基於骨架特徵與 Transformer 架構之連續手語辨識與翻譯系統設計

Design of a Transformer-Based Continuous Sign Language Recognition and Translation System Using Skeleton Features

指導教授:韓欽銓 教授

學生: 王俊元 賴煒稚

國立聯合大學 資訊工程學系

苗栗市南勢里聯大 2 號

{cchan, U1124007, U1124023 }@nuu.edu.tw

摘要

本研究結合電腦視覺與人工智慧技術 ，提出一套具延展性的連續手語辨識系統架構，以增進聾啞人士在社會互動中的可能性。方法整合姿態關鍵點擷取、BERT 概念與自監督學習策略，使模型能在一般個人電腦環境中進行訓練，並在有限資源下維持良好效能，有效降低訓練難度與成本。本研究最終開發出可進行即時手語辨識與影片手語解析之平台，實證系統的可行性。

關鍵詞：電腦視覺、人工智慧、聾啞人士 、手語辨識

Abstract

This study integrates computer vision and artificial intelligence techniques to propose a scalable architecture for continuous sign language recognition, aiming to enhance communication possibilities for individuals with hearing impairments. The proposed method incorporates pose keypoint extraction, BERT-inspired representations, and self-supervised learning strategies, enabling the model to be trained on standard personal computers while maintaining strong performance under limited computational resources. These design choices effectively reduce the difficulty and cost of training. Finally, we developed a platform capable of real-time sign language recognition and video-based sign language analysis, demonstrating the feasibility of the system.

Keywords: Computer Vision, Artificial Intelligence, Deaf and Hard-of-Hearing Individuals, Sign Language Recognition

第一章 研究動機

為了協助聽障者與其他無障礙族群更順利地融入社會與職場生活，本研究旨在開發一套基於骨架特徵的手語自動辨識與翻譯系統。透過深度學習模型，即時地將手語動作轉譯為文字，降低溝通障礙，促進資訊平權。期許未來本系統能廣泛應用於就業輔助、醫療服務、公共服務窗口等情境，使聽障者不再因身體條件受限而在求職或生活上處於不利地位，真正實現無障礙智慧社會的願景。

第二章 相關研究

2.1 專題內容說明

本研究旨在建構一套能夠自動偵測與辨識影片或影像中連續手語內容的深度學習系統，期望透過結合骨架特徵擷取技術與語意序列建模，實現連續手語辨識系統 (Continuous Sign Language Recognition, CSLR)。

圖 1 連續手語時間平行數據圖

2.2 手語辨識發展趨勢

手語辨識(Sign Language Recognition, SLR)技術依據任務定義與輸入內容不同，主要可分為「孤立手語辨識（Isolated SLR）」與「連續手語辨識（Continuous SLR）」兩大類。

孤立手語辨識指模型的輸入為單一手語動作片段，每段皆對應一個固定的 gloss 標籤。此任務假設手語的起迄點明確、無語境連續性，因此資料標註與模型訓練相對容易。儘管在限定詞彙下可取得高準確率，但無法處理連續語句中的詞彙邊界與語意關係，應用範圍受限。

連續手語辨識需從未分段的手語序列中同時完成動作切割與詞彙辨識，更貼近自然手語使用情境。挑戰包括：缺乏明確動作邊界、手勢快速連貫、同一 gloss 可能具有多種表現，以及需處理多詞序列對齊，因此資料標註與模型訓練相對困難。

第三章 研究方法

3.1 訓練流程

本研究的整體訓練流程可分為三個主要階段，分別為：資料前處理、預訓練遮罩、微調 S2G（Skeleton-to-Gloss）模型建構。

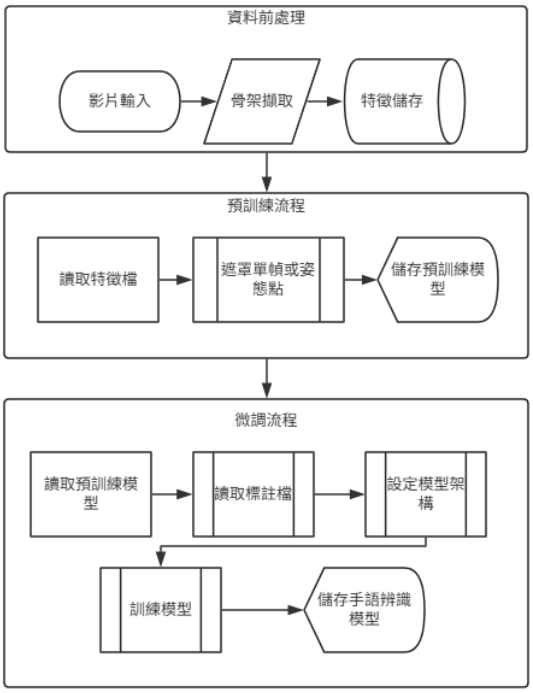

圖 2 訓練流程圖

3.2 資料集

本研究採用 CSL-Daily（Chinese Sign Language Corpus）作為模型訓練與測試的資料來源。該資料集共包含 7,411 個句子，由多位中國手語簽署者錄製而成，每個句子約重複三次，總計約 20,655 支影片，資料總量約 100GB。

圖 3 原始手語圖

※實驗資料集〔1〕：CSL-Daily（Zhou 等人建立之非公開連續手語資料集，經授權使用）。

3.3 資料前處理與特徵擷取

使用 MediaPipe Hands 模組擷取每一幀的 2D 骨架關鍵點，每一幀中左右手各 21 個關鍵點，共 42 個關節點，再加上用到臉部 6 個關鍵點、軀幹 6 個關鍵點總共 54 點作為輸入。每一幀維度(x, y, z)乘以 3，維度共 162 維。

3.4 Transformer 模型架構

基於 Transformer 編碼器之架構，其目標為將輸入的骨架特徵序列映射為對應於 gloss 詞彙空間的時間步分類預測，並最終透過 CTC 解碼器輸出對應的手語詞幹序列。

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

表格 1 模型架構

層次 (Layer)	輸入形狀 (B×T×D)	輸出形狀 (B×T′×D′)	Kernel Size	Stride/Pool	參數數量 (K)
Input	B×T×162	-	-	-	-
Linear + BN + ReLU	B×T×162	B×T×256	-	-	42.3
Conv1D-Block 1	B×256×T	B×256×⌊T/2⌋	5	MaxPool(2)	327.9
Conv1D-Block 2	B×256×⌊T/2⌋	B×256×⌊T/4⌋	5	MaxPool(2)	327.9
Output	B×256×⌊T/4⌋	B×⌊T/4⌋×256	-	-	-

3.5 預訓練與微調策略

在預訓練階段（Pretraining Stage），採用自監督學習方法進行 Masked Frame Modeling，隨機遮蔽輸入骨架序列中的部分時間步，並要求模型根據上下文資訊預測被遮蔽的動作特徵，進一步促使模型學習時間依賴結構與動作序列語意邏輯。

專題使用 CTC（Connectionist Temporal Classification）損失函數來處理連續手語無法對齊 gloss 的問題，主要是把重複字或空白符號，做 collapse（壓縮）處理，會移除連續重複字元與空白符號。

$$L_{CTC} = -log \sum P(\pi, x) \quad (2)$$

3.6 模型評估

Word Error Rate（WER）：WER 為語音與手語識別中常見之指標，用以衡量模型輸出之詞序列與真實標註詞序列之間的差異程度，包含替換（Substitution）、刪除（Deletion）與插入（Insertion）三類錯誤操作。WER 值越低代表辨識準確度越高。

$$WER = \frac{S + D + I}{N} \quad (3)$$

3.7 語言模型訓練流程

本研究採用語言模型微調(Fine-tuning)的方法，在預訓練中文語言模型的基礎上，利用自行整理之手語 gloss–中文平行語料進行訓練，讓模型學會將手語詞序列(Gloss)轉換為自然且語意通順的中文句子。透過本研究自行建立的微調流程，模型逐漸學習手語詞序與中文語法之間的對應關係，進而具備自動語言重組與句式生成的能力。流程圖如下:

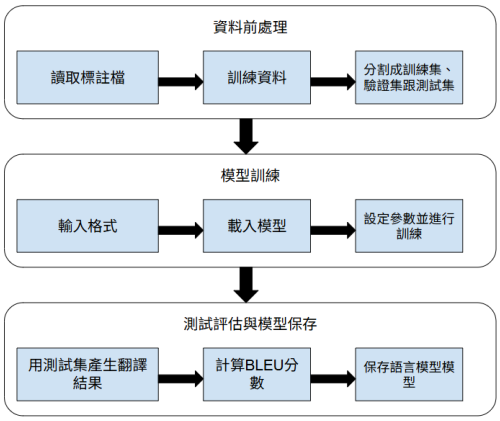

圖 4 語言模型訓練流程圖

3.8 語言模型資料集

語言模型訓練資料集來自 gloss_map 並且從中產生標註檔 gloss-中文平行語料檔 gloss_to_text.json，其中 input 欄位為手語 gloss 序列，output 欄位為對應的中文句子。首先移除任一欄位為空的樣本，只保留同時具有 gloss 與中文標註的配對資料。本研究先從原始資料中擷取前 10,000 筆樣本作為實驗語料，後續所有訓練與評估均在此子集上進行。接著進行資料集切分，比例為訓練集 80%、驗證集 10%、測試集 10%。

3.9 模型選擇與模型架構

本研究並非從零研究模型，而是以公開的中文預訓練序列到序列模型 Mengzi-T5 base(Langboat/mengzi-t5-base)為基礎，載入其預訓練權重與 tokenizer，並在此基礎上進行 gloss→中文任務的微調。除了自行微調之 gloss→中文語言模型外，本研究亦預留與雲端手語翻譯 API（例如 Gemini）整合的介面，作為備援或輔助翻譯模組，以便未來在不同場景下彈性切換或比較不同模型之翻譯效果。

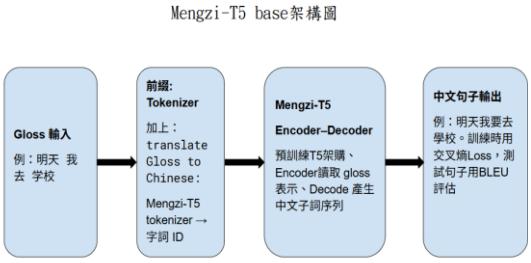

圖 5 語言模型架構圖

除了自行微調之 gloss→中文語言模型外，本研究預留與雲端手語翻譯 API(例如 Gemini)整合的介面，作為備援或輔助翻譯模組，以便未來在不同場景下彈性切換或比較不同模型之翻譯效果。

3.10 語言模型測試集評估與指標

本研究採用機器翻譯領域常用的 BLEU (Bilingual Evaluation Understudy)作為唯一的自動評估指標。BLEU 透過比較模型輸出句子與參考句之間的 n-gram 重疊情形，BLEU 分數介於 0 至 100 分之間，數值越高代表模型輸出在字串層面越接近參考句子。BLEU 公式包含:n-gram 精確率 p_n 跟長度懲罰 Brevity Penalty(BP)，BLEU 通常會對 1~4-gram 取幾何平均。

$$p_n = \frac{\sum_{\text{每個句子}} \text{重疊的} n-gram \text{數量}}{\sum_{\text{每個句子}} \text{模型輸出的} n-gram \text{總數}} \quad (4)$$

$$BP = \begin{cases} 1 & if c > r \\ e^{\left(1-\frac{r}{c}\right)} & if c \le r \end{cases} \quad (5)$$

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (6)$$

第四章 成果展示

4.1 手語辨識模型數據

首先訓練遮罩的自監督預訓練，是使用 Transformer 自編碼器預訓練，在這種自監督學習的預訓練任務中，並沒有一個像分類或檢測那樣的「準確度」或「判斷標準」來判斷模型是否「學會了識別手語」。

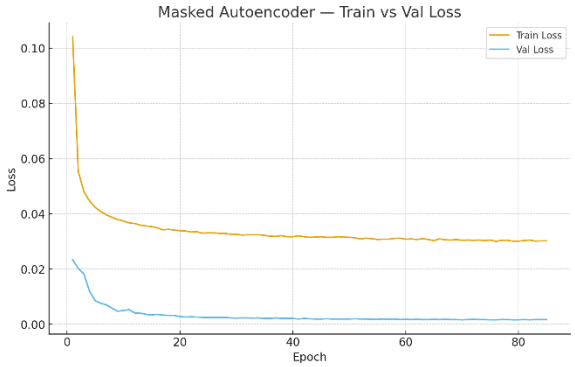

圖 6 訓練遮罩的自監督預訓練

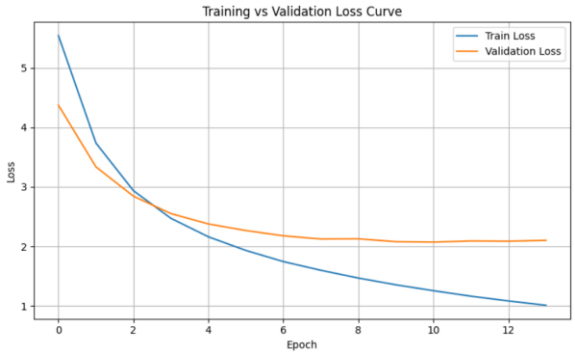

圖 7 微調訓練圖

圖 7 顯示了微調訓練過程的表現，再練過程中雖然訓練集超過驗證集，但是驗證集還在持續下降，代表還沒過擬合，訓練直到不在下降為止才停止訓練。

在每次訓練週期（epoch）結束時，固定使用 CTC 解碼，隨機抽取一段驗證樣本進行質性觀察，輸出「預測序列、參考序列（正解）與對應的 WER」。並同步回報替換/刪除/插入的比率，方便定位主要錯誤來源。

```
Epoch 8 | Train Loss: 1.60 | Dev Loss: 2.12
S002025_P0007_T00 → 預測：下雨 快 来 防 做 今天 | 正解：下雨 快 来 防止 洪水 工作 开始
  ↳ WER: 0.57
Epoch 9 | Train Loss: 1.47 | Dev Loss: 2.13
S001173_P0008_T00 → 預測：驾驶 我 这 手指 茶叶 还 这 | 正解：驾驶 我 在 前 红 绿 灯 这 下
  ↳ WER: 0.67
Epoch 10 | Train Loss: 1.35 | Dev Loss: 2.08
S004284_P0000_T00 → 預測：高兴 圈 房子 里 联系 连 | 正解：玩 高兴 园 惊吓 鬼 房子 里 鬼 连
  ↳ WER: 0.56
Epoch 11 | Train Loss: 1.25 | Dev Loss: 2.07
S000573_P0004_T00 → 預測：为什么 什么 选择 我们 公司 工作 工作 | 正解：请 你 说 为什么 选择 我们 公司 工作
  ↳ WER: 0.62
Epoch 12 | Train Loss: 1.16 | Dev Loss: 2.09
S004525_P0008_T00 → 預測：网 上 1 有 多 外国 视频 字幕 翻译 组 | 正解：网 上 有 多 外国 翻译 视频 字幕 组
  ↳ WER: 0.33
Epoch 13 | Train Loss: 1.08 | Dev Loss: 2.09
S003524_P0002_T00 → 預測：姐姐 开始 后 他 到 但是 做 玩 | 正解：姐姐 毕业 后 到 山 区 教 工作
  ↳ WER: 0.75
Epoch 14 | Train Loss: 1.01 | Dev Loss: 2.10
S006770_P0000_T00 → 預測：银行 恋爱 限制 自己 所以 人 不行 改变 | 正解：银行 卡 限制 自己 身份 人 不行 借
  ↳ WER: 0.38
```

圖 8 訓練週期圖

最後為了驗證本研究模型在 CSL-Daily 資料集上的表現，我們將結果與 Zhou 等人(2021)的 SignBT 論文[1]所報告的 CSLR 基準模型進行比較。根據其 Table 5 所示，最佳的 S2G 組合在測試集的 WER 為 32.2%。該論文發表於 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR 2021）。

Table 5. Evaluation of the S2G network combinations on WER (the lower the better).

S2G combination		PH2014T		CSL-Daily	
Sign Embedding	Encoder	Dev	Test	Dev	Test
I3D	Transformer	32.6	33.2	45.4	44.3
TIN	Transformer	26.2	27.5	36.1	35.7
BN-TIN	Transformer	23.0	24.1	33.6	33.1
BN-TIN	Conv1D	24.7	25.1	33.4	33.3
BN-TIN	Bi-GRU	**22.7**	**23.9**	**33.2**	**32.2**

圖 9 資料來源：Zhou et al.(2021) Table5

本研究採用的 S2G (Sign-to-Gloss)架構基於優化的 BN-TIN-Transformer 模型。在最終測試集上的性能評估結果顯示，平

均詞錯誤率（WER）為 37.99（共 1176 筆樣本），與作為比較對象的基準論文結果（平均詞錯誤率（WER）為 32.2）相比，絕對差距約為 5.7。

表格 5-1 本專題在驗證集和測試集上的表現

	TEST	DEV
WER	37.99	39.19

值得注意的是，本研究在特徵提取環節進行了重大改進：

1. 特徵源改變： 基準論文依賴 CNN 架構從原始影像中學習視覺特徵。

2. 本研究方法： 我們採用 MediaPipe 預先擷取標準化的人體姿態（Pose）關鍵點作為輸入特徵。

這一策略性變動，不僅大幅降低了模型在影像辨識上的訓練複雜度與計算資源需求，更有效地將研究重點聚焦於 Transformer 模型對時序骨架資訊的建模能力，從而驗證了本模型的有效性。

4.2 前後端架構圖

本系統的流程如圖所示，啟動後首先判斷輸入來源為即時鏡頭或影片檔，並從畫面中擷取影格。

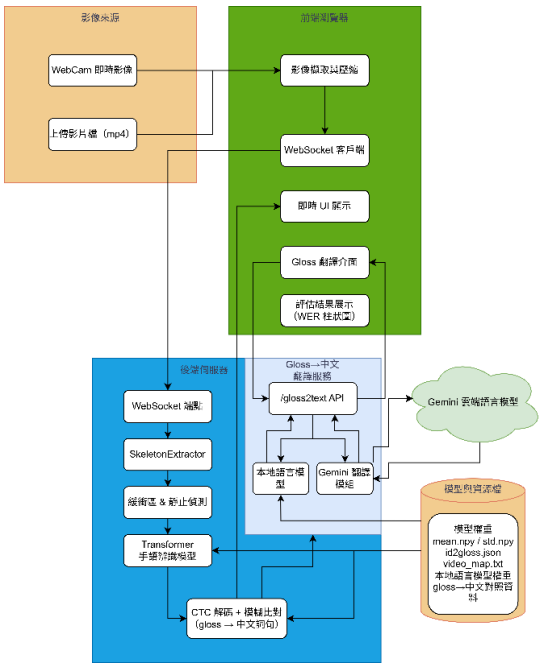

圖 10 前後端架構圖

4.3 即時影像辨識

本系統提供使用者即時手語辨識的功能。使用者可透過前端網頁開啟攝影機進行辨識，畫面中會即時顯示使用者的影像及模型辨識出的手語詞彙，並同時記錄歷史辨識結果。

圖 11 即時手語辨識+翻譯系統

4.4 影片上傳辨識

本系統提供使用者上傳手語影片的功能，前端在影片播放的同時，會將每一幀畫面即時傳送至後端伺服器。當影片播放結束後，後端模型會對整段影格序列進行推論，最終在前端的輸出介面中呈現辨識結果。

圖 12 影片上傳辨識+翻譯系統頁面

圖 13 模型性能比較

本系統整合手語 gloss 至中文句子的翻譯模組，於介面中提供 Gemini 模型與本研究自訓模型兩種翻譯方式。當系統輸出手語 gloss 辨識結果後，使用者可選擇

其中一種模型並按下「翻成句子」按鈕，即可產生對應的中文原始句子。

圖 14 用 gemini 模型翻譯

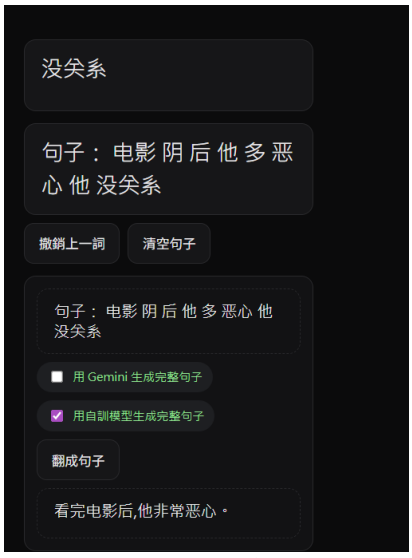

圖 15 用自己訓練模型翻譯

4.5 微調語言模型數據

在語言模型部分，本研究以預訓練模型 Mengzi-T5 為基底，利用約 10,000 筆手語 gloss－中文句子配對資料進行微調，建立專用的 gloss→中文翻譯模型。實驗中將資料依 8：1：1 分割為訓練集、驗證集與測試集，完成訓練後，於測試集上使用 sacreBLEU 計算 BLEU 分數，結果約為 87.46 分，顯示模型已具備將手語詞序列轉換為語意通順中文句子的能力。

圖 16 語言模型句子測試跟模型分數

第五章 結論與未來展望

5.1 結論

儘管最終模型的準確度與評估指標仍有不足，但整體系統已能完成從骨架擷取、模型推論到字幕顯示的完整流程。未來希望能將本研究應用於台灣手語辨識領域，結合在地化資料與更多樣的手勢樣本，以提升模型的泛化能力與準確性。

5.2 未來展望

若能在後續階段持續蒐集與擴充更多高品質的手語影片資料，特別是涵蓋更多手勢、句型與不同環境下的拍攝條件，將能使模型在詞彙辨識與句子生成上更為準確與穩定，並顯著提升系統於真實應用場景中的可靠性與泛化能力。

若硬體資源允許，可適度提高模型參數規模與容量，通常能帶來更豐富的特徵表徵與更穩健的時序關聯建模，預期可進一步提升辨識效能

第六章 專題影片

https://youtu.be/tmYT7g9VOpw

參考文獻

〔1〕Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li, "Improving Sign Language Translation with Monolingual Data

by Sign Back-Translation," IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
〔2〕Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li, "SignBERT+: Handmodel-aware Self-supervised Pre-training for Sign Language Understanding," IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI), 2023.
〔3〕A Chinese Continuous Sign Language Dataset Based on Complex Environm ents *Submitted on 18 Sep 2024*
〔4〕自監督學習在臺灣手語辨識上之應用研究 第 64 屆--民國 113 年
〔5〕HaGRID - HAnd Gesture Recognition Image Dataset *Submitted on 16 Jun 2022*