framework combines computer vision techniques (using custom-trained CNNs on the FER2013 dataset) with feature-based symbolic music generation guided by extracted characteristics such as tempo, mode, and chord progression. By leveraging a multimodal pipeline, our system demonstrates how emotionally aware media interactions can be both intuitive and personalized. This work highlights the potential of affective computing to enhance user-centered experiences in digital media applications.

Methodology

Face Recognition Dataset

- The data consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centred and occupies about the same amount of space in each image.
- Each image is categorized into one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). The training set consists of 28,709 examples and the public test set consists of 3,589 examples.


Happy


Fear


Angry

Music Demo Dataset

- The MediaEval Database for Emotional Analysis of Music dataset consists of 1802 excerpts and full songs annotated with valence and arousal values both continuously (per-second) and over the whole song.
- We define the corresponding emotion of the music based on the valence and arousal values.

Facial Emotion Recognition

-We preprocess the images by resizing them to 48×48 (if necessary) and normalizing pixel values to the [0, 1] range.
-Our model is a custom-designed **Convolutional Neural Network** built from scratch in TensorFlow. The architecture includes several convolutional layers with ReLU activations, batch normalization, max-pooling layers, and a final dense classification head with softmax activation.
-We use categorical cross-entropy as the loss function and the Adam optimizer for training.
-Early stopping and model checkpointing are employed during training to prevent overfitting and save the best-performing model based on validation accuracy.

Music Features Extraction

- We use librosa package to analyze the features of each music, and the output architecture is as below.

- Tempo: The min-max range is estimated using librosa.beat.beat_track.
- Pitch (chroma): Mean and std are used to approximate pitch range.
- Velocity (loudness): Estimated from RMS energy; approximates MIDI velocity range.
- Note density: Estimated from onset strength.
- Mode: Detected via correlation between chroma and predefined templates.
- Chord progression: Simplified — uses "I-IV-V-I" for major, "i-VI-III-VII" for minor.
- Rhythm pattern: Set as "irregular" if the emotion is related to tension (angry, fear, surprise); otherwise "regular".

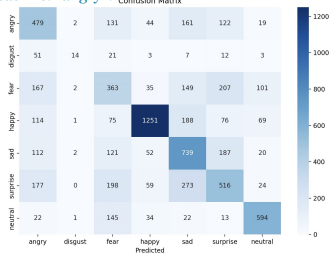
Music Generation

- We use the mido library to create a MIDI file with:
- Tempo track: Random tempo is sampled within the provided range.
- Chord track: It translates abstract chords to actual MIDI notes.
- Pitch: Random melodic pitches are chosen within the defined range.
- Velocity: We add dynamic expression by changing note loudness.
- Note density: It determines how many notes are generated per bar.

Result

Facial Emotion Recognition

Our facial emotion recognition model achieved an overall accuracy of 55% on the FER2013 test set (baseline accuracy 24.71%). It performs best on the “happy” and “neutral” classes, and worst on the "disgust" and "fear" classes. The confusion matrix reveals frequent misclassifications between similar emotions such as “fear” vs. “surprise” and “sad” vs. “angry”.



	precision	recall	f1-score	support
angry	0.43	0.50	0.46	958
disgust	0.64	0.13	0.21	111
fear	0.34	0.35	0.35	1024
happy	0.85	0.71	0.77	1774
sad	0.48	0.60	0.53	1233
surprise	0.46	0.41	0.43	1247
neutral	0.72	0.71	0.72	831
accuracy			0.55	7178
macro avg	0.56	0.49	0.50	7178
weighted avg	0.57	0.55	0.55	7178

Music Generation

We tested a total of 42 images evenly distributed across 7 emotional categories: angry, disgust, fear, happy, neutral, sad, and surprise. The system demonstrated generally accurate emotion-to-music mappings, but some confusion patterns were observed, for example, the music generated for fear and sad tended to sound more similar. Overall, the system successfully demonstrated the ability to generate emotion-aligned symbolic music from facial expressions, while also revealing areas for further refinement in distinguishing subtle emotional nuances.

Discussion

Lesson learned

- Images are 2D grids (pixels), often treated as tensors of shape (height, width, channels).
- Convolutional Neural Networks (CNNs) are extremely effective due to local spatial correlation.

Limitations

- The facial emotion recognition model achieved only about 60% accuracy, which limits the reliability of the downstream music generation.
- The labels and evaluation for music generation are fuzzy. It is hard to define “correct” music — it’s subjective.

Future Work

- We can use some ML-Based Music Generation techniques (e.g., LSTMs, Transformers, or VAEs trained on emotional music).
- We can enable support for changing emotional states over time, allowing the music to evolve and follow emotional trajectories (e.g., from “sad” to “hopeful”).

Reference

- Face Recognition Dataset: https://www.kaggle.com/datasets/msambare/fer2013
- Music Generation Dataset: https://cvml.unige.ch/databases/DEAM/