

國立陽明交通大學
數據科學與工程研究所
碩士論文

Institute of Data Science and Engineering
National Yang Ming Chiao Tung University
Master Thesis

使用深度學習進行基於影片的台灣手語辨識
Taiwan Sign Language Recognition for Video
Using Deep Learning Techniques

研究生：黃明翰 (Huang, Ming-Han)

指導教授：孫春在 (Sun, Chuen-Tsai)

中華民國 一一〇 年 六 月

June 2021

使用深度學習進行基於影片的台灣手語辨識
Taiwan Sign Language Recognition for Video
Using Deep Learning Techniques

研 究 生：黃明翰
指 導 教 授：孫春在 博士

Student : Ming-Han Huang
Advisor : Dr. Chuen-Tsai Sun

國立陽明交通大學
數據科學與工程研究所
碩士論文

A Thesis
Submitted to Institute of Data Science and Engineering
College of Computer Science
National Yang Ming Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in
Computer Science

June 2020
Hsinchu, Taiwan, Republic of China

中華民國 一一〇 年 六月

使用深度學習進行基於影片的台灣手語辨識

研究生：黃明翰

指導教授：孫春在 博士

國立陽明交通大學數據科學與工程研究所碩士班

摘要

隨著硬體設備的進步，深度學習在各領域的技術日漸成熟，例如在手機上使用人臉辨識取代指紋辨識成為新一代的生物辨識系統，或是像 Apple Siri、Amazon Alexa 等行動智慧助理，都讓生活更加的方便。然而，科技鮮少為聾啞人士們帶來便利性，很多技術和新產品都因為身障的關係而無法使用。因此本研究搭建一個台灣手語辨識模型，旨在使聾啞人士更無障礙的與他人溝通。

在過去的手語辨識研究中，大部分的研究都是使用穿戴式設備或深度攝影機來獲取人體的姿態資訊。然而，使用這兩種設備非常不方便攜帶且需要付出的成本太高，很難普及到社會之中。因此本研究將專注於只使用 RGB 影像來進行手語辨識，如此只要使用手機的相機或普通的 webcam 等輕量型設備就可以正確地與聾啞人士進行溝通。

本研究首創可供訓練的台灣手語資料集，參考台灣教育部的常用手語辭典，從中選出 40 個常用手語辭彙並自行拍攝手語影片。模型由兩個深度學習模型所組成，首先使用 DarkPose[10]對 RGB 影像進行全人體姿態估計，從影像中抽取出人體的關鍵點，再分別對 RGB 影像和人體關鍵點進行手語的辨識，最後再將兩個模型的預測加權平均得到最終的結果。實驗結果顯示，透過模型集成的方法可以對 40 個台灣手語達到 98.64% 的辨識率。

關鍵字：深度學習、手語辨識、台灣手語、3DCNN、人體姿態、模型集成

Taiwan Sign Language Recognition for Video Using Deep Learning Techniques

Student : Ming-Han Huang

Advisor : Dr. Chuen-Tsai Sun

Institute of Data Science and Engineering
National Yang Ming Chiao Tung University

ABSTRACT

In the past researches on sign language recognition, most of the researches used wearable devices or depth cameras to obtain the human pose information. However, these two devices are both very inconvenient to carry and the cost is too high. This research focus on using only RGB video for sign language recognition, so that we can use only lightweight devices such as mobile phone or webcams to correctly communicate with deaf-mute people.

This study builds first Taiwan Sign Language dataset for model training, which refers to the Taiwan Sign Language Dictionary of the Ministry of Education of Taiwan and selects 40 common vocabularies from it and shoots sign language videos ourselves. The model is composed of two deep learning models. First, DarkPose [10] is used to estimate the whole body pose of the RGB video, and the key points of the human body, including body, hand and face, are extracted from the image. The weighted average of the predictions from the two models yields the final result. The experimental results show that the recognition rate of 40 Taiwanese sign languages can reach 98.64% through the method of model ensemble.

Keywords: Deep Learning, Sign Language Recognition, Taiwan Sign Language, 3DCNN, Human Pose, Model Ensemble

誌謝

回首碩班這兩年，很榮幸能加入孫教授的實驗室，真的很感謝教授除了在研究上以及留學申請上給予我非常多的指導和幫助，每周教授在咪挺中分享的內容也都使我受益良多。感謝口試委員們的建議，幫助我將研究做得更加完善。感謝宣閔學長、實驗室夥伴，冠群、偉綸、阿勝、思妤、子軒在我的碩士兩年中添加了許多色彩。

特別感謝妹妹、台北王先生、二稗、媽媽、弟弟參與錄製手語影片，幫助我完成研究中非常重要的一個部分，沒有你們的二話不說地幫忙不知道實驗什麼時候才能做完。感謝 Jack 在我沒日沒夜寫論文時凱瑞我公司的事情，希望你未來申請研究所也能順利申請到好學校；感謝楷楷借我一個地方住，讓我在出國前能在台北爽；感謝四個肥宅天天講幹話陪我玩遊戲做笑果，希望大家的體重不要再直線上升；感謝女友毫無怨言地陪伴，才能讓我在期限前把論文完整的生出來；感謝黃知恩，我會趕快幫妳買新衣服。最後感謝家人無條件的支持與信任，在被媽媽通知今年能領兩張畢業證書時還真的沒把握能夠畢業 QQ

謝謝正在讀這篇論文的你，希望本篇研究能真的幫助到需要幫助的人。

目錄

摘要.....	I
ABSTRACT	II
誌謝.....	III
目錄.....	IV
圖目錄.....	VI
表目錄.....	VII
一、緒論.....	1
1.1 研究動機.....	1
1.2 研究背景.....	2
1.2.1 手語.....	2
1.2.1.1 手語的區域性.....	2
1.2.1.2 手語結構.....	3
1.2.2 感測式設備.....	4
1.2.3 深度攝影機.....	6
1.2.4 人體姿態估計.....	7
1.2.5 神經網路模型.....	8
1.3 研究目標.....	10
1.4 研究重要性.....	10
二、文獻探討.....	12
2.1 手語辨識.....	12
2.2 人體姿態估計.....	13
2.3 DARKPOSE.....	15
三、研究方法.....	18
3.1 研究架構.....	18
3.2 資料蒐集.....	18
3.3 資料前處理.....	20
3.3.1 資料分布情形.....	20
3.3.2 影片轉換為 RGB 圖片.....	21
3.3.3 訓練集及驗證集.....	22
3.4 全人體姿態估計.....	23
3.4.1 全關鍵點.....	23
3.4.2 特徵關鍵點.....	24

3.5	預測模型	25
3.5.1	關鍵點模型	25
3.5.2	RGB 模型	26
3.6	模型集成	26
四、	實驗結果與討論	28
4.1	實驗環境	28
4.2	關鍵點模型	28
4.2.1	全關鍵點	28
4.2.2	特徵關鍵點	30
4.3	RGB 模型	31
4.4	模型集成	33
五、	結論與未來展望	37
5.1	研究結論	37
5.2	研究限制	37
5.3	未來展望	38
	參考文獻	39
	附錄一	42

圖目錄

圖 1 在衛服部的疫情直播中，手語老師沒有配戴口罩是為了更明確地傳達非手勢訊息.....	3
圖 2 相似的台灣手語，(左)台灣，(右)玉米	4
圖 3 (左)感測手套設備及(右)加速器。	5
圖 4 (左) ZHANG ET AL. 的自製裝置及(右) KIM ET AL. 的肌肉貼片位置。	5
圖 5 MICROSOFT KINECT 應用 PRIMESENSE 的 LIGHT CODING 技術來獲取 3D 深度圖像。	6
圖 6 OPENPOSE 全人體姿態估計，包含臉部、手部及肢體的關節點。	8
圖 7 RNN 模型架構示意圖。	9
圖 8 (左)傳統 2D-CNN 模型結構與(右)3D-CNN 模型結構示意圖。	10
圖 9 3D-CNN 應用於手語辨識的模型結構，此模型使用了 3 層 3D-CNN 以及 2 層部分取樣層。	13
圖 10 被觀察者的右手因為跟身體的部分重疊造成右手部件被不正確的匹配。	14
圖 11 DEEPOSE 的 DNN 網路架構以及回歸器示意圖。	14
圖 12 熱圖預測方法是在關節點生成一個 2D 的高斯機率圖。	15
圖 13 DARK 針對熱圖解碼為座標點時的優化。	17
圖 14 研究架構示意圖。	18
圖 15 台灣手語資料集製作拍攝示意圖。	20
圖 16 40 個手語分類資料分布情形。	20
圖 17 影片幀數分布圖。	21
圖 18 影片轉換為 RGB 圖片流程圖，(A)從影片中抽取出全身關鍵點，(B)對影片做大小的裁切和縮小， (C)根據基準幀數對影片做影片長度的裁切或補正。	21
圖 19 手語辭彙「我們」完整動作轉換成連續圖片。	22
圖 20 OPENPOSE(左)在畫面中看不到手肘時偵測不到手部姿態，而 DARKPOSE(右)則不被影響。	23
圖 21 從影像中抽取的人體關鍵點及其編號。	24
圖 22 模型輸入示意圖，藍點為人體關鍵點，藍線為骨架，綠線為時間維度對應點的連結。	25
圖 23 RESNET2+1D 模型架構。	26
圖 24 121 個全關鍵點訓練時(左)驗證資料的 TOP1、TOP3、TOP5 準確度趨勢和(右)LOSS 的下降趨勢。 ...	29
圖 25 全關鍵點訓練下的混淆矩陣。	29
圖 26 特徵關鍵點訓練時(左)驗證資料的 TOP1、TOP3、TOP5 準確度趨勢和(右) LOSS 的下降趨勢。	30
圖 27 特徵關鍵點訓練下的混淆矩陣。	31
圖 28 3D-CNN 訓練時的準確度趨勢，藍線為訓練準確度，橘線為驗證準確度。	32
圖 29 3D-CNN 訓練時的 LOSS 下降趨勢，藍線為訓練 LOSS，紅線為驗證 LOSS。	32
圖 30 RGB 模型訓練後的混淆矩陣。	33
圖 31 模型集成後的混淆矩陣。	34

表目錄

表 1 比較 DARKPOSE 與其他先進的 HPE 模型，以 COCO TEST-DEV 資料集為基準。.....	16
表 2 本研究蒐集的 40 個台灣手語與其分類。.....	19
表 3 人體特徵關鍵點。.....	25
表 4 比較各階段模型與模型集成後的準確率。.....	34
表 5 不知道與其相似手型：不是、痛苦、了解。.....	35
表 6 立刻、不是與謝謝比較。.....	36

一、緒論

1.1 研究動機

因應多元化的社會，在全球性的新聞或直播報導的角落有即時的手語翻譯已經成為標準配置，而在日常生活或影視媒體上，我們也經常看到聾啞人士們互相使用手語溝通、使用手語與會手語的普通人士對話或藉由手語翻譯與其他無障礙的人士進行互動的畫面。然而，這些身障人士的家屬經常需要付出額外的時間成本或人力資源來幫助他們打理生活，例如學習手語或額外聘請會手語的看護；或是這些身障人士在獨自遭遇到緊急狀況時，由於手語並不是一個普及的語言，造成他們無法與不會手語的人進行溝通而無法獲得即時適當地處理。因此，本研究將使用最新的電腦視覺與深度學習技術來進行手語辨識(Sign Language Recognition, 以下稱 SLR)，使得這些聾啞人士可以在不需手語翻譯的陪同下與其他人進行互動，並且在緊急情況下普通人也可以給予這些聾啞人士適當的協助。

目前在手語辨識的技術上我們較常看到的應用主要是使用穿戴式的設備或是深度攝影機來獲取手語者的資料[31]。雖然這些設備獲取的額外資料將提供很好的輔助，大大降低了辨識的難度，然而這些設備並不具備便攜性與普遍性，若是想普及化到社會或個人的使用上，除了需要聾啞人士配合穿戴額外厚重的配備來進行辨識，也需要耗費龐大的金錢來購買這些額外的設備，非常不適合日常的使用且花費的成本非常高。因此，本研究將專注在如何單純地透過影像資訊，使用全人體姿態估計的 Darkpose[10]模型抽取出手語者的全身骨架資訊，再結合預先訓練好的神經網路模型來進行手語的辨識。再者，因為手語的地區性，目前尚無可供模型訓練的台灣手語資料集。本研究將首創台灣手語資料集，供本研究及未來的相關研究訓練。最後因為此模型不需特殊的設備即可完成資料的獲取，此系統即可應用到所有人都可以使用的便攜式輕量級設備，例如手機 app 或簡易 webcam，達成手語者與其他人的無障礙溝通。

1.2 研究背景

1.2.1 手語

手語(Sign Language)是一種使用視覺模式來進行溝通的語言，是聽覺能力障礙者及言語能力障礙者用來溝通的重要工具。其主要使用手勢的變化並搭配上半身的肢體來打出手語，再結合臉部表情來傳達手語者更深層的語意或語氣。

1.2.1.1 手語的區域性

手語就如口語一樣也有區域性，各地區會依據自己當地的語言發展出一套不同文法及不同詞彙表達方式的手語系統，其主要分為三種語系：法國手語系、中國手語系以及日本手語系，同個語系中的手語因在發展時互相影響，手語的手勢及文法都比較類似。法國手語系包含法國手語、美國手語，法國手語為目前可考察到的始祖手語，而後影響了美國的手語發展；中國手語系主要在中國大陸使用，更可以細分為南北兩大類，並有一個類似於拼音的系統；台灣手語則和日本手語及韓國手語等三種手語一起被歸類於日本手語系中[2]。

台灣手語(Taiwan Sign Language, TSL)是中華民國地區的聾啞人士用以溝通的主要工具，其發展主要是源自日本手語，日治時期日本在台北及台南分別設立了啞學校並教導手語，使得日本手語在台灣開始普及，戰後日本退出台灣，大部分的聽障教育開始加入大量的中國手語詞彙，也因此跟原本的手語混合發展出了台灣手語。因為這樣的發展，台灣手語開始漸漸分歧為北部手語及南部手語，除了分別源自於日本手語中的東京系統及大阪系統，北部手語也受到較多中國手語的影響。教育部為了減少溝通上的困難，於民國 69 年 9 月正式頒布手語畫冊，統一台灣各地的手語。2018 年 12 月 25 日，《國家語言發展法》正式立法將台灣手語列為中華民國的國家語言，2019 年教育部的線上手語辭典正式啟用，收編台灣手語的數字手勢、音標手勢、詞彙手勢，更有會話與短文的手語範例。

1.2.1.2 手語結構

手語的組成包括了手勢的變化以及非手勢訊息，手勢的變化代表了不同的詞語，而非手勢訊息指的是在打手語時所伴隨的肢體動作以及面部表情，台灣的非手勢訊息主要有抬眉皺眉、點頭搖頭、唇部動作、身體動作以及頸部動作[1]。如同說話時口語中有抑揚頓挫，非手勢訊息是在傳達手語者的情緒和語意，在打手勢的同時加上非手勢訊息可以加強某一個手語，例如唇部吸氣的動作可以加強「吃驚」；非手勢訊息也可以代表語氣詞，例如「ㄟ」的嘴型可以代表為「喔」。非手勢訊息同時也會影響句子的類型，例如疑問句的表示方式為側頭或抬眉。



圖 1 在衛服部的疫情直播中，手語老師沒有配戴口罩是為了更明確地傳達非手勢訊息

資料來源：<https://www.youtube.com/watch?v=HtZxzPLA5SU>

在手語中以手勢動作為最主要的傳遞訊息方式，不同的手勢動作代表不同的詞彙與意義，而手勢動作更可以細分為手形(handshape)、位置(location)、動作(movement)以及手心或指尖朝向(orientation)四個要素[1]。手形指的是單手或雙手手部的形狀和樣態，目前台灣手語的手形共有 56 種；位置指的是比手形時手所在的位置，位置的垂直範圍為頭頂至腰部，且主要在身體前面，在台灣手語中共有 22 種不同的手形位置；動作指的是打手語時手勢上的變化或轉動，例如手指彎曲或雙手畫圈等；朝向則是指打手語時手心或指尖的朝向。如果兩個手語

辭彙的位置、動作跟朝向都相同，但手形不同就會造成詞彙的不同，例如「台灣」和「玉米」都是手放口前左右轉動，但一者是握拳，另一者則是五指彎曲。



圖 2 相似的台灣手語，(左)台灣，(右)玉米

資料來源：<http://140.123.46.77/TSL/>

1.2.2 感測式設備

感測式設備是指藉由在身體的各個部位裝上感測器來蒐集使用者個關節資訊，因為獲得的資料直接是來自身體的關節點，因此獲得的身體運動軌跡是最穩定可靠的。感測式設備在手語辨識方面的應用可以分為感測手套(data glove)、肌電圖技術(Electromyography, EMG)以及無線網路和雷達(Wifi and radar)共三種[22]。

感測手套是指在各手指的關節點都設置固定的感測器，並裝置於手套上以利配戴。使用感測手套時通常會搭配手腕處的加速器(accelerometer)來提供手的旋轉和移動等資訊，以此得到更精確的整體手部移動軌跡[27]。Liang et al.[33]的研究就使用感測手套來獲得手語者的手指、位置、角度及動作資料，並在辨識 250 個台灣手語單詞達到了 89.5%的準確率。

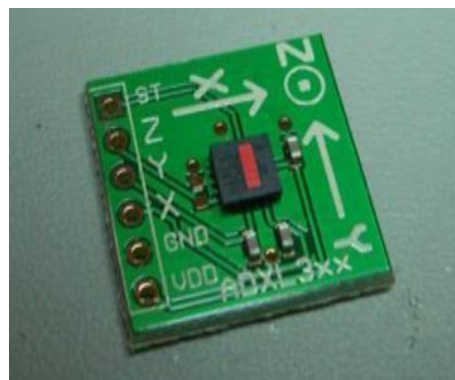
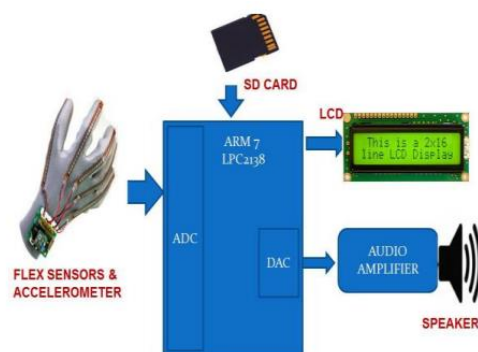


圖 3 (左)感測手套設備及(右)加速器。

資料來源：Lokhande et al.[27]

肌電圖技術是指以肌電波紀錄器紀錄黏貼在皮膚上的電極貼片或是插入肌肉之中的電極傳回的資料來得到肌肉組織的電氣活動而獲得感測部位運動資料的技術。在過往的研究中，Zhang et al.[39]使用一個自製的裝置—五頻道的 EMG 搭配一個三軸加速器來獲取手語者的手部資訊並在辨識 72 個中國手語上獲得了 93.1%的準確率，而 Kim et al.[17]則是使用了貼在手臂上的肌肉貼片來獲得手指及手部的運動軌跡，包含握拳、水平移動及旋轉等，並在辨識 20 個不同的手勢上獲得了 94%的準確率。

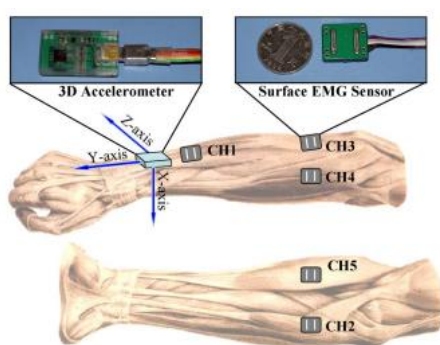


圖 4 (左) Zhang et al. 的自製裝置及(右) Kim et al. 的肌肉貼片位置。

資料來源：Zhang et al.[39] & Kim et al.[17]

無線網路技術則是指以 Wifi 訊號為基底開發出來的辨識動作技術，稱為 Wised [30]。Wised 主要是使用都卜勒位移(Doppler shift)的原理，當受測者在佈滿 Wifi 訊號的房間中作出不一樣的動作時，Wised 會接受到不同的 Wifi 反射訊號，並以此來判

斷受測者的動作。因為此技術是藉由 Wifi 傳遞，因此不需要配戴任何裝備，也突破了使用像 Kinect 或 webcam 等光學式裝置時受測者不能被遮蔽的限制。研究者在提出此項技術時也針對了自定義的九個動作做了動作辨識，結果獲得了還不錯的 94%準確率。

1.2.3 深度攝影機

以動作分析來說，使用攝影機拍攝受測者是最直覺且最無負擔的方式。深度攝影機通常是在普通的 RGB 彩色攝像頭旁再追加多個鏡頭，以多個攝影機來計算場景中物件的深度，使得我們可以同時獲得 RGB 彩色影像以及 3D 深度影像。常見的深度攝影機有 Microsoft Kinect 以及 Intel RealSense depth camera，其中以 Kinect 為攝影機的研究又更為普遍[14]，Kinect 的鏡頭群是以中間的 RGB 彩色攝影機以及由兩旁的紅外線發射器和紅外線 CMOS 攝影機一起構成的 3D 深度感應器，透過 PrimeSense 的 Light Coding 技術對測量空間進行編碼，最後運算獲得深度圖像。

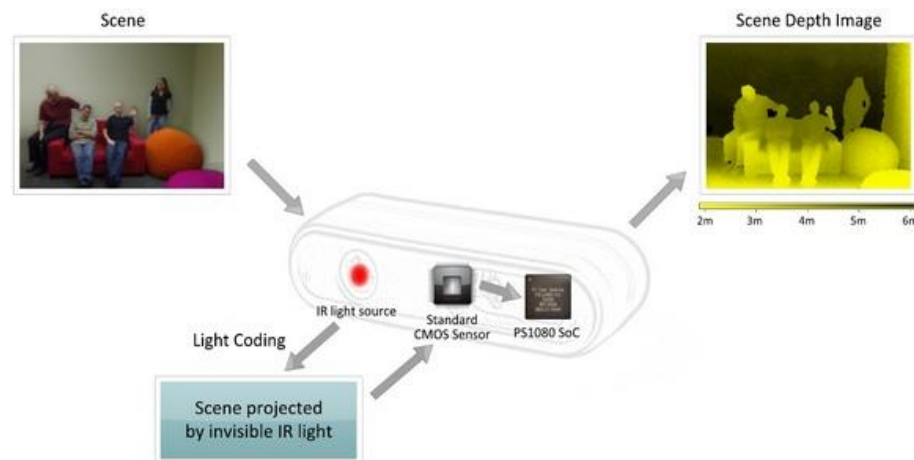


圖 5 Microsoft Kinect 應用 PrimeSense 的 Light Coding 技術來獲取 3D 深度圖像。

資料來源：<http://www.primesense.com/CH/?p=739>

深度攝影機獲得的 3D 深度影像對於動作識別的領域來說無疑是一強大的利器，有了 3D 深度影像除了可以迅速的將觀測對象從複雜的背景中分離出來，成功解決電腦視覺領域中的一大難題，也可以透過公開的 Azure Kinect SDK 進行實時的人體追蹤，並從中直接獲得人體的骨架及關節資料。Pigou et al.[21]的研究就使用 Kinect 來獲取

資料，在使用深度的影像進行背景分離及雜訊處理後，和原本的影像一起進行辨識達到了 91.7% 的準確率。在 Huang et al. [16] 的研究中也透過 Kinect 獲得的深度影像，經過 SDK 取得人體骨架資訊再和深度圖像、RGB 圖像一起進行辨識，對於 25 個手語達到了 94.2% 的準確率。

1.2.4 人體姿態估計

人體姿態估計(Human Pose Estimate, HPE)是電腦視覺領域中非常火熱的一個研究問題，目的是定位圖像或影片中人體的關節點(例如手肘、膝蓋、頭部等)，成功定位後我們便可以將關節點互相連線成為人體的姿態，進而分析人體的動作以及行為。因為關節點互相連線後的圖看起來就像是人體的骨架(skeleton)，也時常有人稱姿態資料為骨架資料。

通常在 HPE 的領域中會將人體的部位分為四個部分並視為四種不同的任務，分別為肢體(body)、臉部(face)、手部(hand)以及足部(foot)。肢體任務為偵測人體四肢的關節，以身體為初始點，包括頭部、肩膀、手臂、手腕、膝蓋以及腳踝；臉部任務為偵測人臉部五官的變化，包括眉毛、眼睛、鼻子以及嘴巴；手部任務為偵測五指各關節，包括手掌及手指的第一關節、第二關節和第三關節；足部任務則專注在腳跟及腳趾的位置。

對於手語者來說，除了肢體的移動外，手指的細節變化以及臉部的表情都會影響到一個手語要表達的意義，因此在本研究中將會選用偵測範圍涵蓋肢體、手部以及臉部的 HPE 模型，此類模型稱為全人體(whole body)姿態估計模型。由卡內基梅隆大學(Carnegie Mellon University, CMU)在 2017 年所開發的 OpenPose[42]則是非常成熟的全人體姿態估計模型之一。

OpenPose 是第一個實時的多人全人體姿態估計模型，他提供了包含肢體、手部、臉部以及足部共 135 個關節點的偵測並在 COCO test-dev 資料集獲得了 64.2 的

AP[41]。然而 OpenPose 使用了多個不同的模型分別針對臉部、手部及肢體做 HPE 來實現全人體姿態估計犧牲掉了非常多的效能，隨著這幾年的演進，本研究將使用 2020 年在 COCO test-dev 資料集刷新了 AP 紀錄到 77.4 並且效能更好的 DarkPose[10]來作為全人體姿態估計模型，此模型在使用 Top-down¹方法及熱圖技術的 HRNet[19]上作為基準模型，改進了熱圖座標點轉換的問題來達到更好的模型效能。

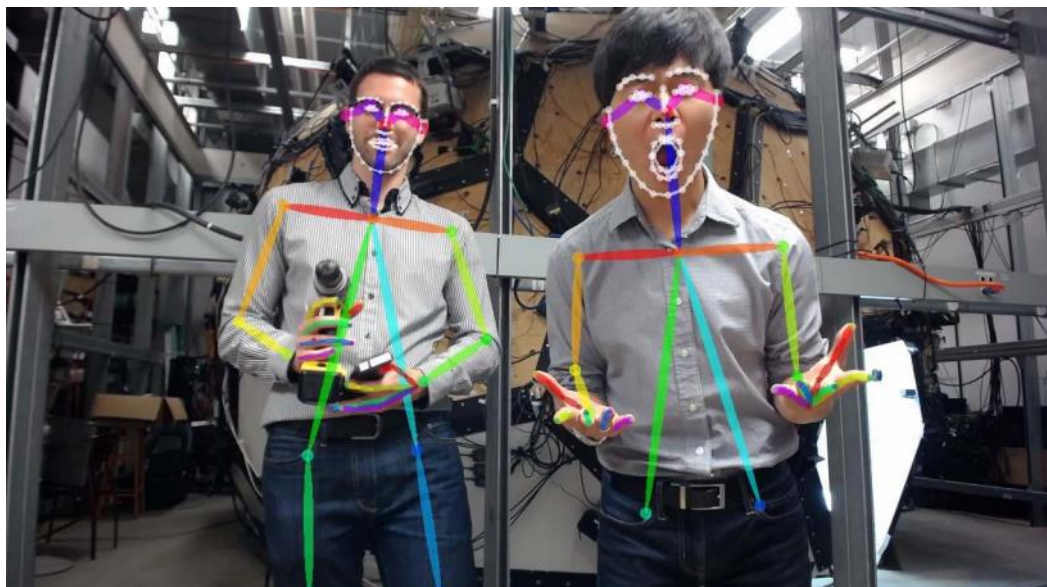


圖 6 OpenPose 全人體姿態估計，包含臉部、手部及肢體的關節點。

資料來源：OpenPose[41]

1.2.5 神經網路模型

「動作」是手語的組成四要素之一，手語並不是只是定格的手勢，而是一個連續的手勢變化，因此在此項任務中，我們無法單純使用在處理影像上表現突出的 2D 卷積神經網路 (2D Convolutional Neural Network, 2D-CNN) 來對手語進行辨識，而是要藉由遞歸神經網路 (Recurrent Neural Network, RNN) 的幫助來學習手勢的前後動作變化。RNN 的運作原理為模型在每一個時刻 t 都會結合此時刻的輸入 X_t 和前一個時刻運算的結果來得到輸出 O_t ，而過去資料影響力會隨著距離此時刻的拉長而慢慢衰減，因

¹ 先檢測出圖像中的人，再針對每個人預測關鍵點。

為這樣的特性，RNN 時常被應用在降雨量預測、股價預測等時序任務上。在 Shi et al. [5] 的研究中就運用了 CNN+RNN 的模型架構來辨識手語，他們首先使用全連接的 CNN 模型來抽取出影像的特徵圖(feature map)，之後將抽取出來的特徵圖當作每一個時刻 RNN 的輸入並計算得到最終的預測。

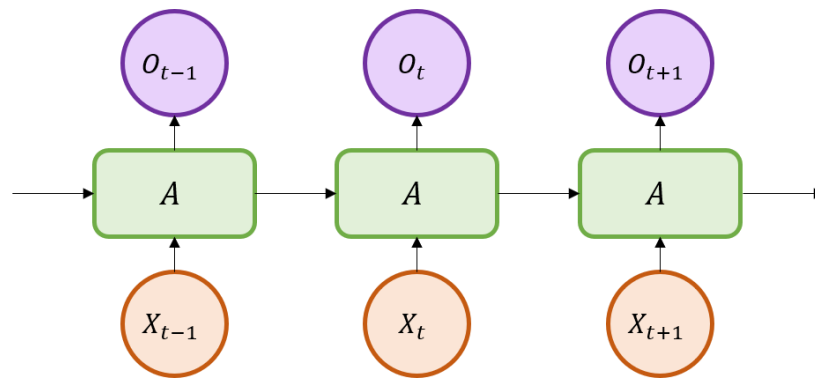


圖 7 RNN 模型架構示意圖。

資料來源：本研究製作。

在動作識別的領域中除了使用 CNN+RNN 這種組合模型外，Ji et al. [34] 提出的 3D-CNN 模型改進了 2D-CNN 無法考慮到時序的限制，將影像處理和時序分析融合到一個模型之中。在傳統的 2D-CNN 中，模型對於影片中每一幀的空間維度進行卷積而得到一個特徵圖；而在 3D-CNN 中，透過將多個連續的幀組合成一個立方體，模型就可以對立方體進行 3D 卷積計算而得到特徵圖，也就是說，每一個特徵圖都是源自於影片中的多個連續的幀，進而同時獲得空間以及時序的資訊。[17, 25] 就使用 3D-CNN 來進行手語辨識並分別獲得了 94.2% 和 97.7% 的準確率。

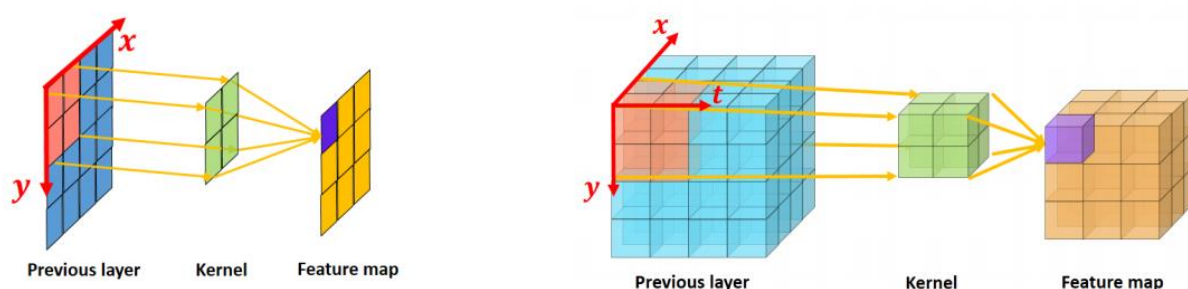


圖 8 (左)傳統 2D-CNN 模型結構與(右)3D-CNN 模型結構示意圖。

資料來源：Huang et al.[16]

1.3 研究目標

過去由於技術上的限制，我們只能通過穿戴式設備來獲取人體的骨架資訊或使用深度攝影機來獲取深度資訊進而獲取人體骨架資料，再利用骨架資訊結合 RGB 影像資訊來針對手語者比出的手語進行辨識，且大部分的研究都是針對其他手語語系的詞彙，除了資料集的缺乏，使用台灣的手語做為目標的研究實在是少之又少，因此本研究旨在打造一個屬於台灣手語的低門檻手語辨識系統，將台灣手語更加普及化。研究的主要目的主要為以下兩點：

1. 根據教育部常用手語辭典，精選 40 個台灣手語詞彙生成資料集並針對此 40 個手語進行辨識。
2. 單透過影片的 RGB 影像資訊，捕捉影片中的人體特徵資訊，進而完成手語者的手語辨識。

1.4 研究重要性

在這個人工智慧快速發展的時代，語音辨識、自動翻譯、行動助理等技術日漸成熟，許多人都因為這些技術的發展而獲得更方便的生活，但身為社會少數的聾啞人們卻遲遲無法從中獲利，減輕生活上的負擔。根據台灣衛生福利部統計處統計，2020 年

全台灣聽覺機能障礙者為 124,825 人，聲音機能或語言機能障礙者為 15,462 人，總計 140,827 人，佔據全台總人口數約 0.6%。聾啞人雖然在台灣的人口比例上占比不高，但基數卻有將近 14 萬人，藉由手語辨識系統，聾啞人不用再隨時需要手語翻譯陪同，只要對方擁有手機，聾啞人就可以與不會手語的人進行對話；或是只要店家或安養院等設施有裝設此套系統，就算服務人員不會手語，他們依舊可以給予聾啞人士適當地協助。而對於他們的家庭來說，聾啞人的親戚也不再需要另外學會手語就可以與聾啞人溝通，促進家族的和睦。

本研究期望提出一個成本更低、更普及且不失準確性的新模型，以台灣手語為資料集，作為專屬於台灣的手語辨識的應用，不但可以為這 14 萬名聾啞人的生活帶來更多的便利性，更可以擴展到他們的家庭，為更多人帶來幫助。同時也促進手語領域的發展和普及化，讓更多人可以無障礙地與聾啞人進行互動。

二、文獻探討

2.1 手語辨識

手語辨識在電腦視覺的領域中包含了很多不同領域的問題，[22]整理出近年來各種不一樣方法的優缺點以及遇到的挑戰，而在資料獲取上面，目前廣泛使用的技術有使用帶有偵測器的手套、accelerometer、Microsoft Kinect 和 Intel RealSense(有深度鏡頭)以及相機(webcam 或多視角相機)[31]，而目前的大宗還是使用深度鏡頭來獲得 3D 的資料[14]。

在手語辨識系統中，首先是背景分離的問題。前處裡的方法很多種，[20]從皮膚來思考，透過膚色將手部及臉部從背景分離出來並且可以持續追蹤手部的動作，而這兩個部分也是手語中非常重要的一環。[7] 提供了一個很好的方法讓我們從複雜的背景中將人體以 blob 的形式分離出來，以此可以除去很多雜訊，達到更好的辨識率。但若使用 Kinect，結合 Kinect 給的 RGB 和深度資訊，可以做到很好的背景分離，再使用 CNN 進行訓練就能有很好的結果[16, 21]。

在視角方面，相對於第三人稱視角，也有研究使用第一人稱視角來設計他們的系統，但得到的資訊量也因此變少且辨識更困難，因此需要其他器具(手環)來輔助[13]。

背景分離後，在辨識的細部就是各個手指的辨識問題。手指的動作在手語中是非常重要的一環，[28]專注於手指的辨識，透過參考點(reference point)得知手語者是伸出哪幾根手指。因為同一個手勢在不同旋轉平移之後會看起來完全不一樣，若我們有深度的資料，即使是手指互相擋住或混淆，我們可以透過深度的資料得到他們 3D 的位置並預測出各個手指的骨架，而剩下的就是如何辨識手部的位置和翻譯[24, 38]，或是使用現有的技術(Intel RealSense depth camera)直接獲取手部骨架的資料，進

而辨識出手勢[29]。跟文字一樣，手語中也會有看起來的很類似的手勢造成辨識上的混淆，[3]使用兩層的 multi-layered random forest(MLRF)在處理這個問題上達到了更好的辨識率，一層負責偵測手部位置，另一層則負責辨識手的動作。

在辨識的模型上面，因為隱藏式馬可夫模型(Hidden Markov Model, HMM)是一個分析連續動作的傳統方法，[37]就藉由使用 HMM 來追蹤手的動作來辨識 40 個英文單詞，但忽略了細部的手指動作。對於目前非常火熱的深度學習模型，直接使用 CNN 對於手語者比出一個單詞的影片已經可以達到很高的辨識率[11]，但因為手語是連續性的資料，RCNN 可以起到更好的效果，但若沒有足夠多的資料，此模型就很容易過度擬化，造成效能下降[32]。若我們能夠擁有 3D 骨架的資料，使用 CNN+LSTM 模型也可以對於 3D+時間連續的動作做出很好的辨識[15]。[34]提出的 3D-CNN 模型改進了傳統 CNN 模型無法參考時序資料的限制並廣泛地被應用到動作辨識的領域上，[8, 17]很好地應用了 3D-CNN 模型來進行手語辨識。

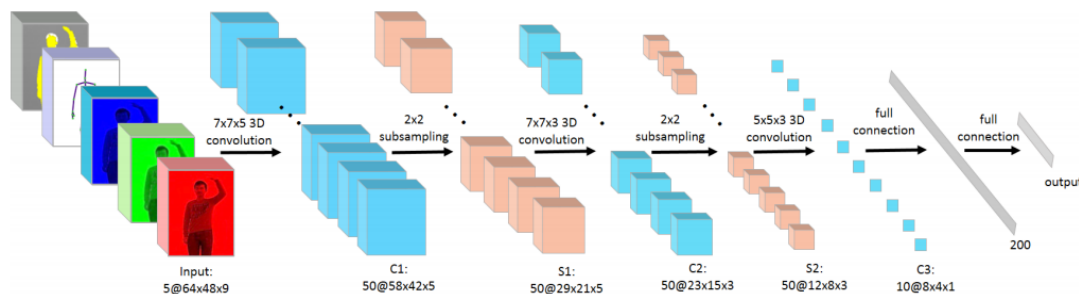


圖 9 3D-CNN 應用於手語辨識的模型結構，此模型使用了 3 層 3D-CNN 以及 2 層部分取樣層。

資料來源：Huang et al.[16]

2.2 人體姿態估計

早期 HPE 的研究方向為將人體視為部件的組合而非偵測人體的關節點，在 Felzenszwalb et al.[26]的研究中，先透過背景分離將人體從背景中抽離出來並得到一個二元的圖像，然後再將人體的部件(方框)透過匹配的方式配對到人的肢體上來獲得姿態估計的結果。此方法的限制是由於肢體物件的匹配是基於背景分離後的二元圖

像，當肢體互相重疊時從二元的圖像中看不出被覆蓋的肢體正確的位置，因而無法將部件正確的配對到肢體上。

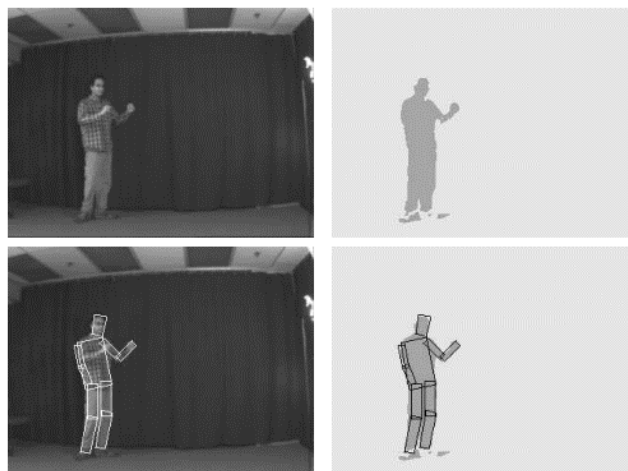


圖 10 被觀察者的右手因為跟身體的部分重疊造成右手部件被不正確的匹配。

資料來源：Felzenszwalb et al. [26]

在深度學習迅速發展後，DeepPose[4]是第一個在 HPE 領域應用深度學習的重要研究，打破傳統研究方法的限制也得到了突破性的結果。DeepPose 的模型是使用一個七層的深度神經網路 DNN 來對圖像進行辨識，並連接一個輸出為 (x, y) 座標的最終層作為關節點的位置。除此之外，DeepPose 還加入了一個回歸器將預測更加精細化，此回歸器會將每一個階段預測的結果座標送回此階段的原始圖像，再根據這個座標對原始圖像進行裁切後再送入下一個階段，以此讓模型可以學習到更細節的特徵，進而獲得更精準的結果。



圖 11 DeepPose 的 DNN 網路架構以及回歸器示意圖。

資料來源：DeepPose[4]

然而，DeepPose 中的 DNN 雖然使用了池化層和部分取樣層來減少計算量和過度擬化的問題，卻也降低了預測關節點位置的精準度，另外，座標回歸方法也因為座標點只關心二維的局部座標位置，缺少空間及與環境相關的資訊，使得模型無法正確地訓練及學習且在關節重疊的部分表現不佳而導致效能下降。這些缺點使得往後的研究改為使用熱圖(heatmap)預測的技術。

熱圖預測技術第一次是在 Tompson et al[18]的研究中提出，此方法的核心思維是對於圖像並行多解析度的處理以實現滑動窗口(sliding window)探測器來尋找目標關節點，對於每一個關節點會生成一個熱圖，此熱圖以目標關節點位置為中心形成一個二維的高斯分布。這樣的高斯分布使得模型在訓練時更可以藉由學習關節點周遭的環境讓模型在複雜的背景下或當關節點被遮擋、重疊時有更好的表現。許多近年的主流先進研究如 CPN[34]、SimpleBaseline[32]、HRNet[19]等皆為應用熱圖預測技術的模型。

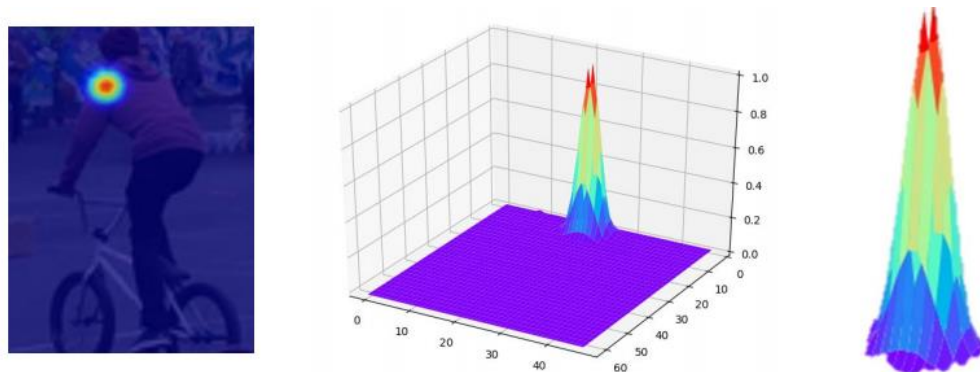


圖 12 熱圖預測方法是在關節點生成一個 2D 的高斯機率圖。

資料來源：DarkPose[10]

2.3 DarkPose

DarkPose[10]是由中國電子科技大學在 2019 年 10 月 14 日針對熱圖技術進行優化後並分別在 COCO 和 MPII 兩個資料集上驗證訓練而推出的一個與模型無關的插件，此技術提升了所有現階段在全人體姿態估計模型的效能，其中將 HRNet 作為基礎模型時

有最好的結果。此研究發現在將熱圖和關節座標點互相轉換時對於 HPE 的訓練及準確度有非常大的影響，因此他們分別對於將熱圖解碼為座標點以及將座標點編碼為熱圖這兩項任務進行計算上的優化並推出全新的 DARK (Distribution-Aware coordinate Representation of Keypoint)算法。

Method	Backbone	Input size	AP	AP ⁵⁰	AP ⁷⁵	AP ^M	AP ^L	AR
Bottom-up								
Openpose [42]	-	-	61.8	84.9	67.5	57.1	68.2	66.5
MultiPoseNet[23]	-	-	69.6	86.3	76.6	65.0	76.3	73.5
Top-down								
G-RMI [12]	ResNet-101	353x257	64.9	85.5	71.3	62.3	70.0	69.7
CPN [40]	ResNet-Inception	384x288	73.0	91.7	80.9	69.5	78.1	79.0
SimpleBaseline [6]	ResNet-152	384x288	73.7	91.9	81.1	70.3	80.0	79.0
HRNet [19]	HRNet-W48	384x288	77.0	92.7	84.5	73.4	83.1	82.0
DARK [10]	HRNet-W48	384x288	77.4	92.6	84.6	73.6	83.7	82.3

表 1 比較 DarkPose 與其他先進的 HPE 模型，以 COCO test-dev 資料集為基準。

資料來源：本研究製作。

在熱圖解碼為座標點的部分，首先他們發現模型預測出來的熱圖經常會有多個高峰，因此他們先藉由將熱圖和一個與測試資料分布相同的高斯核進行卷積來得到平滑化的新熱圖(圖 13a)，接著使用泰勒展開式來計算出關節點正確的位置 (圖 13b)，最後再將熱圖的高峰運算回到與原圖像相同的空間中並轉換為正確的目標關節座標點 (圖 13c)。而在關節點座標點編碼為熱圖的部分，跟解碼時有相同的量化問題。在傳統的編碼方法中，當原圖像被降低解析度時，關節點座標可能會被取整數造成誤差，而 DARK 則透過直接將熱圖的中心設在非量化位置來解決這個問題。因為座標點編碼通常是指將基準值(ground truth)編碼為熱圖後供模型學習，因此對於模型的訓練有大量的優化。

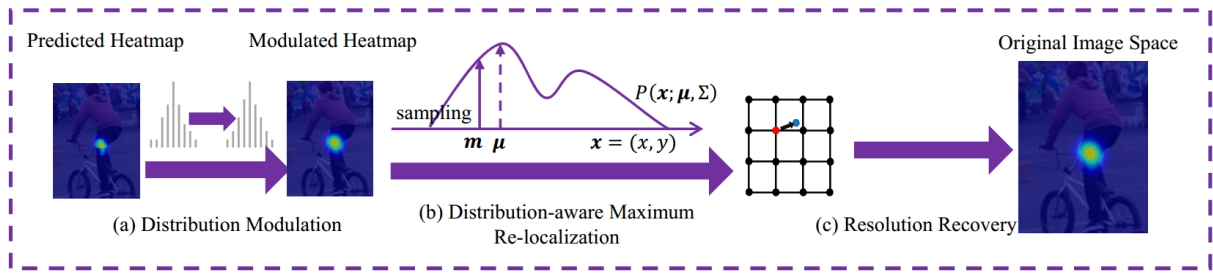


圖 13 DARK 針對熱圖解碼為座標點時的優化。

資料來源：DarkPose[10]

三、研究方法

3.1 研究架構

本研究提出一個台灣手語辨識系統，輸入為手語者比台灣手語的影片，輸出為 40 種台灣手語的預測結果。首先本研究將使用自行拍攝台灣手語的影片組成資料集，接著影片會經由 HPE 模型將身體、手部以及臉部的關鍵點抽取出來，同時將影片的 RGB 影像輸入到 3D-CNN 中進行訓練並得到影像部分的預測，最後再將兩者的結果加權平均輸出最後的台灣手語預測。

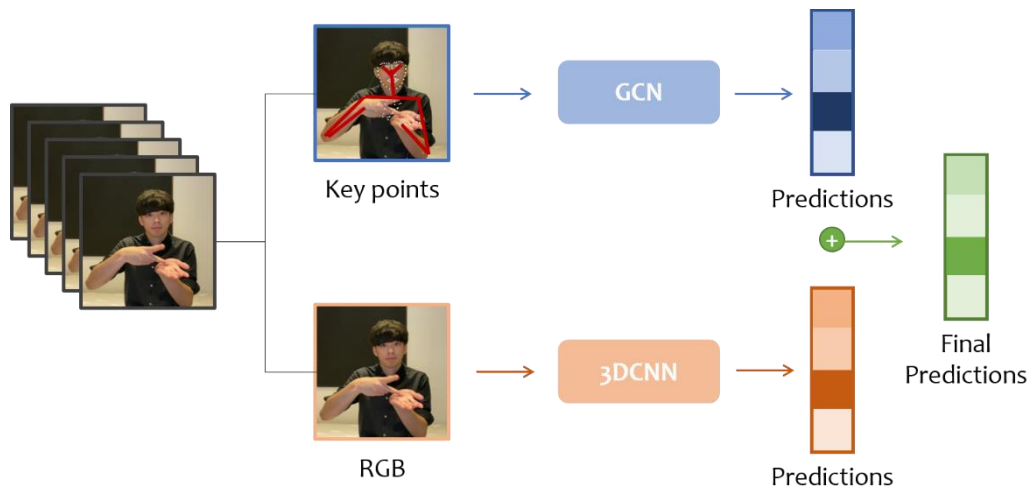


圖 14 研究架構示意圖。

資料來源：本研究製作。

3.2 資料蒐集

因為手語有地區性，每個地區的手語都不盡相同，手語資料集也非常地分散，目前市面上的手語資料集以美國手語(ASL)和中國手語(CSL)較為常見。對於本研究所專注的台灣手語，目前尚無台灣手語的資料庫可供訓練，只有台灣教育部所發行的線上

常用手語辭典以及國立中正大學編撰的台灣手語線上辭典兩者。雖然兩者對於每個手語辭彙都附有一部示範影片，但若訓練手語辨識的模型，這樣的資料量是遠遠不夠的，例如中國手語資料庫 DEVISIGN 中對於 2000 個不同的手語，每個分類皆有 12 部影片，且都是由不同的手語者比出。因此本研究將依據台灣教育部發行的線上常用手語辭典為基準，並參考 ASLLVD 和 DEVISIGN 兩個國外手語資料庫自行拍攝台灣手語的影片。

本研究旨在開發出一個台灣手語辨識系統能夠：

1. 幫助聾啞人士與他人無障礙地溝通。
2. 當聾啞人士遇到緊急狀況需要求救時他人能給予即時且適當的幫助。

基於以上兩點，本研究歸類出感受類、求救類、溝通類以及日常類四種符合任務需求的類別，並根據這四種類別在教育部線上常用手語辭典中精選 40 個台灣手語辭彙來當作本研究的資料集辨識目標。40 個台灣手語詞彙可參考表 1，手語的手勢與動作請參考附錄一。

分類	詞彙
感受	害怕(fear)、高興(glad)、厭惡(dislike)、痛苦(painful)
求救	失蹤(disappear)、尋找(search)、搶劫(rob)、頭痛(headache)、飢餓(hungry)、遺失(lost)、助聽器(hearing aid)、受傷(wounded)、感冒(catch a cold)、昏眩(dizzy)、求救(ask for help)、危險(danger)
溝通	我們(we)、不行(cannot)、不是(not right)、不要(don' t want)、不知道(don' t know)、沒關係(never mind)、小心(careful)、了解(understand)、立刻(at once)、可以(can)、同意(agree)、忘記(forget)、抱歉(sorry)、歡迎(welcome)、請求(request)、謝謝(thank)、非常(very)、鼓勵(encourage)
日常	吃飯(eat)、喝水(drink)、口罩(respirator)、租借(rent)、電話(telephone)、休息(relax)

表 2 本研究蒐集的 40 個台灣手語與其分類。

資料來源：本研究製作。

本研究拍攝的每部影片皆為 1920*1080 畫素，幀率為 30 幀，拍攝的角度皆為手語

者的正面，但稍有不同偏斜使模型學習時更為強健，垂直方向為頭部到腰部，與手語的規定範圍相符。總計由 6 位不同的手語者比出手語，包括 4 位男生與 2 位女生，年齡範圍為 19~45 歲，影片的背景則包含 7 種不同的背景。

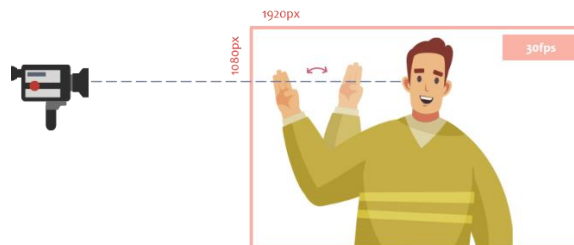


圖 15 台灣手語資料集製作拍攝示意圖。

資料來源：本研究製作。

3.3 資料前處理

3.3.1 資料分布情形

經整理後，本研究可用影片為 746 部影片，40 個台灣手語分類的分布如圖 16，平均每個分類有 19 部影片，資料分布非常均衡。每個手語詞彙的複雜度都不盡相同，有些詞彙只包含一個手勢，打出手語的時間非常短，例如：助聽器、了解；而有些辭彙則是由好幾個手勢動作組合而成，例如：非常、喝水，打出手語的時間比較長。因此每部影片的時間長短不同，根據圖 17，資料集中的影片幀數平均約為 78 幀，最少為 38 幀而最多為 152 幀。

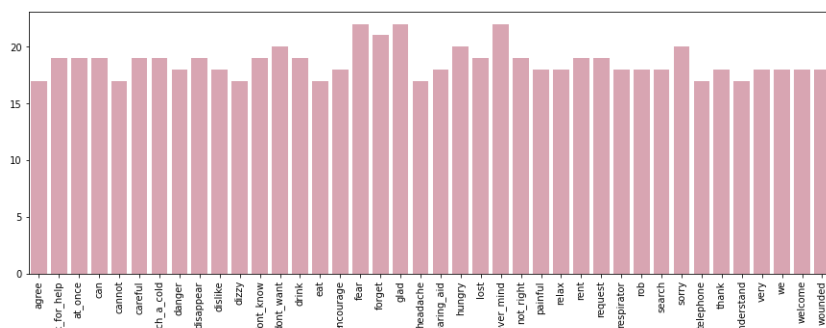


圖 16 40 個手語分類資料分布情形。

資料來源：本研究製作。

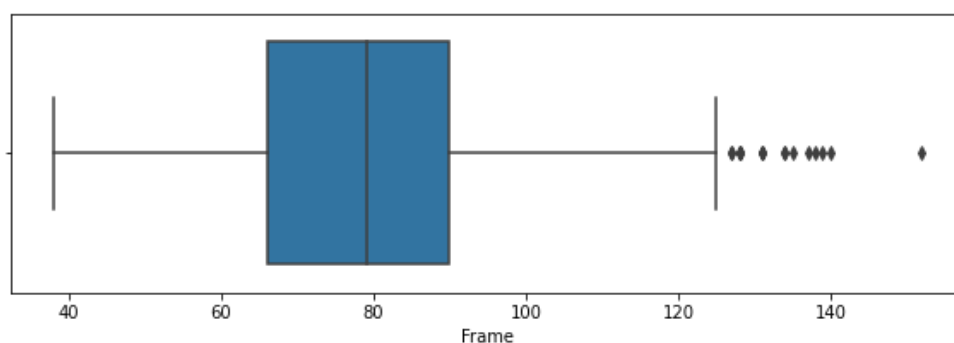


圖 17 影片幀數分布圖。

資料來源：本研究製作。

3.3.2 影片轉換為 RGB 圖片

因為本研究的 3D-CNN 模型的最後一層為全連結的分類層，因此將影片餵給模型訓練之前需先轉換為 RGB 圖片並確保每筆資料的圖片數量(幀數)皆相同。首先會先使用 HPE 模型從影片中抽取出人體的全身關鍵點向量，由全身關鍵點向量找出影片中手語者的最大行動範圍後，根據此範圍將圖片裁切為專注在手語者本身的正方形圖片並將圖片縮小為方便訓練的 256x256 大小。在訓練的過程中，為了使得每部影片的圖片數量都一樣，考慮 GPU 能夠運算的記憶體空間，取約平均的幀數當作基準(本研究取 70 幀)對每段影片做裁切或是補足。若影片的幀數大於基準幀數，則裁切影片中間的部分作為結果；若影片的幀數小於基準幀數，則重複影片直到與基準幀數相同。

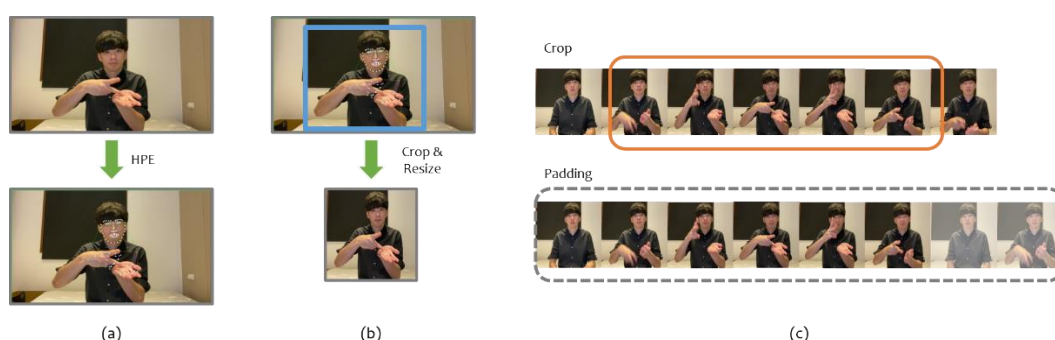


圖 18 影片轉換為 RGB 圖片流程圖，(a)從影片中抽取出全身關鍵點，(b)對影片做大小的裁切和縮小，(c)根據基準幀數對影片做影片長度的裁切或補正。

資料來源：本研究製作。

每段影片經過轉換後，共得到 59610 筆原始影像作為 RGB 模型的資料集，如圖 19，每段影片轉換後的圖像會放在與影片同名的資料夾中，每張圖像從 1 開始編號，代表此圖像為影片中的第幾幀，例如 0031.jpg 為影片中的第 31 幀。

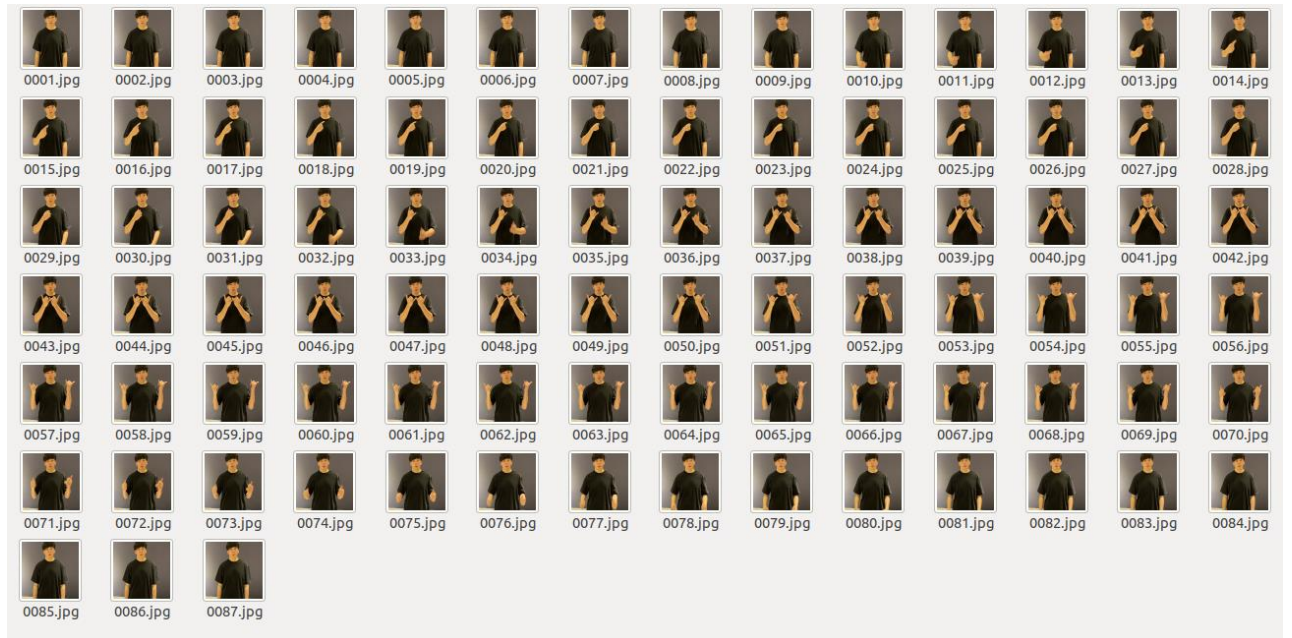


圖 19 手語辭彙「我們」完整動作轉換成連續圖片。

資料來源：本研究製作。

3.3.3 訓練集及驗證集

模型的好壞建立在當訓練完成的模型遇到沒看過的資料時是否能正確地辨識出結果，因此在實驗開始之前，為了正確地訓練和測試模型，會將資料集分成訓練集和驗證集，訓練集用來訓練模型，而驗證集用來驗證模型的效能。

打散時首先將資料隨機分散，再以 4:1 的比例將資料集平衡地分配成訓練集以及驗證集，並確保訓練集和資料集中擁有同比例的每一個手語分類的資料量，不會有訓練集或資料集沒有其中一個分類的資料的情況。最後訓練集有 619 筆影片，驗證集則有 127 筆影片。

3.4 全人體姿態估計

目前市面上成熟且開源的全人體姿態估計有 CMU 開發的 OpenPose 和使用 HRNet 作為基礎模型的 DarkPose。OpenPose 雖然有非常詳細的參數說明及完善的系統，但在效能以及準確度上則落後最新的 DarkPose 一截。經本研究測試，在畫面中看不到被觀察者的手肘時，OpenPose 會完全偵測不到被觀察者的手部姿態。另外，在同時偵測臉部、手部以及肢體姿態的情況中，OpenPose 的幀率會大幅度的下降。在上述兩種狀況下，DarkPose 的精度及效能皆沒有受到影響，因此本研究選擇 DarkPose 為全人體姿態估計的模型。

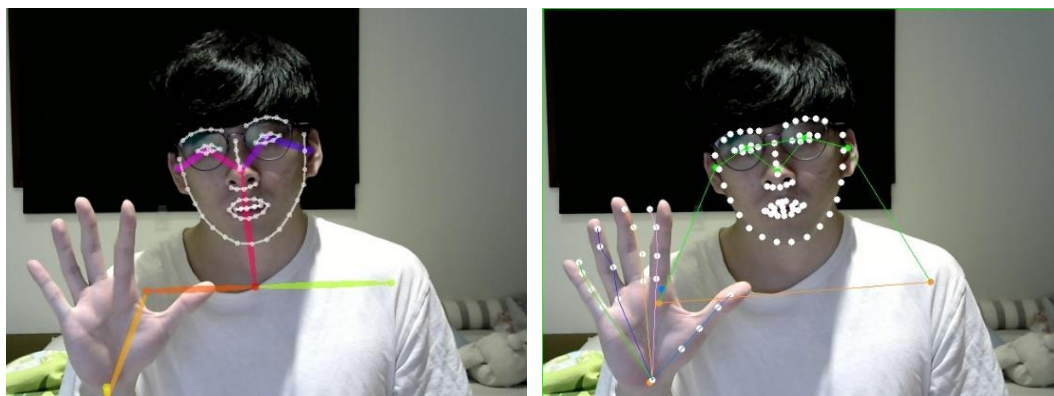


圖 20 OpenPose(左)在畫面中看不到手肘時偵測不到手部姿態，而 DarkPose(右)則不被影響。
資料來源：本研究製作。

3.4.1 全關鍵點

在全人體姿態關節點的抽取上參考 COCO 全人體資料集[35]抽取 133 個關鍵點，分別為 17 個肢體關節點、6 個足部關鍵點、68 個臉部關鍵點及左右手各 21 個關鍵點。對於每個關鍵點將產生一組三維資料，前兩個數字為關鍵點的二維座標(X , Y)，分別代表關鍵點在影像中的水平座標點以及垂直座標點；第三個數字為模型對於此關鍵點的信心值，信心值為 0 到 1 的浮點數。

因為本研究的手語資料集並不涵蓋腰部以下的範圍，因此腰部(12、13)、膝蓋

(14、15)、腳踝(16、17)和 6 個足部關鍵點將略過，不放入預測模型之中。在此部分的實驗中，會將所有人體關鍵點一起訓練，總共為 121 個關鍵點。

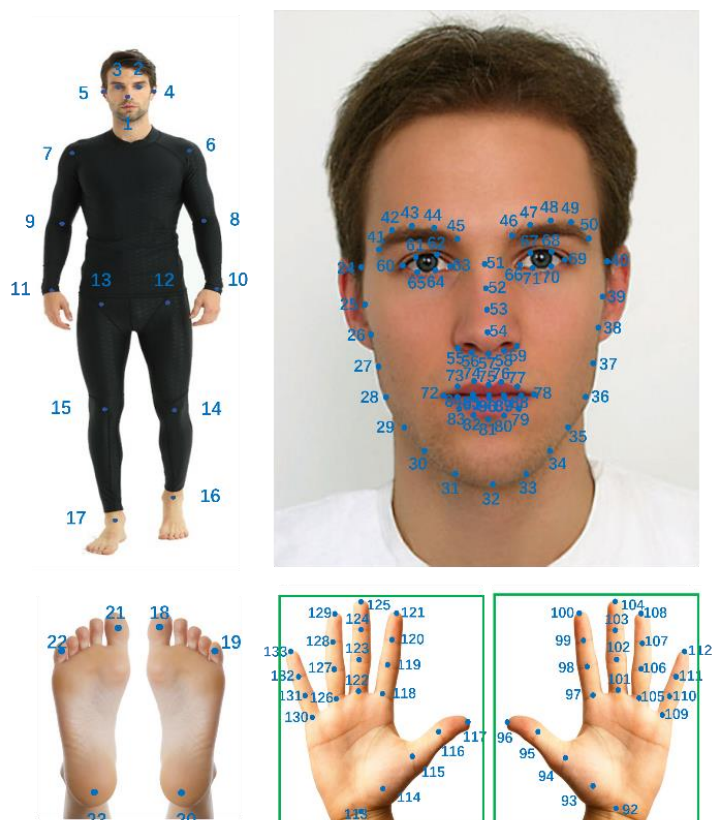


圖 21 從影像中抽取的人體關鍵點及其編號。

資料來源：COCO-WholeBody [35]

3.4.2 特徵關鍵點

121 個全人體關鍵點可以互相連線得到人體每個重要部位的大致輪廓，並由此擷取到人體的每一個細微動作，但在訓練模型的時候，太多的資訊可能會造成模型效能的下降或是過度擬合。因此在此部分的實驗中將嘗試從每一個關鍵部位抽取部分關鍵點，稱為特徵關鍵點。特徵關鍵點雖然無法呈現關鍵部位的細微變化，但依舊可以代表關鍵部位的移動和變化。特徵關鍵點包含肢體 7 點，雙手各 11 點以及臉部 11 點共 39 點。

分類	部位(關鍵點)
肢體(7)	鼻子(1)、耳朵(4, 5)、肩膀(6, 7)、手肘(8, 9)
臉部(10)	眉毛(41, 43, 45, 46, 48, 50)、嘴巴(84, 86, 88, 90)
左手(11)	手腕(92)、大拇指(94, 96)、食指(97, 100)、中指(101, 104)、無名指(105, 108)、小拇指(109, 112)
右手(11)	手腕(113)、大拇指(115, 117)、食指(118, 121)、中指(122, 125)、無名指(126, 129)、小拇指(130, 133)

表 3 人體特徵關鍵點。

資料來源：本研究製作。

3.5 預測模型

3.5.1 關鍵點模型

首先會使用人體關鍵點來進行手語的辨識，本實驗使用 Yan et al. [36] 提出的時空間 GCN(Graph Convolution Network) 模型作為此部分的基礎模型。為了要從人體的關鍵點中獲得空間中的人體姿態，要先將關鍵點根據自然人體並依據編號互相連結成人體骨架，可以視為一個在 2D 空間中由點和邊組成的圖。再來，為了要獲得時間維度上動作的變化，再將相鄰幀上對應的點互相連接，以此作為模型的輸入。

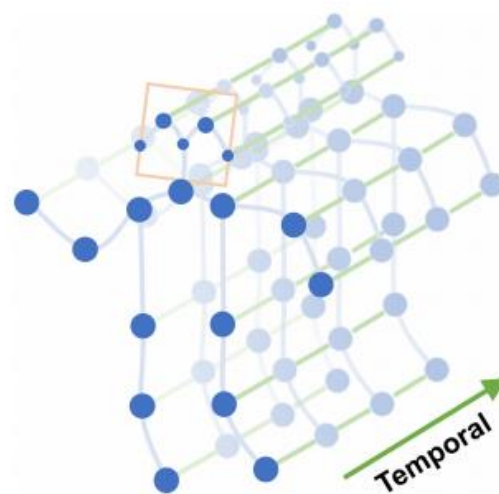


圖 22 模型輸入示意圖，藍點為人體關鍵點，藍線為骨架，綠線為時間維度對應點的連結。

資料來源：Yan et al. [36]

3.5.2 RGB 模型

在 RGB 影像的部分，使用 Tran et al. [9] 在 2018 年提出的 ResNet2+1D，為在 3D ResNet 上應用 1D 卷積的變體，並使用其在 Kinetics 資料集上預先訓練的權重來做為基礎模型。此模型將原本的 $T \times H \times W$ 的 3D 卷積核分為 $1 \times H \times W$ 的 2D 卷積核來處理空間的特徵以及 $T \times 1 \times 1$ 的 1D 卷積核來處理時間上的特徵，增加了非線性的層數也使得模型的訓練錯誤率更低。

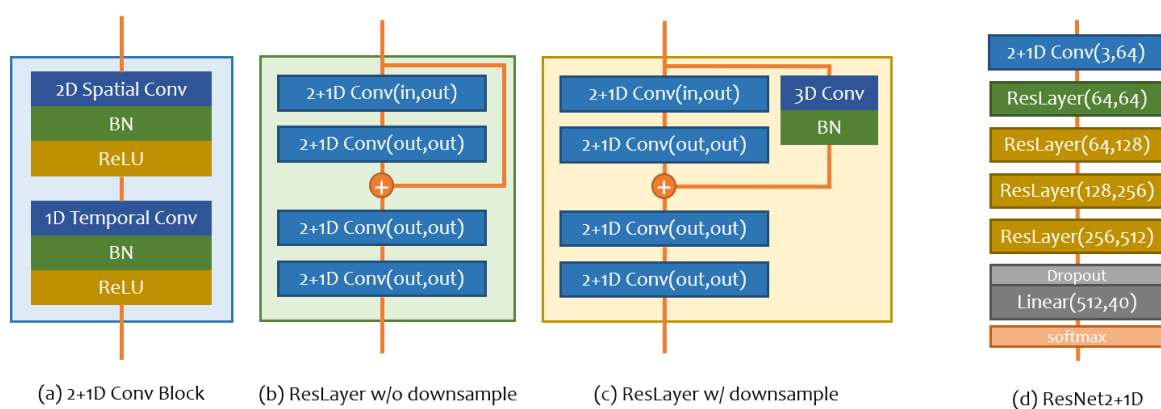


圖 23 ResNet2+1D 模型架構。

資料來源：本研究製作。

3.6 模型集成

每個模型在訓練時會從相同的資料集中學習到不同的特徵，給予每個特徵的權重也不盡相同，因此不同的模型時常會預測出不同的結果。模型集成(model ensemble)即利用此特性將不同模型的結果透過投票或加權平均，使得輸入資料的每個特徵都能部分地反映在最後的結果上，是在深度學習中為提升模型效能的常用技法。

根據本研究的模型架構，最終會獲得人體姿態模型的預測輸出以及 RGB 影像的預測輸出，兩個模型的輸出皆為一個長度為 40 的向量，分別代表輸入影片為 40 個手語的機率。本研究將透過給予兩個預測結果不同的權重並加總來得到更精準的最終預測結果。

$$\text{predict}_{final} = \alpha \times \text{predict}_{pose} + \beta \times \text{predict}_{RGB}$$

其中predict代表模型的輸出， α 、 β 分別為兩個模型的權重。本研究將根據兩個模型在驗證時的準確率調整權重的分配來得到最好的結果。

四、實驗結果與討論

4.1 實驗環境

CPU：AMD Ryzen7 3700X

GPU：GeForce RTX 2070

作業系統：Ubuntu 18.04

程式語言：Python 3.7

深度學習框架：Pytorch 1.8.1

4.2 關鍵點模型

4.2.1 全關鍵點

第一階段的實驗會使用 COCO-WholeBody 資料集標註的 133 個全人體關鍵點去除掉下半身包含腰部(12、13)、膝蓋(14、15)、腳踝(16、17)和 6 個足部關鍵點等 12 個關鍵點後剩餘的 121 個關鍵點來進行模型的訓練以及預測。本實驗將輸入這 121 個關鍵點的垂直、水平座標點，訓練 100 個 epoch 並獲得 Top-N 準確率²，結果如圖 24。

從結果可看出訓練時 Top-3 和 Top-5 的準確率在約 20 個 epoch 後達到約 95%的準確率，而 Top-1 的準確率則在 60 個 epoch 後才趨於穩定。從 loss 的下降趨勢圖中則可看出，因為 121 個關鍵點提供太多冗雜的訊息，loss 一直無法趨於穩定，時常有大幅度的升降。此階段實驗的 Top-1 準確率為 94.90%，Top-3 和 Top-5 的準確率皆為 99.32%。

² 預測信心值前 N 高的分類中命中正確分類的比例。

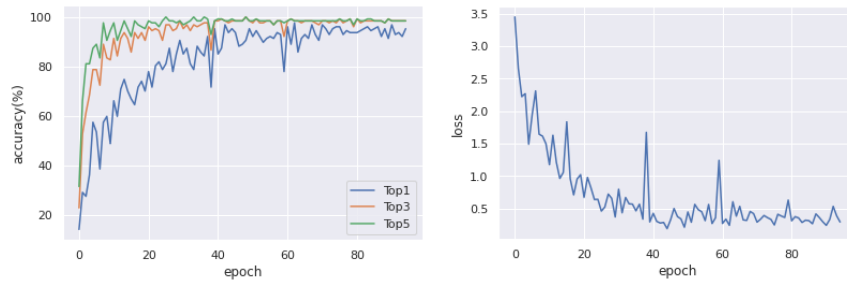


圖 24 121 個全關鍵點訓練時(左)驗證資料的 Top1、Top3、Top5 準確度趨勢和(右)loss 的下降趨勢。
資料來源：本研究製作。

圖 25 為此階段實驗結果的混淆矩陣，其中「立刻(at once)」、「不是(not right)」和「謝謝(thank)」為一組表現較差的分類，此三個分類皆為動作相似但手勢略為不同，推測為太多的特徵點雜訊分散了手勢部分的偵測準確率。另外，「不知道(don't know)」以及「厭惡(dislike)」也為一組表現較差的分類，從此分類中可看出雖然在動作和表情上都非常相似，但在手勢上卻非常不同。從這兩個表現較差的大類可看出有牽涉到表情的變化的手語和動作類似的手語被多餘的關鍵點影響較大，因此在此階段的實驗中表現不佳。

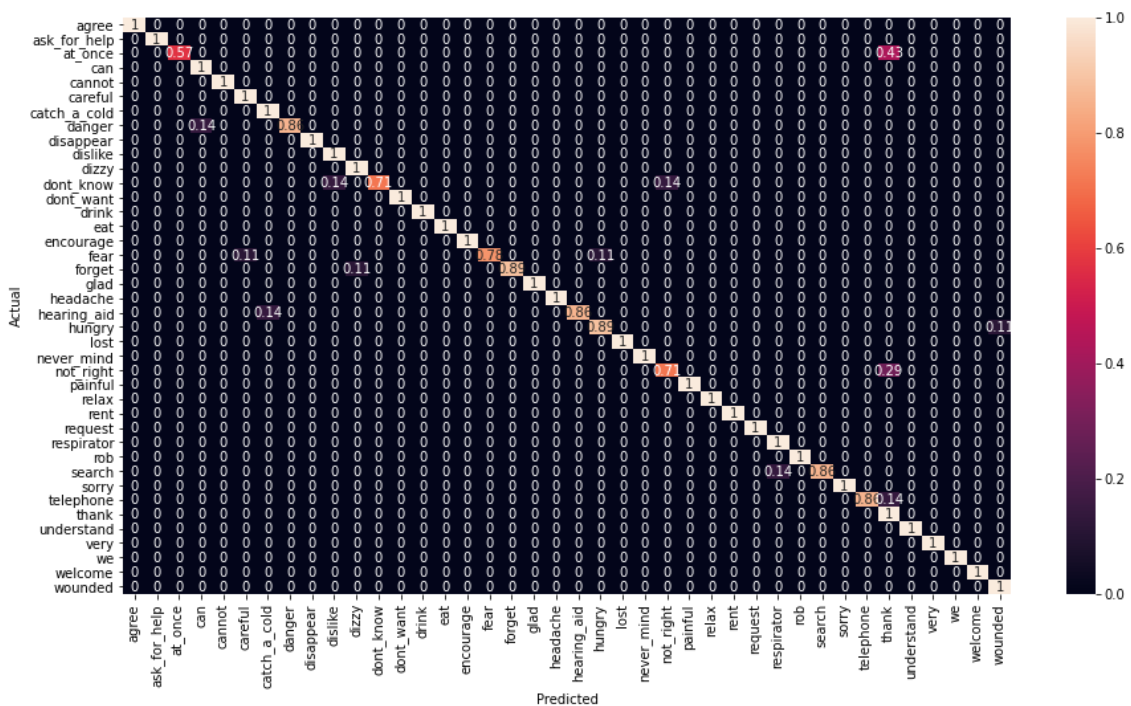


圖 25 全關鍵點訓練下的混淆矩陣。
資料來源：本研究製作。

4.2.2 特徵關鍵點

第二階段的實驗會從第一階段使用的 121 個人體關鍵點中選擇 39 個足以代表人體關鍵部位姿態的特徵關鍵點作為訓練和測試的資料，並觀察去除冗餘的資訊後模型是否表現較佳。本實驗將輸入這 39 個關鍵點的垂直、水平座標點並訓練 100 個 epoch，結果如圖 26。

從結果可看出訓練時 Top-3 和 Top-5 的準確率在約 20 個 epoch 後達到約 95% 的準確率就趨於穩定，而 Top-1 的準確率較第一階段的 60 個 epoch 更快，在 40 個 epoch 後就趨於穩定。從 loss 的下降趨勢圖中則可看出，使用 39 個特徵關鍵點時 loss 的幅度升降較第一階段有平滑的趨勢，訓練的過程較穩定。此階段實驗的 Top-1 準確率為 97.96%，Top-3 和 Top-5 的準確率皆為 100.00%。

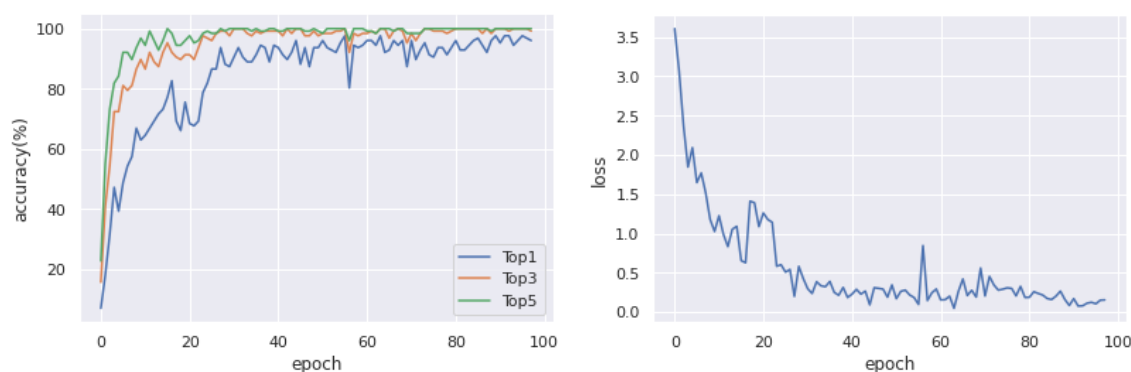


圖 26 特徵關鍵點訓練時(左)驗證資料的 Top1、Top3、Top5 準確度趨勢和(右) loss 的下降趨勢。

資料來源：本研究製作。

圖 27 為此階段實驗結果的混淆矩陣，其中在第一階段中分類較差的「不知道(don't know)」以及「厭惡(dislike)」在此階段表現更好，而「立刻(at once)」和「不是(not right)」、「謝謝(thank)」兩個分類也可以較正確地區分開來，因此模型更聚焦於手勢變化。

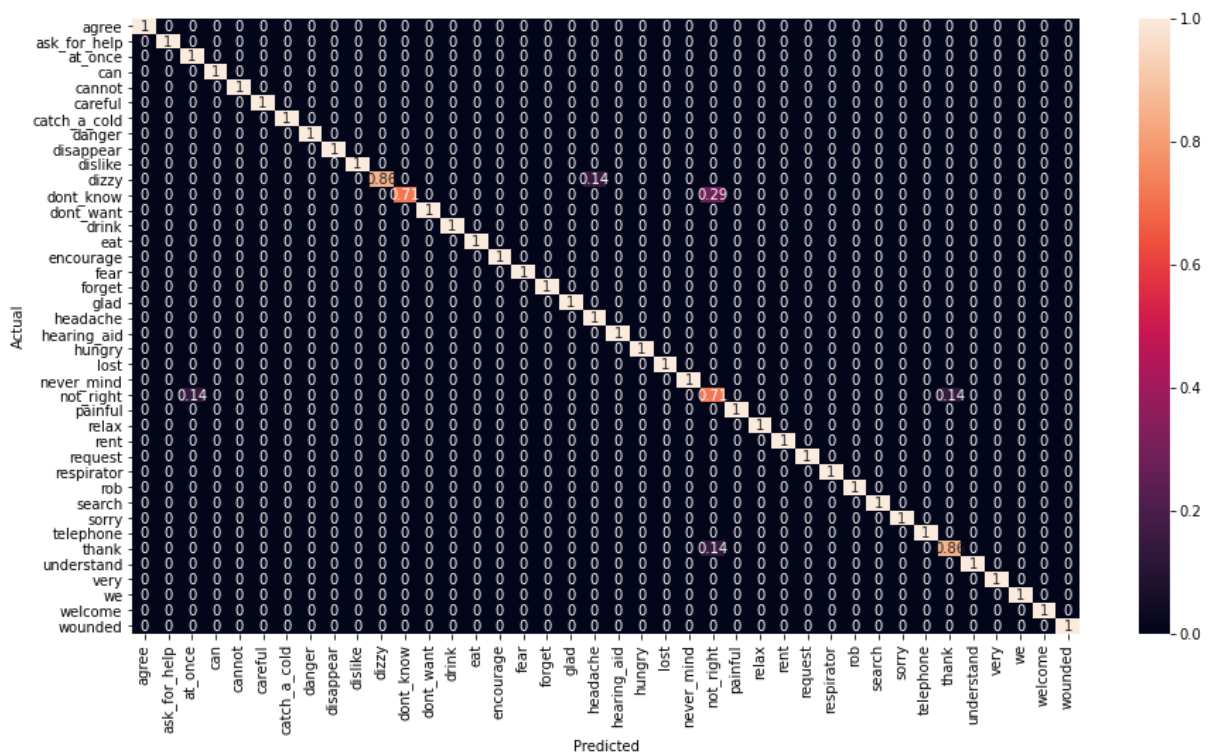


圖 27 特徵關鍵點訓練下的混淆矩陣。

資料來源：本研究製作。

4.3 RGB 模型

第三階段的實驗會將使用 3D-CNN 針對經 3.3.2 處理好的連續 RGB 圖形進行直接辨識。本階段使用的 3D-CNN 模型為 R(2+1)D 模型，為將 3D-ResNet 中的 3D 卷積核拆解為 2D 卷積核+1D 卷積核的變體，分別處理空間和時間的特徵資訊，以此來優化模型的訓練過程。本實驗將輸入 59610 筆原始影像資訊，模型的輸入維度為(channel, frame, size_x, size_y) = (3, 70, 64, 64)，學習率(learning rate)為 0.001，並訓練 100 個 epoch，結果如圖 28、29。

從結果可看出訓練和驗證準確度在約 70 個 epoch 後達到約 95%的準確率就趨於穩定，而驗證 loss 則持續平滑地下降，訓練的過程較穩定。此階段實驗的 Top-1 準確率為 97.62%，Top-3 和 Top-5 的準確率皆為 99.32%。

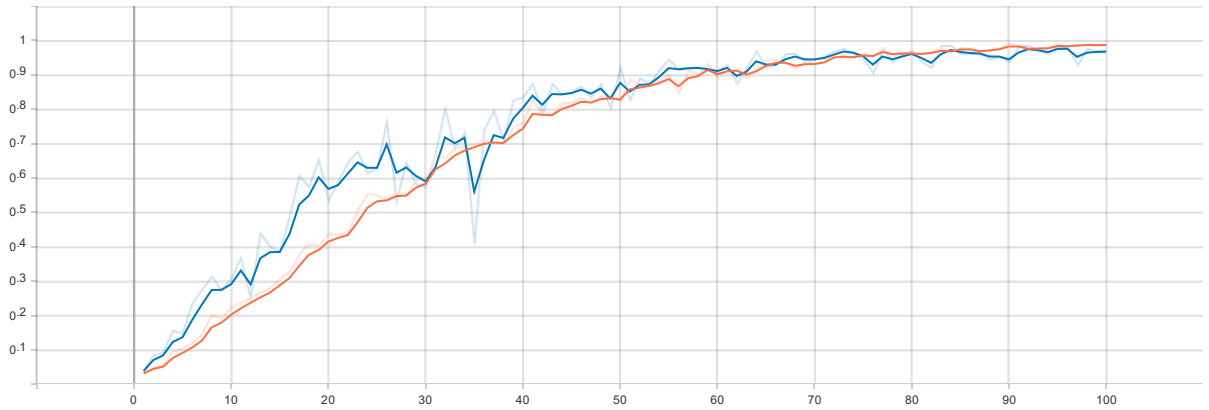


圖 28 3D-CNN 訓練時的準確度趨勢，藍線為訓練準確度，橘線為驗證準確度。

資料來源：本研究製作。

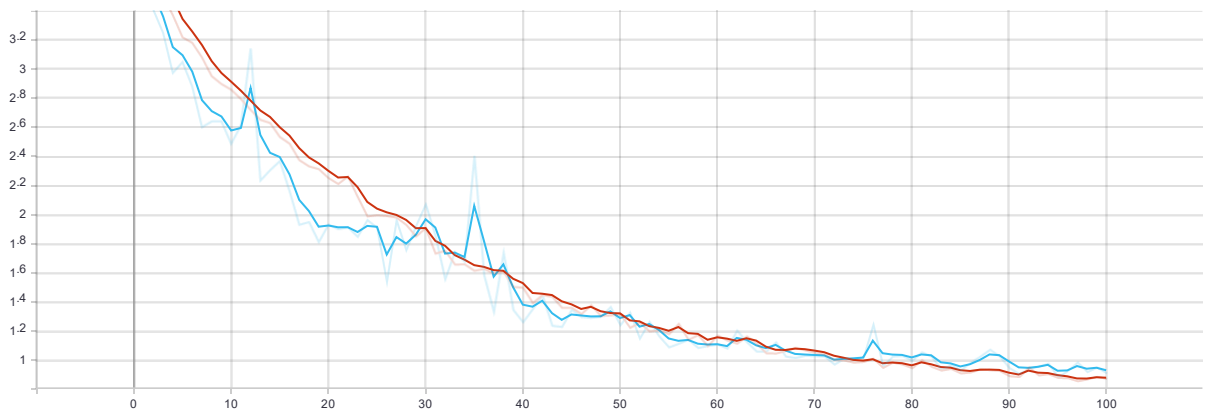


圖 29 3D-CNN 訓練時的 loss 下降趨勢，藍線為訓練 loss，紅線為驗證 loss。

資料來源：本研究製作。

圖 30 為此階段實驗結果的混淆矩陣，其中「謝謝(thank)」、「不知道(don' t know)」為兩個表現較差的分類。與第二階段相比可看出 RGB 模型在遇到相似的手語且有表情變化的手語上表現更不穩定，例如「不知道(don' t know)」被各有 14% 被預測為「不是(not right)」、「痛苦(painful)」和「了解(understand)」，這四個手語皆為單手在身前揮動的手語，在動作和位置都非常相似，只差在手勢有些微不同，且「痛苦(painful)」對比另外幾個明顯有痛苦的表情，如表 5。又如「飢餓(hungry)」和「休息(relax)」，兩者在動作上雖相似，但在表情上亦有明顯不同。然而，從「立刻(at once)」、「不是(not right)」和「謝謝(thank)」這組第一、二階段表現較差的分類可看出，RGB 模型在動作變化上更為敏感，也表現得更好。

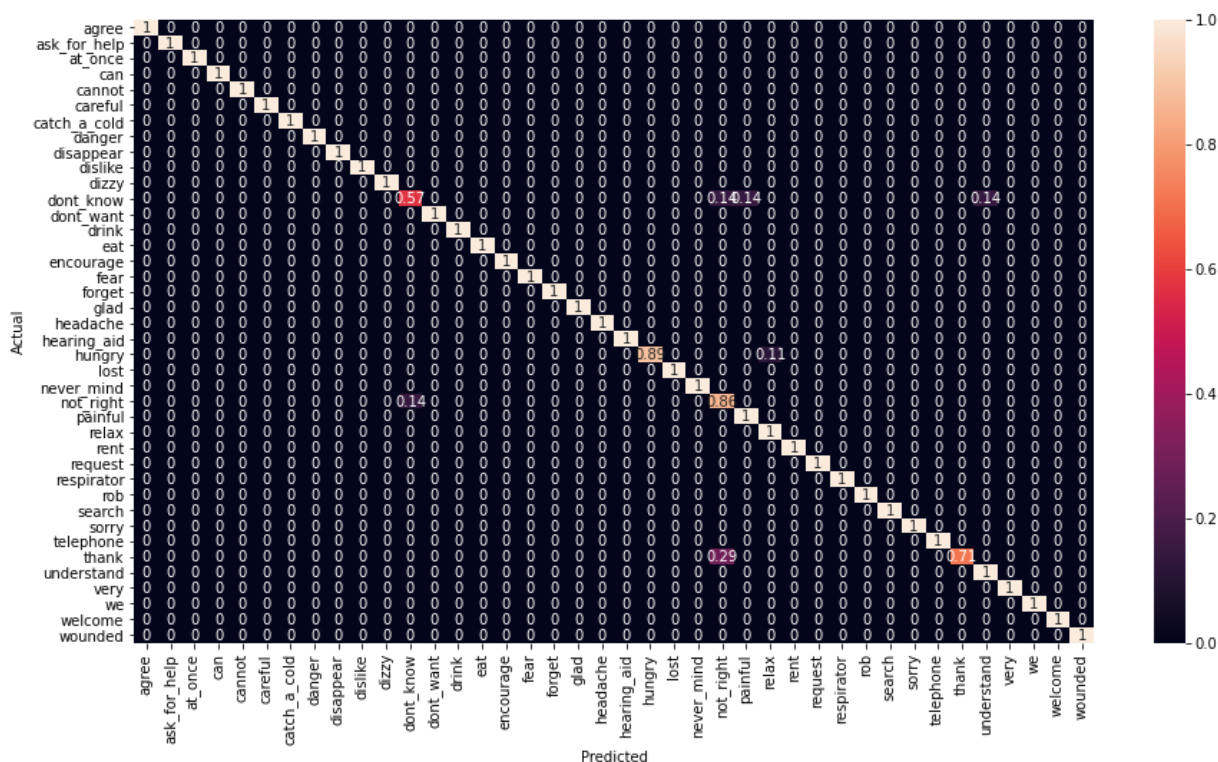


圖 30 RGB 模型訓練後的混淆矩陣。

資料來源：本研究製作。

4.4 模型集成

此階段的實驗會將訓練完的姿態模型以及 RGB 模型做模型集成來看看是否得到更好的預測準確度。根據前面的實驗的結果，本實驗測試當兩個模型的集成權重分別為 0.55 和 0.45 時效能最佳，算式如下：

$$\text{predict}_{final} = 0.55 \times \text{predict}_{pose} + 0.45 \times \text{predict}_{RGB}$$

模型集成後的模型準確率為 98.64%，Top-3 和 Top-5 的準確率皆為 100.00%，模型的比較如表 4。

Method	Top-1	Top-3	Top-5
Joint-121	94.90%	99.32%	99.32%
Joint-39	97.96%	100%	100%
RGB	97.62%	99.32%	99.32%
Ensemble (RGB+Joint-39)	98.64%	100%	100%

表 4 比較各階段模型與模型集成後的準確率。

資料來源：本研究製作。

圖 31 為此階段實驗結果的混淆矩陣，其中預測錯誤的剩下「不知道(don't know)」、「謝謝(thank)」和「不是(not right)」三個分類。與第二、三階段相比可看出大部分來自 RGB 模型在遇到動作類似但表情、手勢不同的手語上表現不穩定的情況都被改善，而關鍵點模型也藉由 RGB 模型對於動作較為敏感的特性亦有改善，而剩下的錯誤都是來自兩者模型皆有的錯誤。

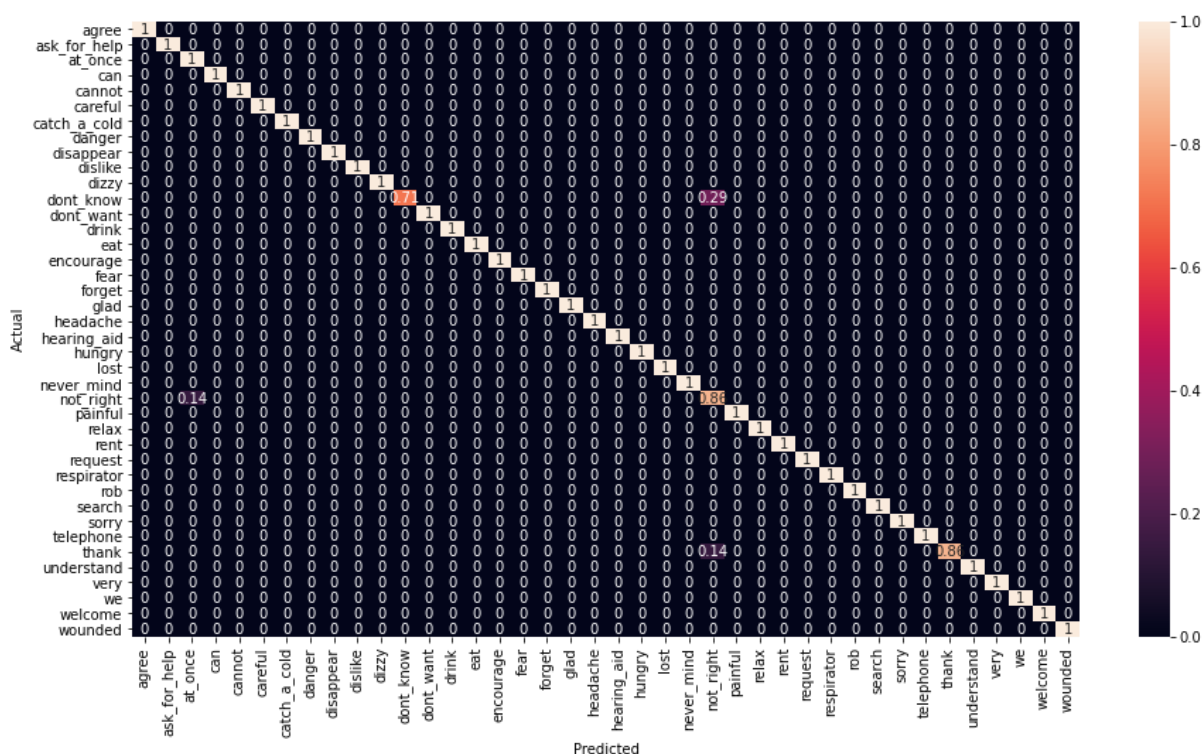


圖 31 模型集成後的混淆矩陣。

資料來源：本研究製作。

「不知道(don' t know)」這個分類在 RGB 模型和關鍵點模型中都被預測錯誤，在兩個模型中都有被誤判為「不是(not right)」，但在 RGB 模型中「不知道(don' t know)」被誤判為「痛苦(painful)」和「了解(understand)」這兩個手語的情況皆被關鍵點模型修正，猜測此原因為「痛苦(painful)」和「了解(understand)」與「不知道(don' t know)」間的相似度不比「不是(not right)」和「不知道(don' t know)」間的相似度。此四種手語皆為單手在身前筆劃的手語，在動作和位置都非常相似，只差在手勢有些微不同，如表 5。


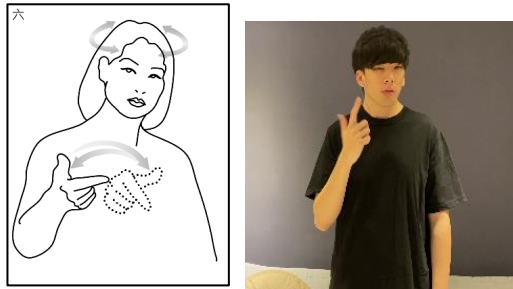
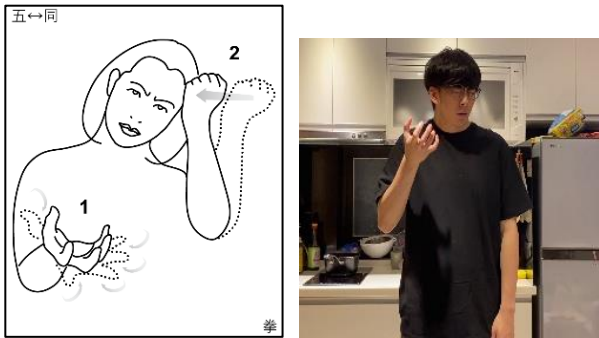
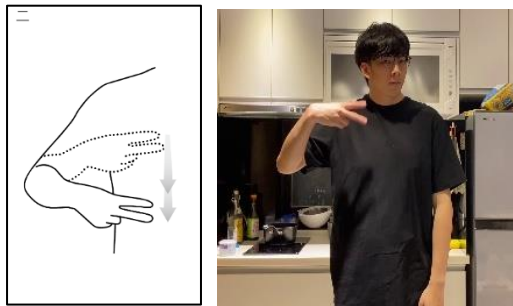
不知道(don' t know)	不是(not right)
	
痛苦(painful)	了解(understand)
	

表 5 不知道與其相似手型：不是、痛苦、了解。

資料來源：本研究資料集和教育部線上常用手語辭典 <https://signlanguage.moe.edu.tw/>。

立刻(at once)	不是(not right)	謝謝(thanks)
		

表 6 立刻、不是與謝謝比較。

資料來源：本研究製作。

五、結論與未來展望

5.1 研究結論

在過去的手語辨識模型中使用者需要配戴穿戴式的設備來獲取手語者的肢體變化，或是使用深度攝影機捕捉場景的 3D 深度資訊來進行背景分離進而分析手語者的動作。本研究提出一台灣手語辨識模型，可直接透過 RGB 影像資料辨識出手語者比出的 40 種台灣手語。此模型由兩個子模型所組成，分別為使用人體關鍵點進行預測的 GCN 模型以及針對 RGB 影像進行辨識的 3D-CNN 模型，最後再集成兩個模型得到辨識的結果。

在實驗的第一階段使用人體上半身的 121 個關鍵點進行模型訓練，Top-1 準確率為 96.85%。第二階段去除多餘資訊造成模型效能下降的狀況，從 121 個上半身關鍵點取出 39 個關鍵點，獲得更高的 Top-1 準確率 97.96%，較第一階段提高了 1%，且 Top-3 的準確率為 100%。第三階段加入整個畫面的資訊，使用 RGB 的影像來直接進行手語辨識，獲得的 Top-1 準確率為 97.62%。第四階段將第二三階段生成的預測結果進行加權相加，讓模型同時參考人體的動作變化以及 RGB 變化，達到模型集成的效果，最終辨識的 Top-1 準確率為 98.64%，Top-3 準確率為 100%，且透過混淆矩陣可以看出加入關鍵點模型的預測修正了單使用 RGB 模型遇到相似手語時辨識錯誤的問題。

5.2 研究限制

在缺乏台灣手語資料庫的情況下，本研究只能透過自行拍攝影片來創建資料集，深度學習模型的效能仰賴著資料量的多寡，越多元，越大量的資料可以使得模型在面對不同的狀況下依舊維持著良好的性能，但在人力和時間因素下只能從眾多台灣手語詞彙中選擇 40 種常用詞彙並獲得平均每種 20 部影片的資料集。雖然實驗最終獲得將

近 99%的辨識準確率，但在遇到不同的手語者，不同的場景或是非正面的角度時都可能造成性能的下降。為了增加模型的強健性，本研究盡量在不同場景下搜集不同性別，不同年齡的手語影片來組建資料集。

5.3 未來展望

如前段所述，由於目前缺乏台灣手語的影像資料庫，因此本研究只針對 40 個手語辭彙進行辨識，離全部手語辭彙的數量還非常遙遠。若在未來台灣手語推廣組織願意協助創建訓練集以搭建台灣手語的大型資料庫，建議沿用本研究拍攝方法，拍攝手語者腰部以上、正面之手語影片，對於每個分類找數名手語者並拍攝至少 15 部影片以上，且盡量在不同背景、光照下進行拍攝，使模型學習不同的資訊。完成後可以直接將本研究的模型對於新的資料進行微調(fine-tuned)得到更全面、更強健的模型。

手語和口語一樣，除了有地區性外，手語也有文法。要能夠將手語辨識的模型應用到日常生活中，除了對於個別詞彙的「辨識」之外，還要能夠將一連串的手語組合「翻譯」成為一個句子，如此才能達到真正無障礙的境界。目前本研究的模型只能達到辨識的功能而不具備翻譯的功能，除了模型要能辨識更多的詞彙之外，如何將連續的手語拆分開來也是一大難題。

透過 HPE 的幫助，本研究已經完成了一個不使用穿戴式設備且不使用深度攝影機，只使用影像即可對台灣手語進行辨識的模型。若未來要普及給社會中的普羅大眾使用，可以將模型放在 Google 推出的 App 後端機器學習平台中，並利用手機的相機鏡頭拍攝手語者的畫面，再打開 App 即可進行手語辨識，讓聾啞人士可以更無障礙地與社會溝通。

參考文獻

- [1] 史文漢，丁立芬，手能生橋第一冊，台北，1995。
- [2] 姚俊英，台灣手語演進，手語教學與應用研討會論文集，2001。
- [3] A. Kuznetsova, L. Leal-Taixé and B. Rosenhahn, "Real-Time Sign Language Recognition Using a Consumer Depth Camera," *2013 IEEE International Conference on Computer Vision Workshops*, pp. 83-90, 2013.
- [4] A. Toshev and C. Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1653-1660, 2014.
- [5] B. Shi, A. M. D. Rio, J. Keane, D. Brentari, G. Shakhnarovich, and K. Livescu, "Fingerspelling recognition in the wild with iterative visual attention," *IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [6] B. Xiao, H. Wu and Y. Wei, "Simple Baselines for Human Pose Estimation and Tracking", *Proceedings of the European Conference on Computer Vision*, pp. 466-481, 2018.
- [7] C. R. Wren, A. Azarbayejani, T. Darrell and A. P. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780-785, July 1997.
- [8] D. Guo, S. Wang, Q. Tian and M. Wang, "Dense temporal convolution network for sign language translation", *Proc. 28th Int. Joint Conf. Artif. Intell.*, pp. 744-750, Aug. 2019.
- [9] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun and M. Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6450-6459, 2018.
- [10] F. Zhang, X. Zhu, H. Dai, M. Ye and C. Zhu, "Distribution-Aware Coordinate Representation for Human Pose Estimation," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7091-7100, 2020.
- [11] G. A. Rao, K. Syamala, P. V. V. Kishore and A. S. C. S. Sastry, "Deep convolutional neural networks for sign language recognition," *2018 Conference on Signal Processing And Communication Engineering Systems (SPACES)*, Vijayawada, pp. 194-197, 2018.
- [12] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, et al., "Towards accurate multi-person pose estimation in the wild", *Proceedings of the IEEE conference on computer vision and pattern Recognition (CVPR)*, pp. 3711-3719, 2017.
- [13] H. Brashear, T. Starner, P. Lukowicz and H. Junker, "Using multiple sensors for mobile sign language recognition," *Seventh IEEE International Symposium on Wearable Computers*, pp. 45-52, 2003.
- [14] H. Cheng, L. Yang and Z. Liu, "Survey on 3D Hand Gesture Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 9, pp. 1659-1673, Sept. 2016.

- [15]J. C. Núñez, R. Cabido, J. J. Pantrigo, A. S. Montemayor and J. F. Vélez, "Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition", *Pattern Recognition*, vol. 76, pp. 80-94, Apr. 2018.
- [16]J. Huang, W. Zhou, H. Li and W. Li, "Sign language recognition using 3d convolutional neural networks", *ICME*, pp. 1-6, 2015.
- [17]J. Kim, S. Mastnik and E. Andr, "EMG-based hand gesture recognition for realtime biosignal interfacing", *Proc. 13th Int. Conf. Intell. User Interfaces*, pp. 30-39, 2008.
- [18]J. Tompson, R. Goroshin, A. Jain, Y. LeCun and C. Bregler, "Efficient object localization using Convolutional Networks," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 648-656, 2015.
- [19]J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, et al., "Deep high-resolution representation learning for visual recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Apr. 2020.
- [20]K. Imagawa, Shan Lu and S. Igi, "Color-based hands tracking system for sign language recognition," *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, pp. 462-467, 1998.
- [21]L. Pigou, S. Dieleman, P.-J. Kindermans and B. Schrauwen, "Sign language recognition using convolutional neural networks", *Proc. Eur. Conf. Comput. Vis. Pattern Recog. Workshops*, pp. 1-6, 2014.
- [22]M. J. Cheok, Z. Omar and M. H. Jaward, "A review of hand gesture and sign language recognition techniques", *International Journal of Machine Learning and Cybernetics (IJMLC)*, vol. 10, no. 1, pp. 131-153, Jan. 2017.
- [23]M. Kocabas, S. Karagoz and E. Akbas, "MultiPoseNet: Fast multi-person pose estimation using pose residual network", *Proc. ECCV*, pp. 417-433, Sep. 2018.
- [24]M. Oberweger, P. Wohlhart and V. Lepetit, "Hands deep in deep learning for hand pose estimation", 2015.
- [25]N. Sripairojthikoon and J. Harnsomburana, "Thai Sign Language Recognition Using 3D Convolutional Neural Networks", *Proceedings of the 2019 7th International Conference on Computer and Communications Management*, pp. 186-189, Jul. 2019.
- [26]P. Felzenszwalb and D. Huttenlocher, "Pictorial Structures for Object Recognition", *International Journal of Computer Vision (IJCV)*, vol. 61, no. 1, pp. 55-79, 2005.
- [27]P. Lokhande, R. Prajapati and S. Pansare, "Data Gloves for Sign Language Recognition System", *IJCA Proceedings on National Conference on Emerging Trends in Advanced Communication Technologies NCETACT*, pp. 11-14, Jun 2015.
- [28]P. S. Rajam and G. Balakrishnan, "Real time Indian Sign Language Recognition System to aid deaf-dumb people," *2011 IEEE 13th International Conference on Communication Technology*, Jinan, pp. 737-742, 2011.

- [29]Q. De Smedt, H. Wannous and J. Vandeborre, "Skeleton-Based Dynamic Hand Gesture Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Las Vegas, NV, pp. 1206-1214, 2016.
- [30]Q. Pu, S. Gupta, S. Gollakota and S. Patel, "Whole-home gesture recognition using wireless signals", *Proceedings of the 19th Annual International Conference on Mobile Computing and Networking*, pp. 27-38, 2013.
- [31]R. Anderson, F. Wiryana, M. C. Ariesta and G. P. Kusuma, "Sign language recognition application systems for deaf-mute people: A review based on input-process-output", *Procedia Computer Science*, vol. 116, pp. 441-448, Oct. 2017.
- [32]R. Cui, H. Liu and C. Zhang, "Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, pp. 1610-1618, 2017.
- [33]R. H. Liang and M. Ouhyoung, "A real-time continuous gesture recognition system for sign language", *Proceedings of IEEE International Conference on Automatic Face Gesture Recognition*, pp. 558-567, 1998.
- [34]S. Ji, W. Xu, M. Yang and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231, Jan. 2013.
- [35]S. Jin, L. Xu, J. Xu, C. Wang, W. Liu, C. Qian, W. Ouyang, and P. Luo, "Whole-body human pose estimation in the wild," *Proceedings of European Conference on Computer Vision*, 2020.
- [36]S. Yan, Y. Xiong and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition", *Proc. AAAI*, pp. 1-9, 2018.
- [37]T. Starner and A. Pentland, "Real-Time American Sign Language Recognition From Video Using Hidden Markov Models", *Proc. Int'l Symp. Computer Vision*, 1995.
- [38]X. Chen, G. Wang, H. Guo and C. Zhang, "Pose Guided Structured Region Ensemble Network for Cascaded Hand Pose Estimation," 2017.
- [39]X. Zhang, X. Chen, Y. Li, V. Lantz, K. Wang and J. Yang, "A framework for hand gesture recognition based on accelerometer and EMG sensors", *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* vol. 41, no. 6, pp. 1064-1076, Nov. 2011.
- [40]Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu and J. Sun, "Cascaded pyramid network for multi-person pose estimation", *Proceedings of the IEEE conference on computer vision and pattern Recognition (CVPR)*, pp. 7103-7112, 2018.
- [41]Z. Cao, G. Hidalgo, T. Simon, S. -E. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172-186, 1 Jan. 2021.
- [42]Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *in CVPR*, 2017.

附錄一

害怕	高興	厭惡	痛苦
<p>廿手</p> 	<p>五五</p> 	<p>萬</p>  <p>萬→同</p>	<p>五↔同</p>  <p>拳</p>
失蹤	尋找	搶劫	頭痛
<p>手手</p> 	<p>錢錢</p> 	<p>萬→拳</p>  <p>同同→拳拳</p>	<p>一</p>  <p>五↔同</p>
飢餓	遺失	助聽器	受傷
<p>手手</p> 	<p>〇→手</p>  <p>同同→拳拳</p>	<p>句</p> 	<p>手手</p>  <p>一手</p>
感冒	昏眩	求救	危險
<p>胡</p>  <p>二</p>	<p>萬</p> 	<p>手手</p>  <p>方手</p>	<p>同</p> 

我們	不行	不是	不要
<p>一</p> 	<p>一</p> 	<p>六</p> 	<p>呂→六</p> 
不知道	沒關係	小心	了解
<p>手</p> 	<p>錢錢→五五</p> 	<p>一二</p>  <p>六六</p>	<p>二</p> 
立刻	可以	同意	忘記
<p>二手</p>  <p>六</p>	<p>女</p> 	<p>一</p>  <p>拳</p>	<p>萬→同</p> 
抱歉	歡迎	請求	謝謝
<p>九</p> 	<p>五五</p>  <p>≠</p>	<p>手</p>  <p>手手</p>	<p>男→副</p> 

非常	鼓勵	吃飯	喝水
	<p>手男</p>	<p>棕同</p>	<p>方</p>
口罩	租借	電話	休息
<p>句句</p>	<p>欠手</p>	<p>民</p>	<p>手手</p>

圖片來自於教育部線上常用手語辭典 <https://signlanguage.moe.edu.tw/>