基於深度學習的手語辨識與語音轉譯平台

專題編號：114-1-CSIE-S004
執行期限：113 年第 1 學期至 114 年第 1 學期
指導教授：郭忠義教授
專題參與人員： 111820017 陳嘉祥
111820023 王品翔
111820030 楊承諭
111820033 江子文

一、摘要

本專題旨在開發一個基於深度學習的手語辨識與語音轉譯平台，實現「手 語 → 語音文字」與「語音 → 手語」的雙向轉換。系統設計包含以下幾個核 心部分：

首先，使用 MediaPipe 從影片中擷取手部關鍵點座標，經正規化處理後生 成角度與距離等特徵，並建立統一的手勢特徵資料集。接著，透過隨機森林分類器與 Transformer 模型進行手語辨識訓練，能有效提取時間序列特徵並達到高準確率。

在語音轉手語方面，系統採用 Vosk 進行即時語音辨識，將語音輸入轉換為文字，並 LabelMatcher 演算法進行詞彙分段與模糊匹配；若資料庫中存在對應手語，即可即時於畫布播放骨架動畫，若無則回退至 A–Z 字母手語拼字，以確保完整覆蓋率。資料管理則使用 SQLite 建立手語骨架資料庫，並透過 WebSocket 進行即時資料傳輸。

最終系統以網頁平台形式呈現，前端負責音訊串流、動畫繪製與播放佇列管理，後端以 FastAPI 提供語音辨識與手語查詢服務。整體架構兼具即時性、互動性與容錯性，可應用於手語學習、輔助溝通。

二、緣由與目的

手語是聾啞及聽障人士的重要溝通方式，但由於大眾對手語理解有限，導致其在日常生活中仍面臨溝通隔閡。現有的溝通輔助工具大多僅能支援「語音 轉文字」或「文字轉手語」的單向功能，缺乏即時性與雙向互動，限制了實際 應用情境。此外，傳統手語翻譯需依賴專業人員，成本高且可及性有限。隨著 深度學習與多媒體處理技術的快速發展，如何透過人工智慧自動化地辨識手語 並結合語音辨識技術，建立一個即時、雙向的輔助平台，成為值得投入的重要課題。

三、研究架構

本研究整體流程共分為六個主要步驟，涵蓋資料建立、模型訓練、特徵提取、語音整合與系統應用。各步驟說明如下：

1. 手語資料蒐集與標註:

蒐集公開手語影像 資料集（如 ASSLVD）及自行錄製之手語影片，並自行拍攝 ASSLVD 中提供之常見手勢作為手勢辨識訓練資料。每筆資料包含影像內容、時間戳與語意對應文字，以建立多來源、具代表性的訓練資料集。

2. 骨架特徵擷取與資料前處理:

利用 **MediaPipe Hands** 模組擷取每幀影像之手部關鍵點（21 個關節），生成三維座標資料。後續進行資料清理、平移（以鼻子為基準）、縮放（以肩距為標準）、鏡像統一及角度與距離特徵計算，形成結構化手勢特徵，供模型訓練使用。

3. 靜態手勢辨識訓練與驗證:

針對單一手勢影像，以 **Random Forest** 模型進行訓練與驗證，並透過交叉驗證監控準確率（Accuracy）、召回率（Recall）與 F1 分數。

4. 連續手語辨識模型建立 :

針對連續影像序列，使用 **Transformer Encoder** 搭配 **Masked Autoencoder (MAE)** 進行預訓練，使模型學習時序與空間關聯。再以 **MLP 分類器** 進行下游手語片段分類，透過 Adam 優化器與交叉熵損失函數訓練，最終以準確率、Recall 與 F1 分數評估模型在不同手語句型上的辨識表現。

5. 語音辨識與單字索引整合:

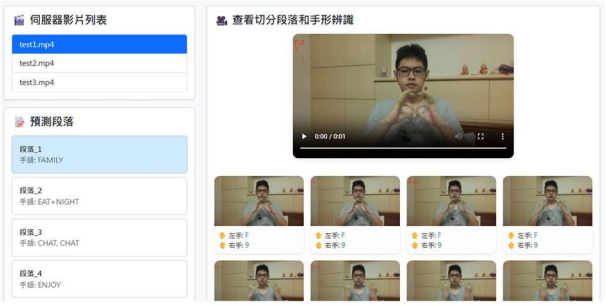
使用 **Vosk 語音辨識模型** 進行語音輸入的即時轉譯，將辨識結果經文字正規化後與手語標籤資料庫比對。再透過 **模糊比對與動態規劃演算法**（Levenshtein Distance + DP）進行詞級匹配，產生與語音對應之手語標籤序列，並輸出至播放模組。

6. **系統整合與即時應用**

將語音辨識模組、單字索引模組與手語播放模組整合為一體化系統。使用者以麥克風輸入語音後，系統即時顯示辨識文字、匹配手語動畫與信心分數，達成語音轉手語的即時互動應用。

四、實驗結果

本實驗涵蓋 **43 種手勢**與 **303 種手語**的辨識任務。手勢辨識表現優異，**mAP@43 約 0.970**，顯示隨機森林模型在靜態手形分類上高度準確且穩定，為手語序列辨識提供可靠基礎。手語辨識面對更多類別與時序變化，**mAP@303 約 0.896**，表明 Transformer Encoder 結合 MLP 分類器能有效學習手語動作的時空特徵，對長序列與複雜手語具有良好泛化能力。



圖一、手語辨識 Demo

圖二、手勢辨識 Demo

圖三、語音轉手語 Demo

五、結論

本專題開發了一個基於深度學習的手語辨識與語音轉譯平台，實現「手語 → 文字」與「語音 → 手語」雙向轉換。系統透過 MediaPipe 擷取手部關鍵點，結合隨機森林與 Transformer 模型，靜態手勢準確率達 97%，手語句子準確率超過 90%。語音轉手語結合 Vosk 即時辨識與匹配演算法，支援資料庫查詢與字母拼字回退，確保溝通完整性。整合於網頁平台後，可即時處理語音串流與手語動畫。研究中發現段落切分不夠精確會影響手語模型預測效果，未來可透過滑動視窗優化、連續性追蹤與後處理策略改善辨識穩定性，同時擴充詞彙與引入多模態資訊提升系統泛用性。

六、參考文件

[1] 論文閱讀 | Masked Autoencoders Are Scalable Vision Learners : https://medium.com/@glowing_sage_deer_60/%E8%AB%96%E6%96%87%E9%96%B1%E8%AE%80-masked-autoencoders-are-scalable-vision-learners-f3858e2ba88

[2] F. Zhang et al., "MediaPipe Hands: On-device Real-time Hand Tracking," Google AI Blog/Docs, 2019. (https://developers.google.com/mediapipe)

[3] S. Ramírez, S. Montiel, and S. Tiangolo,
"FastAPI Framework Documentation," 2019.
(https://fastapi.tiangolo.com)