結合肢體與音訊特徵之個人化鋼琴演奏視覺系統設計

蔡岱臻[1] 黃稚筌[1] 許程胤[1] 張舜涵[1] 李炘穎[1] 溫凱翔[2] 潘則佑[3] 胡敏君[1]
國立清華大學[1] 青鑫藝術[2] 國立臺灣科技大學[3]

Abstract

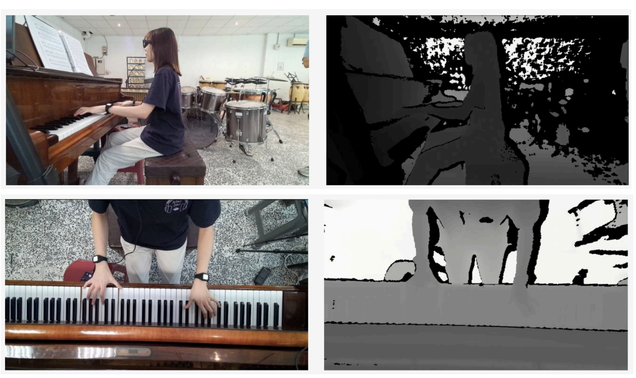
本研究提出一套個人化鋼琴演奏視覺設計系統，整合鋼琴演奏者的肢體動作與音訊特徵，以強化視覺輸出中的情緒深度與表演者意圖之呈現。系統透過即時影像分析推估演奏者所傳遞的情緒狀態，並根據辨識結果切換對應視覺風格。同時，結合音訊特徵即時調整粒子動態、色彩與視覺節奏，呈現多層次的視覺變化。系統亦支援使用者自定義情緒與音訊特徵對應的視覺規則，實現個人化的視覺設計。為支援系統建置與驗證，我們建構一組多模態鋼琴演奏資料集，包含同步的 RGB 影像、音訊與情緒象限標註。本研究展示結合肢體表現與音樂分析進行情緒驅動視覺化的可行性，並為發展更具表演者感知能力與情感深度的音樂互動體驗奠定基礎。

Method

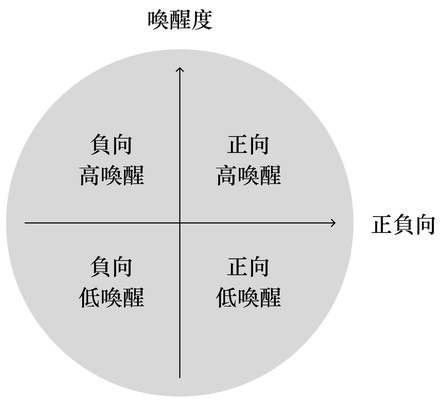
本研究開發一個整合演奏者肢體動作與音訊特徵的多模態鋼琴演奏視覺化系統。研究方法包含三個核心部分：⑴多模態鋼琴演奏資料集建置、⑵情緒辨識模型比較與評估、⑶音訊特徵提取與處理。

多模態鋼琴演奏資料集建置

資料集採用雙 Azure Kinect 攝影機擷取俯視/側視 RGB-D 影像並同步錄製音訊（如圖一）。選用 EMOPIA[1] 資料集中基於情緒環狀模型（如圖二）的四象限樂曲各5首（共20首），該模型以兩個維度描繪情緒：正負向（Valence）與喚醒程度（Arousal）。「正向」與「負向」代表情緒的愉悅程度，如快樂或悲傷，而「喚醒程度」則指情緒的活躍性，例如是激動或平靜。根據這兩個維度，可將情緒劃分為四個象限：正向高喚醒、負向高喚醒、負向低喚醒與正向低喚醒。5位專業鋼琴家（音樂系或10年以上的演奏經驗，4女1男）參與實驗，累積156分鐘多模態數據與主觀情感標註。



圖一、兩台 KINECT 攝影機分別從不同角度同步錄製演奏者的身體姿態手部動作。側面視角（上）和俯視視角（下）的 RGB 影像與對應的深度影像。



圖二、 情緒環狀模型

情緒辨識模型比較與評估

為評估多模態資料於鋼琴演奏情緒辨識任務中的應用潛力，本研究比較 ST-GCN++[2] 與 TSM[3] 於側/俯視視角的情緒分類表現（表一）。ST-GCN++在側身視角與俯視視角間展現相對穩定的表現（F1 分別為 0.50 與 0.49）TSM 則呈現視角敏感性，從側身視角的 0.46 提升至俯視視角的 0.64。TSM 在俯視視角下展現顯著優勢，能完整呈現演奏者雙手動作，包括觸鍵力度、手指運動與手腕角度等情緒表達關鍵特徵。TSM 擅長捕捉影像序列時間動態變化，有效處理局部動作連續性，相較於依賴完整骨架資訊的 ST-GCN++，更適應手部區域專注辨識任務。

音訊特徵提取與處理

現場演出除了鋼琴演奏者的肢體動作外，透過鋼琴所發出的音符構成了情感表達的聲學特徵。為客觀解析出多維度的情感特徵，本研究設計從鋼琴音訊提取多維情感特徵作為視覺化即時調控參數。

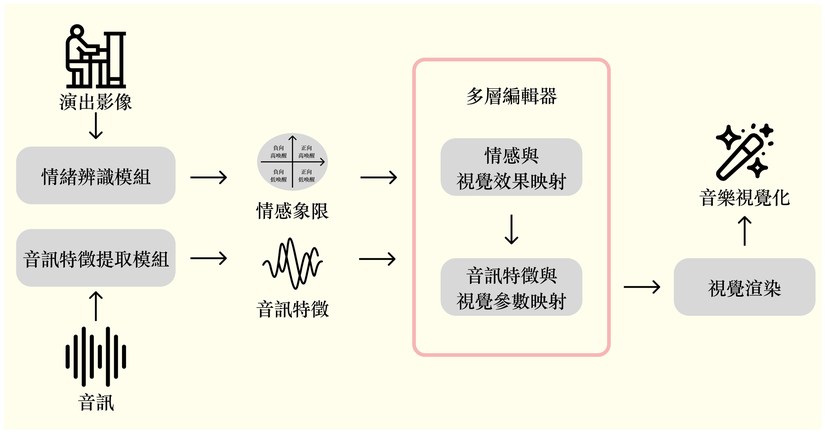
時域特徵：

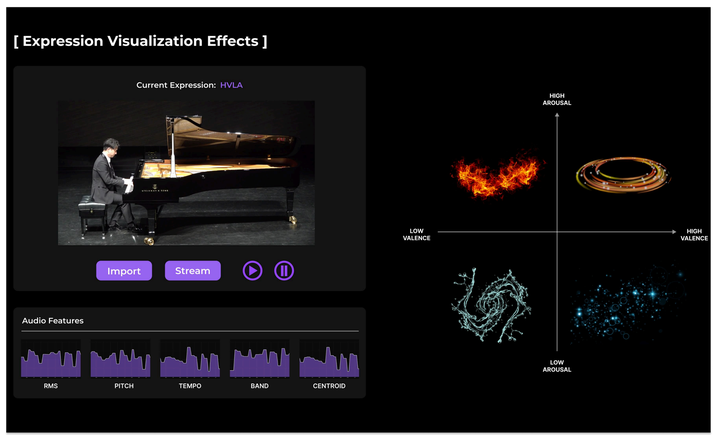
1. 節奏（Tempo）：反映音樂速度變化，影響情緒起伏與張力
2. 均方根能量（RMS）：描述音訊整體強度，反映音樂力度變化

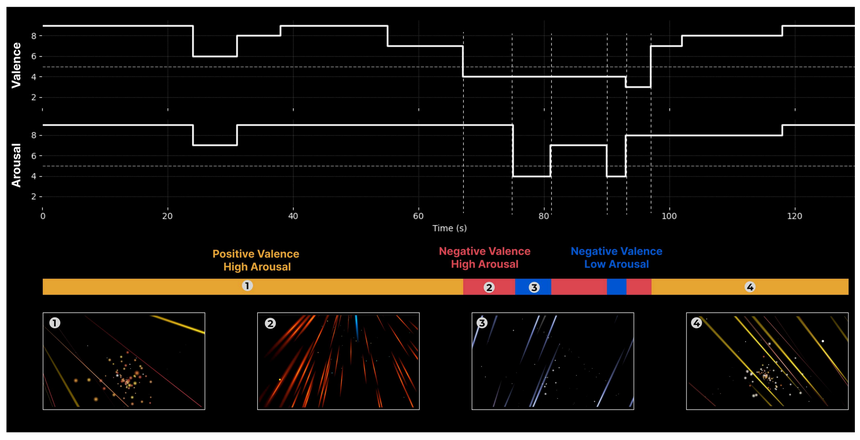
頻域特徵：

1. 頻譜重心（Spectral Centroid）：頻譜能量重心，反映音色明暗特質
2. 音高(Pitch)：音符基頻位置，直接影響情緒色彩

Result







圖三、模態鋼琴演奏視覺化系統。(左) 系統架構圖；(中) 使用者互動介面；(右) 音高驅動的視覺化效果變化示意圖。

基於上述，本研究開發雙流程並行處理架構（圖三左）。身體動作流程透過情緒辨識模型分析肢體動作，由視覺風格選擇模組根據四象限情緒模型自動切換視覺風格；音訊處理流程經音訊特徵擷取模組提取時/頻域特徵，透過視覺參數調整模組即時控制粒子動態、色彩與空間分佈。系統提供直觀介面供使用者預定義情緒象限視覺風格與音訊參數映射規則，實現個人化設定。演奏過程中，系統即時辨識情緒狀態並自動切換視覺效果（圖三中），同時音訊分析持續驅動粒子密度、速度、色彩變化（圖三右）。最終由即時渲染器整合雙流程輸出，生成融合情緒表達與音樂結構的動態視覺效果。

Next Step

本研究建立完整的多模態鋼琴演奏分析框架，以演奏肢體動作與音訊特徵結合形成一個以演奏者情感表達為核心、音訊特徵為輔助的視覺化系統。預計規劃使用者研究估視覺化效果與音樂情緒對應程度、情緒傳達準確性，以及對音樂欣賞體驗的提升效果。同時驗證操作介面直觀性與系統穩定性。

Reference

[1] Hsiao-Tzu Hung, Joann Ching, et al. “EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. ” International Society for Music Information Retrieval (2021)
[2] Duan, Haodong, et al. "Pyskl: Towards good practices for skeleton action recognition." Proceedings of the 30th ACM International Conference on Multimedia. 2022.
[3] Lin, Ji, Chuang Gan, and Song Han. "Tsm: Temporal shift module for efficient video understanding." Proceedings of the IEEE/CVF international conference on computer vision. 2019.