PhD Candidate full name:

Serkan Sulun

Dissertation Title: Video-based Music Generation

Date: 08/10/2025 14:00

Location: University of Porto | FEUP | DEEC | I-105

Higher Education Institution: University of Porto

Doctoral Programme: Doctoral Program in Electrical and Computer Engineering (FEUP)

Abstract or Public Summary: As the volume of video content on the internet grows rapidly, finding a suitable soundtrack remains a significant challenge. This thesis presents EMSYNC (EMotion and SYNChronization), a fast, free, and automatic solution that generates music tailored to the input video, enabling content creators to enhance their productions without composing or licensing music, streamlining creativity and production. Our model creates music that is emotionally and rhythmically synchronized with the video, offering an adaptive and expressive solution for automatic soundtrack generation. A core component of EMSYNC is a novel video emotion classifier. To achieve accurate and efficient video classification, we intelligently fuse pretrained models. We additionally address the data-centric challenges in video classification through cinematic trailer genre classification experiments using a large-scale dataset. By leveraging pretrained deep neural networks for feature extraction and keeping them frozen while training only fusion layers, we reduce computational complexity while improving accuracy. We show the generalization abilities of our method by obtaining state-of-the-art results on Ekman-6 and MovieNet, the largest video datasets for emotion and cinematic genre classification, respectively. Another key contribution is a large-scale, emotion-labeled MIDI dataset for affective music generation. Using annotations from online resources, we build the largest MIDI dataset with valence-arousal labels. We additionally analyze the emotional content of song lyrics within the MIDI files. We then present an emotion-based MIDI generator, the first to condition on continuous emotional values rather than discrete categories, enabling nuanced music generation aligned with complex emotional content. To enhance temporal synchronization, we introduce a novel temporal boundary conditioning method, called "boundary offset encodings," aligning musical chords with scene changes. Integrated into EMSYNC, this method ensures music naturally follows the video's pacing and rhythm, improving the overall user experience. We also explore audio synthesis, focusing on audio bandwidth enhancement due to the scarcity of paired MIDI-audio data. We present a proof-of-concept to highlight and address the challenges in audio synthesis, emphasizing generalization. For the first time, we identify the problem of "filter overfitting," where models trained on specific low-pass filters fail to generalize to

real-world scenarios. To address this, we propose a data augmentation strategy that outperforms standard regularization methods, marking the first step toward developing robust audio enhancement models for real-world use. Combining video emotion classification, emotion-based music generation, and temporal boundary conditioning, EMSYNC emerges as a fully automatic video-based music generator. User studies show that it consistently outperforms existing methods in terms of music richness, emotional alignment, temporal synchronization, and overall preference. As a result, EMSYNC sets a new state-of-the-art in video-based music generation, creating music that is both emotionally and rhythmically aligned with the video.

Principal Supervisor at INESC TEC: Paula Viana

Additional Supervisor: Matthew E. P. Davies

Scientific Domain: [Artificial Intelligence]; [Computer Science and Engineering]

Keywords: deep neural networks; midi generation; video analysis; transformers; multimodal fusion; affective computing