Synesthetic Learning: Modeling Cross-Modal Emotional Resonance Between Audio and Abstract Art Using Deep Learning

Visaviern Mosqueda, Chris Jallaine Mugot, Frency Rayne Montesclaros, and Qylle Christian Quiño

Department of Data Science - College of Information Technology and Computing
University of Science and Technology of Southern Philippines
Lapasan, Cagayan de Oro City 9000, Philippines

1. Introduction

1.1 Background

Emotion is an essential and universal aspect of human perception. It guides decision‑making, memory, and creativity — shaping how people interpret and interact with their surroundings. Humans naturally connect emotions across sensory experiences: a calm melody might remind someone of soft pastel colours, while chaotic sounds may evoke sharp, dark visuals. This interconnection suggests that emotion is not confined to a single sensory channel but instead emerges as a shared affective experience spanning multiple senses.

Despite major advances in affective computing, current artificial intelligence (AI) systems largely analyse emotional content within a single modality — for example, facial expressions in vision, tone in speech, or melody in music. However, few studies attempt to bridge emotion between modalities, especially between auditory and visual domains. This leaves a critical research gap: can machines learn to associate emotions across senses the way humans intuitively do?

This project addresses that gap by investigating the concept of synesthetic learning, inspired by the neurological phenomenon of synesthesia — where stimulation of one sense automatically triggers perception in another (e.g., "seeing" colours when hearing sounds) [1]. Using deep learning, this study aims to model how emotional meaning transfers between sound and visual abstraction, simulating the human‑like resonance between what we hear and what we see.

Abstract art, unlike representational imagery, is a particularly fitting medium for this exploration. Its lack of concrete subjects allows emotion to emerge purely through colour, form, and texture — paralleling how emotion in audio arises from rhythm, tone, and timbre.

By focusing on these expressive and symbolic properties, this research seeks to train a neural model capable of recognising and aligning emotional correspondence between abstract art and affective audio.

Such work lies at the intersection of computational creativity, multimodal affective computing and psychological modelling. It does not merely classify emotion but seeks to understand its resonance across sensory channels — a step closer to machines capable of emotional perception and empathetic artistic synthesis.

1.2 Problem Statement

While significant progress has been made in unimodal emotion recognition, emotional understanding in artificial systems remains fragmented. Most deep-learning models operate within a single sensory input — for example, CNNs for visual emotion detection or RNNs for speech-based emotion recognition. These systems perform well in isolation but fail to capture the cross-sensory dynamics of human emotion, where a sound can evoke a colour or an image can suggest a tone.

Existing research in cross-modal learning primarily focuses on semantic alignment (e.g., matching captions with images or sound events with video clips), rather than emotional alignment. This neglects the affective layer that drives human perception and creative expression. Additionally, emotional content in art and audio is often subjective, making it more complex than standard classification tasks [2,3].

Hence, there is a compelling need to design a framework that:

- Understands emotion beyond modality boundaries, linking auditory and visual cues through shared emotional representation;
- Models emotional resonance – not only matching classes like "happy" or "sad," but learning the subtler gradients of affect that exist in both abstract colour composition and tonal expression;
- Advances affective AI toward creativity and empathy, expanding its use in therapy, design, and human–computer interaction.

This project therefore seeks to build a deep cross-modal model that learns to associate emotion between abstract art and affective sound clips, aligning them through shared embeddings that reflect emotional similarity. By doing so, it contributes to understanding how deep-learning systems can internalise and express human-like emotional synthesis.

2. Objectives

The main goal of this study is to develop a cross-modal deep learning framework that can model emotional resonance between abstract visual art and affective audio samples. Specifically, the project aims to:

1. **Develop a dual-stream deep learning model** — one processing abstract art images and another processing emotion-labeled audio — to extract high-level affective features and align them within a shared latent space.

2. **Model emotional correspondence** between modalities, encouraging emotionally similar pairs (e.g., a sad sound and a somber painting) to converge in embedding space, while emotionally different pairs diverge.

3. **Evaluate the model's emotional alignment capability** through quantitative metrics (embedding similarity, clustering accuracy) and qualitative measures (human judgment of emotional congruence).

4. **Analyze learned affective representations** to interpret which visual or auditory features (e.g., color intensity, spectral energy, rhythm) most strongly influence perceived emotion alignment.

2.1 . Significance

This study contributes both technically and socially to the fields of affective computing and computational creativity. Technically, it introduces a novel framework for cross-modal emotion learning, expanding beyond traditional unimodal emotion classification. By leveraging deep representational learning, the project unifies abstract visual and auditory affect within a shared embedding space—an approach that remains rare in current literature [2,3]. The study advances multimodal neural architectures capable of perceiving emotion holistically rather than as isolated signals.

Furthermore, insights derived from the learned representations can reveal how emotion manifests differently through colour, texture, tone, and rhythm, deepening the computational understanding of human perception.

From a social and creative perspective, emotionally intelligent systems developed through this research can support art therapy by providing emotionally congruent audio-visual feedback that promotes relaxation or mood recovery. In digital art and media design, such systems can generate or recommend emotionally consistent sound-art combinations for exhibitions, interactive installations, or film scoring. The study also reinforces human-centred

AI development, aiming to create systems that do more than process emotion—they resonate with it, fostering empathy, creativity, and psychological well‑being. In essence, this research explores how technology can begin to “feel,” not in a human sense, but by computationally approximating the shared language of emotion across art and sound.

2.2. Expected Outcomes

By the end of the project, the following outcomes are expected:

- **A trained deep learning model** capable of aligning emotional representations between abstract art and audio clips, establishing measurable cross-modal affective similarity.
- **A shared emotion embedding space**, visualized to show how emotions cluster across modalities, revealing correlations such as “warm colors” aligning with “soft harmonics.”
- **Quantitative metrics** demonstrating the model’s emotional alignment performance, validated by both machine and human evaluation.
- **Qualitative insights** into cross-modal affect — how emotional cues in sound and color co-vary, and what this reveals about affective cognition.
- **A reproducible dataset framework** combining the *Audio Emotions Dataset* (Kaggle) and *D-ViSA Abstract Art Dataset* (GitHub), preprocessed and annotated for future multimodal research.

Ultimately, the project aims to contribute a meaningful step toward emotionally aware AI — capable of perceiving and relating human emotion across different sensory forms, much like the natural phenomenon of synesthesia itself.

3 References

[1] J. R. Bock, “A Deep Learning Model of Perception in Colour-Letter Synesthesia,” *Big Data and Cognitive Computing*, vol. 2, no. 1, p. 8, 2018.

[2] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, Y. Zong, “A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face,” *Entropy*, vol. 25, no. 10, p. 1440, 2023.

[3] M. Jia, Z. Sun, “A Survey of Multi-modal Emotion Recognition Based on Deep Learning,” *Highlights in Science, Engineering and Technology*, vol. 119, pp. 533-540, 2024.