國立臺灣大學電機資訊學院電機工程學系

碩士論文

Department of Electrical Engineering
College of Electrical Engineering and Computer Science
National Taiwan University

Master Thesis

# 雙流膨脹卷積網路之台灣手語單字辨識

# Two-Stream Inflated 3DCNN for Taiwanese Sign Language Recognition

楊鎮維

Chen-Wei Yang

指導教授：顏嗣鈞 博士

Advisor: Hsu-chun Yen, Ph.D.

中華民國 109 年 07 月

July 2020

# 中文摘要

在台灣手語辨識的相關研究中，往往都是透過輔助裝置獲得影片中的手部、運動軌跡等資訊並利用傳統特徵擷取算法來實驗與闡述，使得這些研究往往僅能解決在特定情境、特定資料集下的辨識問題，且許多論文並未明確指示出實驗資料集的數量與分布。本論文首先建立了一個針對台灣手語單字影片的資料集，共有 4 位不同人在不同背景比出常用之 63 種手語單字，每種單字種類有 8 至 9 個影片，共有 538 支影片；接著透過近年在影像動作辨識上取得良好效果的雙流膨脹卷積網路以泛化的形式解決手語辨識問題，在測試集上達到 87.3% 的準確率，並透過實驗及錯誤分析給出針對光流和 rgb frame 作為輸入的往後改進方向。最後將訓練完的模型應用到能即時展示辨識結果的程式中。

# 英文摘要

In the related research on Taiwanese sign language recognition, the hand position and motion trajectory information in the sign language video are often obtained through auxiliary devices, and traditional feature extraction algorithms like SIFT、SURF are used for experiment. These make the relates research can only solve the problem for a specific data set or scenario. Furthermore, many papers do not clearly indicate the experimental data set like quantity and distribution. In this paper, we first established a data set for Taiwanese sign language vocabulary videos. A total of 4 different people gesture the 63 commonly used sign language vocabularies in different backgrounds. Each vocabulary has 8 to 9 videos and a total of 538 videos. Secondly, we apply two-stream dilated convolutional network [1], which has achieved good results in video motion recognition recent years, to solve the Taiwanese sign language recognition problem in a generalized form. In the experiment, we reach an accuracy of 87.3% on the test set, and through error analysis, we indicate future improvement on rgb frame and optical flow as input for model. Finally, the trained model is applied to a program that can display the recognition results in real time.

# 目錄

# 圖目錄

# 表目錄

# Chapter 1

# 緒論

## 1.1 背景

手語是聾人間傳遞訊息的媒介，要與聾人進行交談必須雙方皆熟習手語表達或有專業的手語翻譯員進行雙向的翻譯。而手語辨識系統的相關研究則可以免除這兩者條件，一方面保障聾人們資訊獲得的權利，一方面幫助大眾對聾人族群的了解交流。手語就像如同正常的語言般，各國家有自己形式的手語語言，甚至在國家地區內還存在有不同的手語方言，因此手語辨識系統必須依照各個國家與地區去做資料的蒐集與系統建置。在台灣的手語系統中，兩個聾人在一起打的手語為「自然手語」，其語法與中文大不相同，但「中文手語」的語法則完全依循中文文法，只是將中文用一種視覺的方式加以表達 [7]，自然手語如同日常口語般主旨在傳達意思，中間字詞可能會有所省略；而中文手語則是將一個句子中每個字對應的手語表達依序比劃出，手式雖多但有益於聾人學習中文。本論文主要研究的是台灣手語的中文手語單字辨識，輸入一段台灣手語的單字影片，辨識出所表達的單詞。

1.

## 1.2 台灣手語相關研究

關於目前台灣手語辨識方面的研究主要有: Taiwan sign language (TSL) recognition basedon 3D data and neural networks [8] 利用微型反光標記器辨識靜態手勢 (輸入為靜態手勢圖片)；Kinect-based Taiwanese sign-language recognition system [9] 利用 kinect 得到手部位置，將手部位置與手部在前後 frame 的距離、手掌形狀利用 SVM、HMM 等方法分類到預定義好的類別中當作特徵，最後將這些特徵透過與預定義好的 word database 機率表相乘做出單字分類。但此篇沒有明確說出訓練測試資料量，以及利用到許多人為定義的特徵類別；A real-time continuous gesture recognition system for sign language [10] 利用 dataglove 所得的 motion、手部姿勢、方向、位置等資訊結合 HMM 做單字及句子的辨識。手部姿勢使用 613 個樣本訓練、281 個做測試，手部方向使用 143 個樣本訓練、71 個做測試，手部 motion 使用 279 個樣本訓練、40 個做測試，最後單字分別測試在 71、155、250 個單字上準確率為 84%、82%、70%；Taiwan Sign Language Recognition System Using LC-KSVD Sparse Coding Method [11] 利用 kinect 得到深度圖以及關節點位置，利用 HOG 算出手部特徵和關節點位移向量後使用 K-SVD 做字典分類；Sign Language Recognition[12] 利用白色手套抓出手部的軌跡，配合傅立葉描述符選取 key frames 並抓出特徵，和現有的單字特徵算歐式距離做分類。綜上所述可發現，目前針對台灣手語單字辨識的研究都需要額外的輔助條件如 kinect、dataglove、特定的人物與環境顏色等，並透過這些來獲取手部位置和軌跡資訊，利用圖像特徵擷取算法擷取特徵輸入給 SVM、HMM 等傳統機器學習分類器方法去做單字分類。而這些研究當中的問題有

- 沒有數量明確、不需額外輔助條件的台灣手語單字影片資料集來當作比較基準與實驗
- 沒有使用深度神經網路的方法來做一般化的特徵擷取和分類

故本論文的主要貢獻點有 1. 製作台灣手語單字影片的資料集，往後相關研究可以此資料集為骨幹去拓展、實驗和比較精準度 2. 使用深度神經網

1.

路的方法以泛化解決手語單字的辨識問題，且實驗資料集不需要額外的輔助環境、裝置，並列出 baseline 模型的實驗數據和改善方向。

# Chapter 2

# 神經網路模型與光流簡介

隨著近年來資料量與硬體運算力的提升，深度神經網路在許多領域上的表現如圖像識別已經超越各式方法，包括傳統機器學習分類器如 SVM、HMM，rule-based 以及融合 hand craft 特徵的演算法。從以往必須透過許多對問題的洞見以及特定條件限制去提高辨識準確率，發展到近年來透過資料驅動訓練深度神經網路去實現各種任務，使得各領域的問題能以一般且泛化的形式去解決。而在其中關注的重點也轉移到將高階抽象化的假設設計成不同的神經網路模型，接著透過大量的資料去訓練模型使之自行學習到對於解決任務最重要的特徵以及參數。以下分別介紹在手語辨識系統中會運用到的基本模型、網路以及取得影片 motion 的光流演算法。

## 2.1 光流 (Optical Flow)

光流是一種用來描述影片中相鄰 frame 與 frame 之間變化的方法，將相鄰 frame 間物體的移動以二維向量 (x,y) 來表示，分別代表物體在 x,y 方向上的位移。而求解光流的問題，相當於輸入兩個前後關係的 frame，輸出和 frame 大小相同的二維向量場。光流應用除了有 motion based 的物體切割、相機校準、影像壓縮外，在本論文的手語辨識系統中也提供了影像的 motion 資訊。

2.

### 2.1.1 假設前提與光流基本方程式

光流的求解通常會基於兩種假設:

- 假設一: 前後 frame 的明亮度保持不變 (亮度固定)
- 假設二: 物體的移動範圍較小

令 (x, y) 點在第 t 張 frame 的亮度為 f(x,y,t)，且經過了 dt 時間後位移量為 (dx,dy)，則根據假設一可得到

$$f(x, y, t) = f(x + dx, y + dy, t + dt) \quad (式1)$$

透過假設二將 (式 1) 右邊進行一階泰勒展開:

$$f(x, y, t) = f(x, y, t) + \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial t}dt \quad (式2)$$

將 (式 2) 移項並同時除以 dt，可得光流的基本方程式:

$$f_x u + f_y v + f_t = 0 \quad (式3)$$

$$\begin{aligned} f_x &= \frac{\partial f}{\partial x};\ f_y = \frac{\partial f}{\partial y} \\ f_t &= \frac{\partial f}{\partial t} \\ u &= \frac{dx}{dt};\ v = \frac{dy}{dt} \end{aligned}$$

### 2.1.2 Horn&Shunck 方法

透過將光流的基本方程式加上不同的約束條件進行最佳化求解可得到不同約束條件下的結果，Horn&Shunck 方法 [13] 透過光流的平滑約束 (速度的變化率越小越好) 結合 (式 3) 可得到:

$$\iint \{(f_x u + f_y v + f_t)^2 + \lambda((u_x)^2 + (u_y)^2 + (v_x)^2 + (v_y)^2)\}\, dx\, dy$$

希望上式趨近於 0，其中 fx、fy、ft 可透過將不同的 edge detection mask 與前後 frame 作運算並相加最後結果得到，則目前剩餘未知數為 u, v。將

2.

上式透過變分法求等於 0 的值:

$$(f_x u + f_y v + f_t)^2 fx + \lambda(\nabla u)^2 = 0$$
$$(f_x u + f_y v + f_t)^2 fx + \lambda(\nabla v)^2 = 0$$

其中 Lapacian u, v 透過 Lapacian mask 可計算出，最後移項可得

$$u = u_{av} - f_x \frac{P}{D}$$
$$v = v_{av} - f_y \frac{P}{D}$$
$$P = f_x u_{av} + f_y u_{av} + f_t$$
$$D = \lambda + f_x^2 + f_y^2$$

將 $u_{av}$, $v_{av}$ 的值初始化為 0，並迭代運算多次得出最佳的光流解。然而此方法有兩項缺點

- 由於光流的平滑約束，無法精確處理圖像裡的邊界或梯度較大的區域
- 由於假設二，當物體的移動範圍較大時光流會有誤差

針對梯度較大的區域，A Duality Based Approach for Realtime TV-L1 Optical Flow [14] 透過改變平滑約束項為一次項與求解過程改良；而針對光流較大的移動範圍，Coarse to Over-Fine Optical Flow Estimation [15] 利用圖像分層概念，將圖像做多層 down sampling 算出光流後，透過差值將結果迭代回上一層已取得大範圍的光流變化。

## 2.2 卷積神經網路 (Convolutional Neural Network, CNN)

卷積神經網路 [16] 是解決圖形辨識最重要也最基礎的神經網路，相比於一般全連接的深度神經網路，卷積神經網路一方面將人腦辨識視覺的概念轉

2.

移至模型中，一方面透過圖像的性質如特徵的平移不變性、局部性、圖片的多尺度設計卷積與池化操作，大大增加有效的參數量。例如任務為辨識圖片是否存在貓，而貓的獨特特徵如貓耳、貓的鬍鬚可能存在圖片裡的某一個局部，而平移或放大縮小都不會影響此特徵，因此可以透過以下兩個操作來擷取不同的特徵，相比直接使用全連接的深度神經網路減少大量無效的參數和特徵。

### 2.2.1 卷積

卷積運算透過不同大小的 filter 和圖片進行卷積，從而擷取到局部、平移不變性的特徵。卷積操作如 圖 2.1，將圖片和 filter 進行 element-wise 相乘的結果相加作為輸出，透過設定移動步伐由左而右、由上而下掃過整張圖片。不同以往計算圖像特徵的方法如 SIFT、SURF 需人為決定在不同情況分別適用哪種算法，神經網路透過卷積和資料驅動訓練的過程讓 filter 調整適當的參數去自行決定對解決任務最重要的特徵。通常同一大小的 filter 會產生多個和圖片進行卷積，分別對應到不同的特徵幫助辨識，並稱為一卷積層當作輸入給神經網路的下一層。透過反覆堆疊卷積層能組合前一層擷取的線、稜角等低階特徵成考量更大範圍的高階特徵 (形狀、圖案)。而卷積後的結果會再透過激活函數去掉負值、加強物體輪廓並加速神經網路收斂。

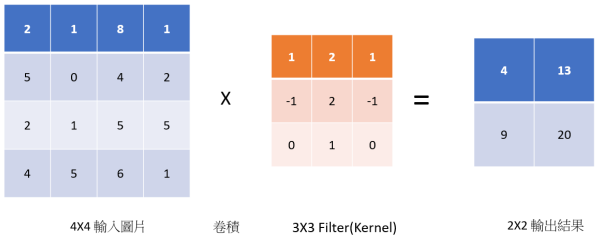



圖 2.1: 卷積運算

2.

### 2.2.2 最大值池化

最大值池化對應到圖片的多尺度，將一張圖片保留重要的特徵並縮放不會影響圖像的識別，如貓經過縮放仍然保持貓的性質，但能減少神經網路所需的參數量，而且池化後的資訊更專注於圖片中是否存在相符的特徵，而非圖片中哪裡存在這些特徵。最大值池化運算如 圖 2.2將輸入矩陣和 filter 做取最大值操作，只保留 filter 大小內的最大值當作輸出。例如將一大小為 4x4 矩陣經過移動步伐為 2、大小為 2x2 的最大值池化運算後，輸出大小為 2x2，分別代表各自區塊內的最大值。最大值池化會接續在卷積層的後面，將卷積擷取的特徵壓縮，稱為池化層。卷積神經網路反覆堆疊卷積

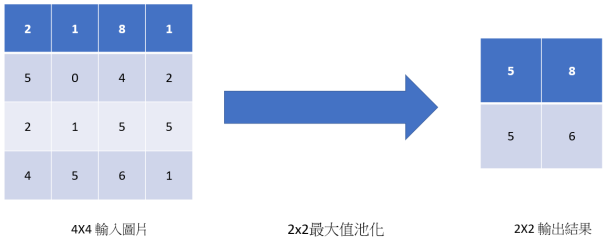



圖 2.2: 池化運算

層與池化層多次後，最後將結果輸入到全連接層進行分類得出最終結果，整體卷積神經網路模型架構如圖 圖 2.3

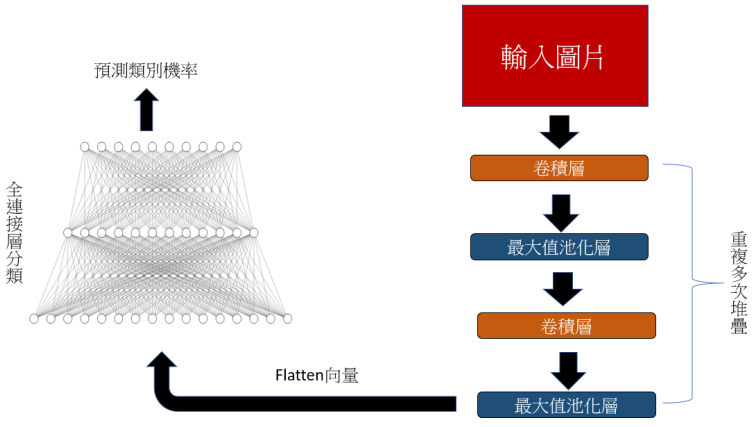



圖 2.3: 卷積神經網路

2.

## 2.3 Inception 網路

### 2.3.1 1x1 卷積

1x1 卷積概念最早是由一篇論文 Network In Network 中所提出來 [17]，當把 filter 的大小設為 1x1 時，卷積後的圖片寬高不變，channel 則變為 1x1 filter 的個數。當輸入圖片的 channel 為 1 時，可以單純的把 1x1 卷積看成將圖片裡面的值放大縮小並做 1x1 filter 的個數次疊起來；而當輸入圖片的 channel 大於 1 時，1x1 卷積等同於對輸入的 pixel channel 做全連接放到對應的輸出位置上，同樣的也做 1x1 filter 的個數次疊起來。概念上為在卷積層中間加入全連接層增加網路深度以提高網路的表現力，示例圖如 2.1。1x1 卷積的概念後來被廣泛使用，主要有兩點功能:

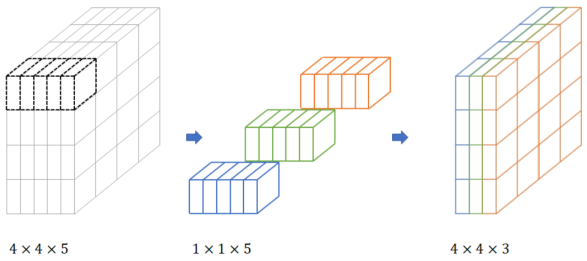



圖 2.4: 1x1 卷積 [2]

- 增加神經網路的深度與非線性程度
- 透過 filter 的個數增加或減少圖片的 channel

Inception 網路中即是透過 1x1 卷積降低圖片的 channel 來減少在深層網路中做卷積的計算複雜度。

### 2.3.2 Inception 模組

相較於普遍深度神經網路的不斷加深，Inception 網路的核心概念為透過 Inception 模組在神經網路廣度上的加寬。以往在堆疊卷積神經網路時需人為決定每一層是否為卷積層或池化層，以及對應使用的 filter 大小，

2.

但往往圖片中含有重要訊息的區域大小不一致；而 Inception 模組透過將不同大小的卷積 filter 與池化 filter(3x3、5x5、1x1) 在廣度上作列舉展開，由神經網路透過參數更新提取不同的特徵，最後再把各個特徵結果在 channel 上堆疊起來傳給下一層，增加網路的適應性，圖 2.2 為上述所提之 Inception 模組。

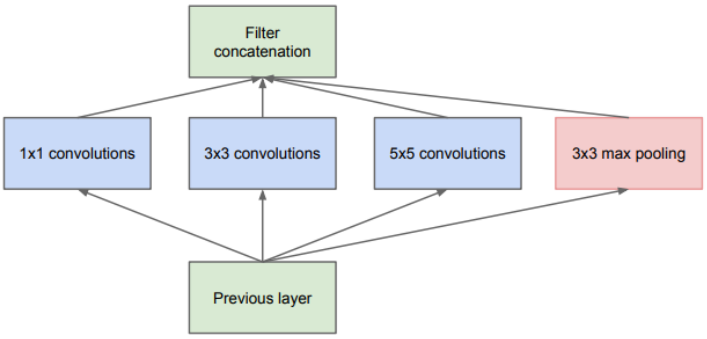



圖 2.5: Inception 模組 [3]

為了將特徵結果在 channel 上堆疊，Inception 模組裡的卷積操作皆是採用 same padding 模式，即圖片大小經過卷積後不會改變。

### 2.3.3 降維 Inception 模組

Inception 網路利用 Inception 模組在神經網路的廣度與深度上做堆疊，伴隨而來的是計算的複雜度的上升，而這邊將 1x1 卷積降維的概念放到 Inception 模組中可以大大的降低計算量 (圖 2.3)。雖然多了一層卷積操作，但由於降低了需要和 filter 卷積的圖片 channel，總體的計算量可以降到原本的十分之一，如下假設:

原始假設

- 輸入層圖片大小: 寬 28、高 28、channel 192
- 輸出層圖片大小: 寬 28、高 28、channel 32
- 5x5 filter 大小: 寬 5、高 5、個數 32

2.

則整體計算量為

$28x28x32$(輸出層$pixel$個數) $x$ $5x5x192$(卷積出$pixel$需做的乘法) $\approx$ 1.2億個乘法

降維假設

- 輸入層圖片大小: 寬 28、高 28、channel 192
- 中間層圖片大小: 寬 28、高 28、channel 16
- 輸出層圖片大小: 寬 28、高 28、channel 32
- 1x1 filter 大小: 寬 1、高 1、個數 16
- 5x5 filter 大小: 寬 5、高 5、個數 32

則整體計算量為

$28x28x16$(中間層$pixel$個數) $x$ $192 + 28x28x32$(輸出層$pixel$個數) $x$ $5x5x16 \approx$ 1200萬個乘法

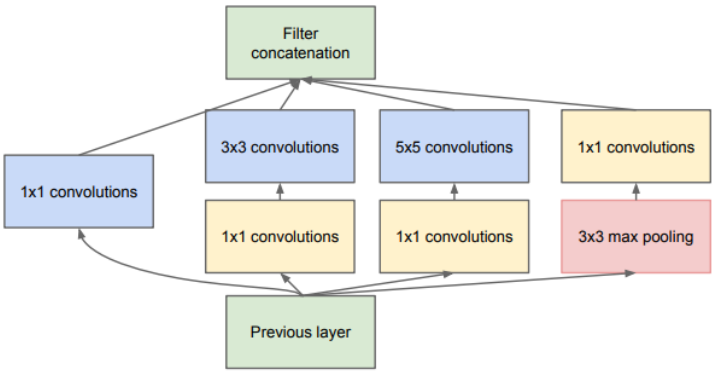



圖 2.6: 降維 Inception 模組 [3]

2.

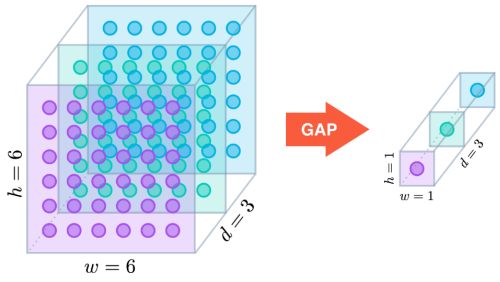



圖 2.7: 全域平均池化 [4]

### 2.3.4 訓練細節

神經網路在最後分類層使用了 Network In Network 中的全域平均池化 (GAP) 圖 2.7的概念來取代全連接層 [17]，用 GAP 的特點有二：

- GAP 在特徵圖和最終的分類間轉換更加簡單，等同於賦予最後特徵圖每個 channel 實際的類別意義
- 免除全連接層所需使用的參數，使神經網路更加健壯，降低神經網路的過擬合。

由於網路的龐大，Inception 網路另外在網路中間引入了兩個輔助分類器，在訓練時將分類層的 loss 和輔助分類器的 loss 乘以 0.3 的權重加起來再進行反向傳播，減少梯度消失的問題，即

總$loss$ = 分類層$loss$ + 0.3 × 輔助分類器1的$loss$ + 0.3 × 輔助分類器2的$loss$

從另一角度也可以看成將中間層的分類結果與分類層的最終結果做 ensemble，而輔助分類器僅使用在訓練時，測試時並不使用。總括來看 Inception 網路整體架構，一開始以基礎的卷積神經網路為主幹，在其上線性疊加 9 個降維 Inception 模組，中間延伸出兩個輔助分類器解決梯度消失的問題，最後在分類層使用全域平均池化層和 softmax 函數得到最終結果，總體網路深度含有 22 層，並在 2014 年 ImageNet 舉辦的 ILSVR 比賽中拿下最高的準確率。

# Chapter 3

# 影像動作辨識架構

針對影像動作辨識架構 (video action classification)，因 video 本質上是一連串的 frame 在時間頻域上的堆疊，故要考量的點有二: (1). 空間上 (2). 時間上。空間上蘊含了場景、人物、物件等資訊；而時間則蘊含物體的 movement 資訊。底下分別列出能擷取出影片時空域特徵並做分類的深度學習模型 [1]。

## 3.1 卷積神經網路 (Convolutional Neural Network, CNN)+ 長短期記憶模型 (Long Short-Term Memory, LSTM)

由於 CNN 在圖像識別上的成功，而 LSTM 是典型處理 time series 的深度學習模型；一個直覺的想法是結合兩者:

Step 1: 每一個 frame 通過 CNN encode 出 latent feature

Step 2: 將 Step 1 產生的 latent features 透過 LSTM 連接

Step 3: 以最後一個 LSTM 的 output 當作預測結果

訓練時針對每一個 LSTM 的 output 以 cross-entropy 當作損失函數算出 loss 更新；測試時僅以最後一個 LSTM 的 output 當預測結果。

3.

缺點: 因為 CNN 中間層的作用分別為提取 low level 到 high level 的圖片特徵，而 LSTM 是拿取 CNN 最後一層的輸出當作 input，因此此模型會缺少 low-level motion 的資訊，而這一點卻是在動作辨識上是很重要的。
模型架構如圖 3.1。

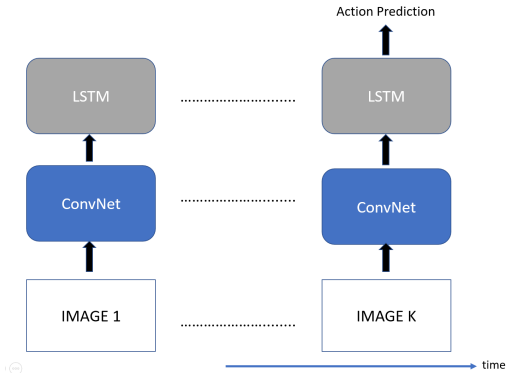



圖 3.1: 卷積神經網路 (Convolutional Neural Network, CNN) + 長短期記憶模型 (Long Short-Term Memory, LSTM)

## 3.2 3D 卷積神經網路 (3D Convolutional Neural Network, 3DCNN)

由 2D Convolutional Neural Network(2DCNN) 衍生而來，在 CNN filter 方面增加一維度 (時間維) 變成三維，每個 frame 依時間順序堆疊起來當成 input，由神經網路同時抓取時間與空間上的特徵。
缺點:

- 模型輸入變為三維，無法直接利用 ImageNet[16] 等大型資料集做預訓練
- 整體模型變為三維，相較於 3.1 的模型參數量變大，較難訓練

模型架構如圖 3.2。

3.

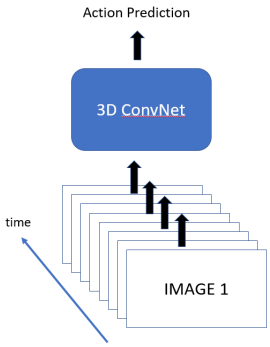



圖 3.2: 3D 卷積神經網路 (3D Convolutional Neural Network, 3DCNN)

## 3.3 雙流卷積神經網路 (Two-Stream Convolutional Networks)

為了解決 3.1 所提到的缺少 low-level motion 資訊，可以將 CNN 分成兩分支，一分支透過 RGB frame 擷取空間上的資訊，一分支透過計算得到的光流擷取時間上的資訊。因為計算出的光流蘊含了 motion 及時間上的資訊，可以減少神經網路在學習上的負擔，在效率及效能上都有不錯的表現。訓練時取一個 RGB frame、若干個所選 RGB frame 後的光流 (經過實驗發現選取 10 個光流 frame 表現最好)，各自經過 CNN 後將 softmax 的分數相加平均得到最終結果；測試時在影片中取樣若干個 frame 並同樣把結果平均得到預測 [5]。其他變形如訓練時取若干個 RGB frame 與對應的光流過雙流卷積神經網路，融合再輸入到 3DCNN 去學習更高階的時空流和空間流的關係 [18] [19]o。

模型架構如圖 3.3。

令光流堆疊結果在 k 時間點的 (u, v) 位置點為 Stack(u, v, k)，光流在 k 時間點的 (u, v) 位置為 Optical(u, v, k)，P1=(u, v)，(x)、(y) 為取向量的 x、y 分量運算子，選定一個 frame 與之後 L 時間點依序的光流圖。則對於光流作為輸入的疊加法有以下:

3.

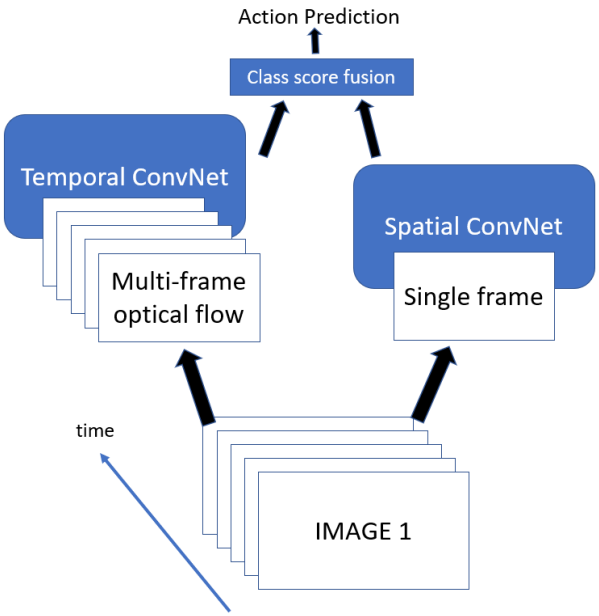



圖 3.3: 雙流卷積神經網路 (Two-Stream Convolutional Networks)

### 3.3.1 光流堆疊

針對一個點 (u, v)，將同一點 (u, v) 上的光流 x, y 方向依時序堆疊。即依序簡單堆疊 frame 之後 L 張時間點的光流圖，則光流堆疊總共會有 2L 張，每一張為對應時間點上光流圖的 x 或 y 方向。

$$Stack(u, v, 2k-1) = Optical(u, v, k)(x)$$

$$Stack(u, v, 2k) = Optical(u, v, k)(y)$$

### 3.3.2 軌跡堆疊

針對一個點 (u, v)，追蹤 (u, v) 點在時序上的光流軌跡。即在 L 時間點的堆疊點 Stack(u, v, L) 為 (u, v) 點從開始 frame 隨著光流軌跡更新至 L-1 時間點後，在 L 時間點最終位置的光流 x, y 方向。

$$Stack(u, v, 2k-1) = Optical(Pk(x), Pk(y), k)(x)$$

$$Stack(u, v, 2k) = Optical(Pk(x), Pk(y), k)(y)$$

$$Pk = Pk-1 + Optical(Pk-1(x), Pk-1(y), k-1)$$

3.

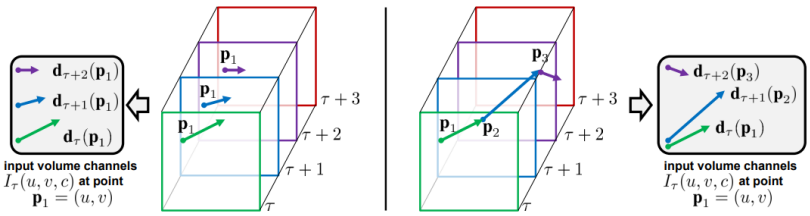



圖 3.4: 光流堆疊 vs 軌跡堆疊 [5]

### 3.3.3 雙向光流堆疊

除了取選定 frame 之後 L 時間點光流圖，也可以在時間序列上做反向的光流運算，分別選取 frame 之後 L/2 與之前 L/2 時間點的光流圖，如此可以考慮到雙向時間序的關係同時維持一樣的光流圖數目。在把光流圖當作輸入進雙流卷積神經網路時，通常把輸入資料中心標準化到以 0 為中心時訓練收斂效果最好，故在此可以把每張光流圖的 x,y 方向位移大小減去光流圖的平均位移，避免因相機移動等因素造成整體光流偏向某一方向。

## 3.4 雙流 3D 膨脹卷積神經網路 (Two-Stream Inflated 3DCNN)

此方法融合了 3.3 提及的雙流策略以及利用了 ImageNet[16] 的 2D 預訓練模型參數和架構 (如 Inception-V1)，解決上述模型提及的各個缺點。

### 3.4.1 將 2D 預訓練模型拓展為 Inflated 3DCNN

假設 2D 預訓練模型之各個 filter 為 N 維，將 filter 加一維度拓展成為 3D filter:

$$N \times N(square) -> N \times N \times N(cubic)$$

而整體架構維持不變，並將此模型稱之為 Inflated 3DCNN，如此可免去重複嘗試不同的 3D 架構的不同表現。

3.

### 3.4.2 由 2D 預訓練模型的參數初始化 Inflated 3DCNN

定義將單個 frame 重複堆疊 N 次為一個 boringVideo；2D 預訓練模型的 filter 為 f2D；Inflated 3DCNN 的 filter 為 f3D，則要使用 2D 預訓練模型的參數在 Inflated 3DCNN 必須滿足:

$$f2D \otimes frame = f3D \otimes boringVideo$$

即 2D 預訓練模型的 filter 與單一 frame 做卷積產生出來的 feature map 必須與 Inflated 3DCNN filter 和 boring video 做卷積得出的結果一致。這裡可以透過將 2D 預訓練模型的 filter 沿著時間軸方向拓展為 3D 後 (N x N -> N x N x N)，再除以 N 做數值縮放達到目標 [proof....]。由於卷積出來的數值相同，在模型後續的 activation layer 與 pooling layer 也能保證相同的數值與特性，完成把 2D 預訓練模型的參數初始化到 Inflated 3DCNN 上。

### 3.4.3 調整特徵的學習範圍域

在此 Inflated 3DCNN 模型中可以自行選擇的參數有

- convolution stride
- pooling stride
- pooling kernel size

調整這些會影響到神經網路在影片上學習特徵的範圍域。在處理 2D 圖片時，kernel 與 stride 通常會有對稱的性質如 (2,2)、(3,3)、(7,7) 等，涵義為同等對待圖片的水平與垂直平面；但把時間也納入考慮因素時，需要根據影片的 frame rate 與 frame 大小來調整 kernel size，有時對稱並不是最佳的。例如當時間變化快過於空間變化時，對稱的 kernel 會核並且破壞物體的邊緣資訊；而當空間變化快過於時間變化時，便無法有效抓取物體的 motion 資訊。

3.

### 3.4.4 加入光流作為輸入

縱使 Inflated 3DCNN 可以直接從時間域中學習 motion 特徵，但實驗上將計算完的最佳化光流也當作輸入，分別訓練模型再取平均，有助於提升不少的效能。

訓練時將光流與堆疊的 frame 分別輸入到 Inflated 3DCNN 分開訓練，測試時把兩個模型的輸出平均得到最終結果。

模型架構如圖 3.4

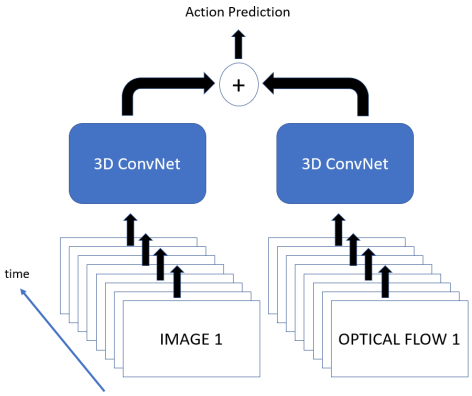



圖 3.5: 雙流 3D 膨脹卷積神經網路 (Two-Stream Inflated 3DCNN)

## 3.5 光流的效用探討

在前面小節中有提到把光流輸入到神經網路中直覺上可以更好的描述影片的 motion 資訊，但在近期的 On the Integration of Optical Flow and Action Recognition 論文 [6] 中提出了相異的看法: 透過光流描述影像中物體外觀的不變性質才是提升影像辨識準確率的主要原因。 表 3.1:

- 將光流作為輸入，打亂光流順序的精準度僅只有下降 8%；打亂計算光流之 RGB 圖像順序精準度雖有下降但仍有不差的效果，說明直覺中光流所含的 motion 資訊並不是決定準確率的關鍵因素
- 將 RGB 和光流分別作為輸入，但測試時將影像顏色做隨機修改，

3.

RGB 作為輸入的準確率降低非常多，而光流準確率變化不大，說明光流對顏色這種性質具有良好的不變性質

其他二個論點為: 1. 只有在位移較小及物體的邊緣處，光流的準確率才和影像辨識的準確率正相關。2. 透過影像辨識的 loss 來 fine tune 光流網路可以讓準確率提升。也是透過相關的實驗來佐證，這些想法可以帶出一些針對影像辨識模型的未來改進方向，如可以針對 RGB 輸入、光流輸入、模型架構去加強學習物體的不變性。

表 3.1: 不同輸入條件的影像辨識準確率 [6]

| 輸入 | 準確率 |
|---|---|
| 光流 | 86.85% |
| RGB | 85.45% |
| 光流 (打亂光流順序) | 78.64% |
| 光流 (打亂 RGB 圖像順序) | 59.55% |
| 光流 (改變顏色) | 84.30% |
| RGB(改變顏色) | 34.23% |

# Chapter 4

# 台灣手語單字辨識系統

## 4.1 手語資料集

### 4.1.1 資料來源

透過抓取台灣手語線上辭典 [20]、台灣手語地名電子資料庫 [21]、台灣手語拾遺 [22] 等線上台灣手語單字影片資源做為主幹，將從不同來源抓取之手語單字聯集結果視為常用的單詞，並做為單字辨識系統的單字分類目標。希望達成的效果為輸入一段手語單字影片，預測出在分類目標中的單字分類。而因線上相關資源的不足，在前述選定好單字分類目標後，另請兩位不同性別的人在不同背景分別針對各個單字拍攝三種有動作差異的手語單字影片，方有足夠的多樣性資料能夠去訓練影像動作辨識的神經網路。由於目前尚未有人嘗試建立台灣的手語單字資料庫以及相關利用神經網路辨識手語單字的研究，往後相關的研究可以以本論文建立的手語資料庫和準確率作為基準去比較與改良模型架構。

### 4.1.2 資料分析

在去除掉一字多義以及單一手語對應到多個單字的情況，最後選定有 63 種單字去做分類，並且確保單字跟手語影片有一對一的關係。切割資料集使訓練集: 驗證集: 測試集為 7:1:1(,.)，大約為針對每個單字隨機選 7 部影

4.

片做訓練，一部做驗證，一部做測試。不同訓練集影片的詳細數據 (影片 frame 長度、影片寬高、影片 fps) 列出如下 圖 4.2。前述之三種影片拍攝目的為增進訓練資料量以及個體多樣性，但保有類似的動作分布。故針對一個單字分類其訓練、驗證、測試影片主要相異在人體的輪廓與衣著、手部放置的位置與動作弧線、影片的背景顏色等 圖 4.1，也是此辨識系統要去學習與克服的困難之處

(a) 訓練範例 (牙齒) [22]　(b) 驗證範例 (牙齒) [20]　(c) 測試範例 (牙齒)

圖 4.1: 影片範例

## 4.2 單字辨識系統流程

### 4.2.1 資料前處理

由於影片 frame 長度的相異，而最長的影片含有 181 個 frame，故先透過取樣將影像固定為 200 個 frame。不同影片之每一個 frame i 取樣為

$$i * (200/\text{影片}\,frame\text{長度})$$

，即影片透過 (200 / 影片長度) 比率延伸或縮短至 200 個 frame。而由於預訓練模型 (Inception 網路) 的原先訓練集圖片大小為 224x224，因此將每一個 frame 的短邊照比例縮放至 240 pixel，再將圖片中心裁剪成 224 x 224 大小與 Inception 網路預訓練圖片大小統一。最後 RGB 的值透過標準化縮放至-1 至 1 區間加速模型的收斂。另外在訓練時利用鏡像法將每個

4.

frame 左右對調，期望學習到同一手語但分別用左右手不同表示的情況。

Algorithm 1: 資料前處理

1 將每部手語影片的 frame i 取樣為

$$i * (200/\text{影片}\, frame\text{長度})$$

統一影片至 200 frame;

2 將 frame 裁剪縮放至 224 x 224 與 Inception 網路預訓練資料集相同;

3 將 frame 的 RGB 值透過標準化縮放至-1 至 1 區間;

4 訓練神經網路時利用鏡像法做資料增量;

### 4.2.2 模型架構

將前處理完成之資料輸入進雙流 3D 膨脹卷積神經網路，將預訓練在 ImageNet [16] 和 kinect [1] 的權重加載到模型中，並先固定參數只訓練 3 個 epoch 在分類層，優化器為 Adam 並且 learning rate 為 0.001；最後在透過 17 個 epoch、learning rate 0.0001 訓練整體模型參數，實驗結果與變因分析在第五章分別列出。

4.

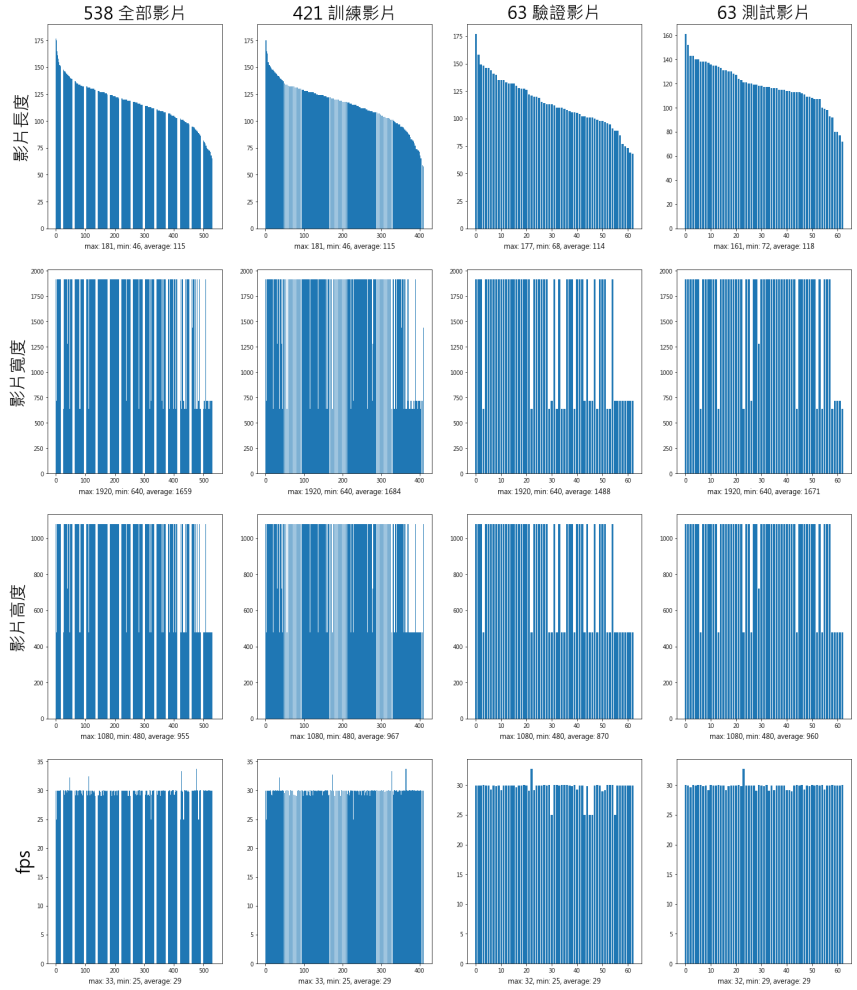

538 全部影片
421 訓練影片
63 驗證影片
63 測試影片
影片長度
影片寬度
影片高度
fps
max: 181, min: 46, average: 115
max: 181, min: 46, average: 115
max: 177, min: 68, average: 114
max: 161, min: 72, average: 118
max: 1920, min: 640, average: 1659
max: 1920, min: 640, average: 1684
max: 1920, min: 640, average: 1488
max: 1920, min: 640, average: 1671
max: 1080, min: 480, average: 955
max: 1080, min: 480, average: 967
max: 1080, min: 480, average: 870
max: 1080, min: 480, average: 960
max: 33, min: 25, average: 29
max: 33, min: 25, average: 29
max: 32, min: 25, average: 29
max: 32, min: 29, average: 29

圖 4.2: 影片資訊分析

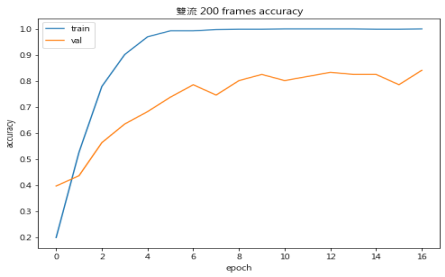
# Chapter 5

# 實驗結果與分析

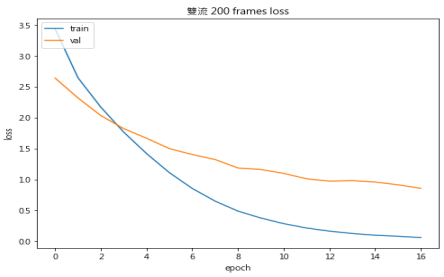
## 5.1 不同模型輸入與取樣比較

### 5.1.1 對影片不同之取樣數量

前述提及將各個影片依照比例伸縮至 200 frames，但可能由於影片 fps 太高亦或是影片關鍵的訊息並不是均勻分布在時間軸上等因素造成太多冗餘的資訊，因此在本節中透過將影片伸縮至不同長度並觀察訓練和測試過程的變化 表 5.1。由下圖 圖 5.1訓練過程可得知在取樣為資料集的平均長度 115 時效果最好，取樣到長度 200 時雖然驗證及測試集的精準度和 115 相差不多，但整體訓練曲線較為震盪，而取樣長度為 40 時則明顯因為降取樣遺失某些 frame 的資訊，進而造成光流計算誤差加大，表現不如以上兩者好。

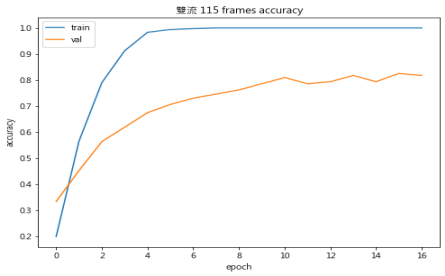



(a) 雙流 200 frames accuracy

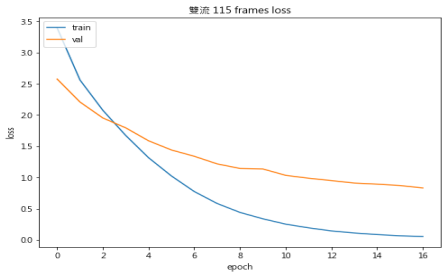



(b) 雙流 200 frames loss

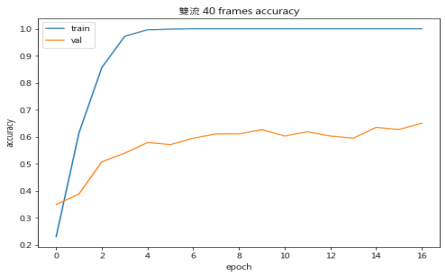
5.



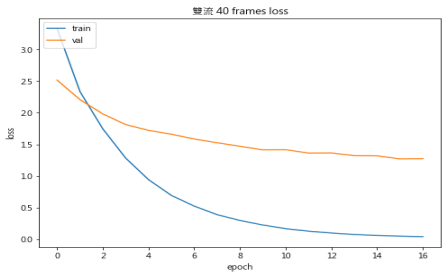

(a) 雙流 115 frames accuracy




(b) 雙流 115 frames loss




(a) 雙流 40 frames accuracy




(b) 雙流 40 frames loss

圖 5.3: 雙流不同取樣數量之測試精準度

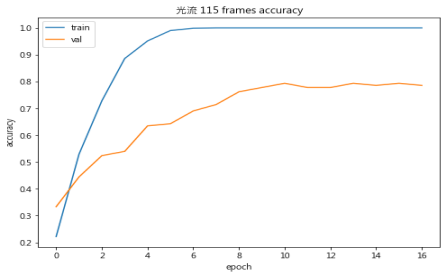
表 5.1: 雙流不同取樣數量之測試精準度

| frame 數量 | top1 測試集準確率 | top3 測試集準確率 | top5 測試集準確率 |
|---|---|---|---|
| 200 | 55 / 63 | 62 / 63 | 62 / 63 |
| 115 | 55 / 63 | 62 / 63 | 62 / 63 |
| 40 | 31 / 63 | 45 / 63 | 45 / 63 |

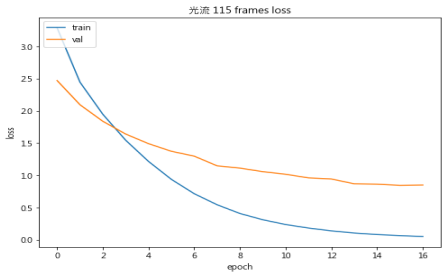
### 5.1.2 RGB vs 光流輸入

固定影片取樣為 115 frames，分別針對單獨 RGB、單獨光流、RGB 融合光流當作輸入給 3D 膨脹卷積神經網路，訓練過程與測試集精確率如下 表 5.2 圖 ??。可以看到光流輸入在訓練過程中的曲線比起 RGB 光滑且穩定上升，並且在大部分值為 0 的情況下 (背景處無光流)，僅透過兩個 channel(x,y 方向的光流) 就能在效能上逼近且超越將原始圖片作為輸入的結果。如同 3.5 節提到的光流是描述人物不變性外觀，相比原始 RGB frame 蘊含了背景、衣著配件等在手語辨識任務中較為不重要的資訊。

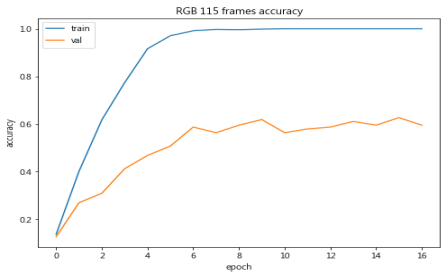
5.



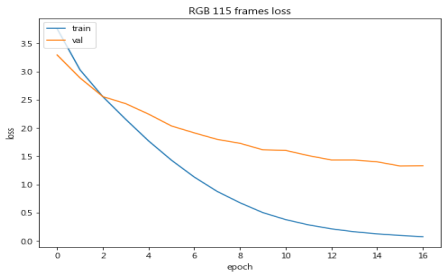

(a) 光流 115 frames accuracy




(b) 光流 115 frames loss




(a) RGB 115 frames accuracy




(b) RGB 115 frames loss

圖 5.5: RGB vs 光流輸入之測試精準度

表 5.2: RGB vs 光流輸入之測試精準度

| 輸入型態 | top1 測試集準確率 | top3 測試集準確率 | top5 測試集準確率 |
|---|---|---|---|
| 單獨光流輸入 | 53 / 63 | 61 / 63 | 61 / 63 |
| 單獨 RGB 輸入 | 51 / 63 | 60 / 63 | 60 / 63 |

## 5.2 錯誤討論與改良

由錯誤表格中 表 5.3可以看出模型的判斷錯誤主要在數字 (手指差異) 和相似的單字手語表示 (井、田)。取測試集之一個月手語單字的關鍵 frame 以及光流結果列出如下分析 圖 5.6 圖 5.7，可以看出光流的結果對 motion 及外觀有較好的不變性，但同時缺乏了細節的表現導致誤判，可能原因是手部跟臉部在過程產生互相遮擋，導致光流在相鄰 frame 最佳化求解產生誤差；而 RGB 作為輸入所含有的不相關訊息太多，導致 RGB 的模型容易過擬合。改良方法則可以從光流及 RGB frame 下手，透過光流進階的去誤

5.

差算法或者同樣使用神經網路模型來計算光流，在訓練時同時學習手語的辨別與光流的計算；RGB frame 可以透過切割手部臉部的演算法去除不相干的訊息與擷取細節特徵，使模型能精準判斷手部細微差異並更具有泛化能力。

(a) 第 12 時間點 frame　(b) 第 47 時間點 frame　(c) 第 73 時間點 frame

圖 5.6: 手語關鍵 frame(一個月)

(a) 第 12 時間點光流　(b) 第 47 時間點光流　(c) 第 73 時間點光流

圖 5.7: 手語關鍵光流 (一個月)

5.

表 5.3: 錯誤討論

| 正確結果 | top3 預測結果 | top3 預測信心機率 |
|---|---|---|
| 一個月 | 一 / 可以 / 一個月 | 82.3% / 4.1% / 1.9% |
| 二 | 一 / 二 / 二十 | 91.5% / 2.1% / 0.7% |
| 七十 | 八十 / 七十 / 四十 | 95.9% / 1.2% / 1.2% |
| 二十 | 三十 / 二十 / 二 | 50.8% / 38.3% / 3.9% |
| 田 | 井 / 田 / 王 | 27.9% / 16.2% / 13.2% |
| 久 | 一同 / 民族 / 打開 | 41.8% / 12.8% / 12.6% |
| 小學畢業 | 一樣 / 小學畢業 / 民族 | 42.0% / 28.6% / 6.5% |
| 父親 | 二十 / 可以 / 父親 | 19.1% / 13.1% / 9.6% |

# Chapter 6

# 結論與未來發展

## 6.1 結論

本文透過建立使用手語資料集以及雙流膨脹卷積網路模型，將目前在影像動作識別上表現良好的深度學習架構轉移到針對台灣手語單字分類的任務上，並透過全面的實驗分析在測試集達到 87.3% 的準確率。而經實驗結果發現透過降低影片的取樣能提升辨識的速度同時保持一定的精準度，如此能將訓練好的模型應用到即時辨識的系統中。

## 6.2 未來發展

除了解決台灣單字手語辨識外，未來更進階的研究為翻譯句子甚至文章的手語。一種方法是使用單字的辨識模型做為骨幹，先將句子中含有的單字辨識出來後結合自然語言模型去進一步重組單字。而對於單字手語辨識本身，除了 5.2 節提到之改善光流算法及 RGB frame 之手臉部位切割；對於泛化重要的點有時間、空間上的變化如人物占比比例、複雜背景與多個人物、手語速度的快慢等。時間上能透過取出關鍵 frame 解決語速的問題；而空間上則能朝語義切割和超解析等方向去研究。最後關於資料集，可以加入更多拍攝者、光線、拍攝角度等以增強多樣性，並針對資料集做完善的統計與變因分析，以求評估不同模型實驗時能更健全客觀。

# Reference


[1] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6299–6308, 2017.

[2] "1-by-1 convolution." [Online]. Available: https://pic4.zhimg.com/v2-e52fc0baf42226e2dfa0f8afc5658c49_r.jpg

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9, 2015.

[4] "global average pooling." [Online]. Available: https://miro.medium.com/max/1538/1*-y6BjnXNHEXaeEdi-eHzvg.png

[5] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," Proceedings of Neural Information Processing Systems (NIPS), 2014.

[6] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, and M. J. Black, "On the integration of optical flow and action recognition," in Pattern Recognition, T. Brox, A. Bruhn, and M. Fritz, Eds. Cham: Springer International Publishing, 2019, pp. 281–297.

REFERENCE


[7] (2010) 手語介紹—手語及台灣手語介紹. [Online]. Available: http://www.cnad.org.tw/ap/news_view.aspx?bid=25&sn=35ef3d27-4e95-4116-a45d-74e76ece9998

[8] Y.-H. Lee and C.-Y. Tsai, "Taiwan sign language (TSL) recognition basedon 3D data and neural networks," Expert Systems with Applications, vol. 36, pp. 1123–1128, 2009.

[9] G. C. Lee, F.-H. Yeh, and Y.-H. Hsiao, "Kinect-based Taiwanese sign-language recognition system," Multimedia Tools and Applications, vol. 75, pp. 261––279, 2016.

[10] Rung-Huei Liang and Ming Ouhyoung, "A real-time continuous gesture recognition system for sign language," in Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 558–567.

[11] C.-T. Hsieh, H.-C. Liou, and L.-M. Chen, "Taiwan sign language recognition system using lc-ksvd sparse coding method," in Proceedings of the 2016 6th International Conference on Machinery, Materials, Environment, Biotechnology and Computer. Atlantis Press, 2016/06, pp. 519–523. [Online]. Available: https://doi.org/10.2991/mmebc-16.2016.112

[12] M. Pahlevanzadeh, M. Vafadoost, and M. Shahnazi, "Sign language recognition," 03 2007, pp. 1 – 4.

[13] B. K. Horn and B. G. Schunck, "Determining Optical Flow," in Techniques and Applications of Image Understanding, J. J. Pearson, Ed., vol. 0281, International Society for Optics and Photonics. SPIE, 1981, pp. 319 – 331. [Online]. Available: https://doi.org/10.1117/12.965761

[14] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l1 optical flow," in Pattern Recognition, F. A. Hamprecht, C. Schnörr,

and B. Jähne, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 214–223.

[15] T. Amiaz, E. Lubetzky, and N. Kiryati, "Coarse to Over-Fine Optical Flow Estimation," Pattern Recognition, vol. 40, pp. 2496–2503, 2007.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Proceedings of Neural Information Processing Systems (NIPS), 2012.

[17] M. Lin, Q. Chen, and S. Yan, "Network In Network," arXiv e-prints, p. arXiv:1312.4400, Dec. 2013.

[18] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1933–1941, 2016.

[19] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in Computer Vision – ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 20–36.

[20] 蔡素娟、戴浩一、陳怡君. (2015) 【台灣手語線上辭典】第三版中文版. 國立中正大學語言學研究所. [Online]. Available: http://tsl.ccu.edu.tw/web/browser.htm

[21] 張榮興. (2011) 台灣手語地名電子資料庫. 嘉義：國立中正大學語言學研究所. [Online]. Available: http://signlanguage.ccu.edu.tw/placenames.php

[22] 林亞秀. (2015) 台灣手語拾遺. [Online]. Available: https://www.twsl.cc/