E-Motion Baton: Human-in-the-Loop Music Generation via Expression and Gesture

Mingchen Ma, Stephen Ni-Hahn, Simon Mak, Yue Jiang, Cynthia Rudin
Duke University
{mingchen.ma, stephen.hahn}@duke.edu

Abstract

We introduce E-Motion Baton, an interactive conducting framework that generates music in real time from a user's gestures and facial expressions. Leveraging computer vision and machine learning, the system tracks motion and emotional cues to dynamically control musical output. Unlike prior work that focuses on either gesture or affect, E-Motion Baton unifies both modalities to create a human-in-the-loop music experience. This positions the system as both a high-level musical instrument and a collaborative tool, with potential applications in music education, therapy, and live performance.

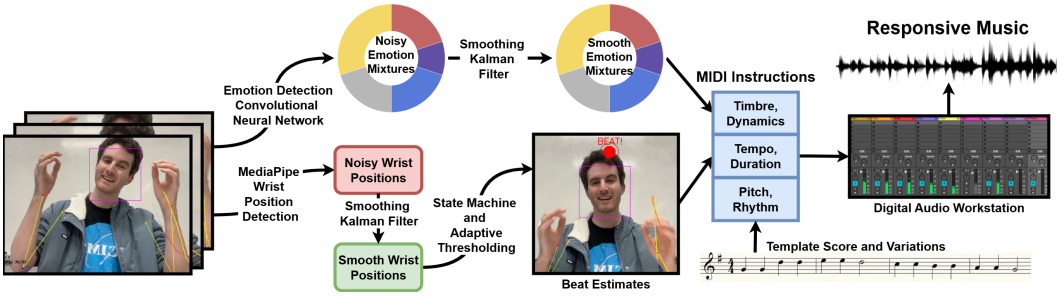

Figure 1: Overview of the E-Motion Baton Framework. Using a live feed from your camera, the model extracts and smooths information about facial expression and wrist position to determine emotional content and beat detection respectively. This information is then combined with several variations of a musical template to create MIDI instructions. The MIDI instructions are then rendered through a Digital Audio Workstation to generate real-time responsive music.

1 Introduction

Have you ever waved your hands, pretending to conduct a symphony? What if the music actually responded to your cues? Our framework, E-Motion Baton, makes these dreams realities by synchronizing music in real time with a user's gestures and facial expressions. Using state-of-the-art computer vision and machine learning, E-Motion Baton detects movements and emotions to shape dynamic musical outputs. Unlike prior gesture- or emotion-only systems, our approach unifies both to create an interactive conducting experience. As such, E-Motion Baton functions as a high-level musical instrument and collaborative partner, with applications in education, therapy, and performance.

Related Work: Prior research on interactive music systems explore how human motion and affect can drive real-time musical experiences. Gesture-based frameworks map hand and arm movements to sounds, enabling embodied interaction with music [7, 10, 2]. Parallel work in dance-to-music generation investigates how large-scale movement patterns shape musical structure [1, 4]. There has

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

also been considerable research in emotion-driven composition systems which translate affective cues into generative music [5, 9, 3]. Together, these efforts illustrate growing interest in multimodal human–music interaction. However, to our knowledge, no system combines facial expression and conducting gestures to mediate musical performance and learning in real time. Our work addresses this gap by developing a real-time conducting platform that leverages both movement and affect.

2 Methodology

E-Motion Baton incorporates emotional modeling, visual emotion detection, motion tracking for tempo control, symbolic music variation generation, and real-time music realization.

Emotion Modeling: Following earlier work [3], we model musical emotion as a mixture of five primary emotions: anger, fear, sadness, joy, and neutral (Figure 2). Concretely, we model each moment t as a mixture of emotions summing to one (e.g., 0.3 sadness, 0.4 fear, and 0.3 anger). By using a mixture of emotions rather than one dominating emotion at a time, we are able to make more nuanced and smooth changes to generated music, granting greater controllability to the user.

Visual Emotion Detection: To determine emotion, E-Motion Baton uses a small five-layer convolutional neural network [8] trained on human facial expressions to detect the most likely emotion at each t. The model takes 48×48 grayscale facial images as input and outputs one of the emotion classes described above. We also estimate the existence of a smile using facial landmarks. In practice, the model's predictions are highly variable from moment to moment. Thus, we implement an emotion-stabilizing Kalman filter [6] that smooths the motion between instances in emotion space. We also find the most prominent emotion within a given window $[t-\alpha, t]$, where α is a hyperparameter determining the time span of the window. The dominant emotion within a window is determined by a time-weighted scoring system, with recent frames given more weight. We also adjust the weight sensitivity for certain emotions based on bias in the dataset and empirical performance.

Beat Detection System: To determine where the beats lie when conducting, we employ a multi-layer protection system that involves a Kalman filter, state machine, and adaptive thresholds. Our Kalman filter smoothly traces users' wrist positions, giving a clear signal for direction and velocity. These motions are interpreted by our state machine comprised of five states: 1) Idle (no motion is taking place), 2) Rising (velocity increasing), 3) Peak (velocity peak), 4) Falling (velocity decreasing), and 5) Cooldown, which prevents immediately triggering a new sequence of states. Intuitively, the state machine cycles through these five states in order for each beat conducted. Finally, using the smoothed velocity information from the state machine, our adaptive threshold system calibrates velocity, acceleration, and energy thresholds based on the user's movements. These thresholds determine when a beat is detected.

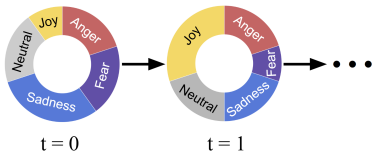

Figure 2: Example emotion mixture changing over time

Symbolic Music Variation: To allow the user to make emotional changes in real time, we require multiple versions for a given piece of music. Thus, we generate musical variations of musicxml scores by including variations on the musical mode (major vs. minor) and note density (more or less embellishment). Joy and neutral are associated with major mode, while fear, anger, and sadness are associated with minor mode; increased fear or anger leads to more embellishment.

MIDI Control System: Finally, we synchronize a MIDI version of the musicxml with the beats detected from the user, generating audio through a Digital Audio Workstation (DAW). Through the DAW, we are able to access different tracks and timbres depending on the emotion and tempo provided by the conductor.

3 Conclusion

E-Motion Baton provides users with fun and intuitive controls for real-time music generation. By merging emotion detection with motion tracking, we extend human-computer interaction into a collaborative, expressive art. This opens opportunities not only for engaging performances but also for applications in education and therapy, where accessibility and emotional engagement are essential.

References

[1] Gunjan Aggarwal and Devi Parikh. Dance2music: Automatic dance-driven music generation. *arXiv preprint arXiv:2107.06252*, 2021.

[2] Google Creative Lab. Semi-conductor. `https://github.com/googlecreativelab/semi-conductor`, 2018. Accessed: 2025-08-18.

[3] Stephen Hahn, Jerry Yin, Rico Zhu, Weihan Xu, Yue Jiang, Simon Mak, and Cynthia Rudin. SentHYMNent: An interpretable and sentiment-driven model for algorithmic melody harmonization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 5050–5060, New York, NY, USA, 2024. Association for Computing Machinery.

[4] Bo Han, Yuheng Li, Yixuan Shen, Yi Ren, and Feilin Han. Dance2midi: Dance-driven multi-instrument music generation. *Computational Visual Media*, 10(4):791–802, 2024.

[5] Fathinah Izzati, Xinyue Li, and Gus Xia. Expotion: Facial expression and motion control for multimodal music generation. *arXiv preprint arXiv:2507.04955*, 2025.

[6] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

[7] Mahya Khazaei, Ali Bahrani, and George Tzanetakis. A real-time gesture-based control framework, 2025.

[8] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

[9] Shilin Liu, Kyohei Kamikawa, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Zero-shot controllable music generation from videos using facial expressions. In *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, pages 1169–1170. IEEE, 2024.

[10] Yue Yang, Zhaowen Wang, and Zijin Li. MuGeVI: A multi-functional gesture-controlled virtual instrument. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 536–541, 2023.