國立臺灣師範大學電機工程學系

碩士論文

指導教授：林政宏博士

應用於人體骨架動作辨識的
結合快慢網路與注意力自適性圖卷積架構

Integration of Slow-Fast Network and Attention Adaptive Graph Convolutional Network for Skeleton-Based Action Recognition

研究生：蔡 旻 諺　撰

中 華 民 國 一 一 一年 二 月

誌　　謝

兩年多的碩士生活中，我的指導教授林政宏老師給予我不間斷的鼓勵以及支持，打磨我對研究與軟體工程的理解以及執行能力，讓我精益求精，感謝老師對我的指導、關心以及理解。

感謝實驗室的同伴們，鐘陽與欣儀從大學就帶著我，鐘陽時常解惑我對程式所產生的疑惑，欣儀非常關心我，如同我的哥哥與姐姐一般。有你們在實驗室讓我感到非常安心與自在。感謝柏永與承憲，帶著我一起作研究與案子，傾囊相授所有的知識，不分晝夜的共同討論與學習中使我更加茁壯。你們總是站在我的前方，作為我的楷模讓我時常鞭策自己繼續前進。感謝尚德與鼎傑，跟你們在研究室做研究非常的有趣，你們的協助非常地給力。

感謝我親愛的家人的照顧與支持，讓我能專心於研究不用擔心生活，順利完成大學和研究所的學業。

蔡旻諺 2021.11.16

應用於人體骨架動作辨識的
結合快慢網路與注意力自適性圖卷積架構

學生：蔡旻諺　　指導教授：林政宏博士

國立臺灣師範大學電機工程學系碩士班

摘　　要

本論文探討了圖像動作辨識與骨架動作辨識任務，近年來骨架動作辨識任務被快速的發展，發展出藉由圖卷積神經網路結合鄰接矩陣表達人體結構的方式，尤其注重於在圖卷積神經網路中的跨距離連結能力，並學習不同型態的骨架資訊在大型數據集達到更高的準確率。我們認為比起學習多樣的資料型態，注重動作的解析同樣重要，因此引入圖像動作辨識的雙流方法，使用高頻率與低頻率分別解析單一型態的骨架序列，從而提取不同的靜態與動態動作資訊。同時兩流分別作為兩種對於關節點的連結策略，分別注重間格性時間與相鄰時間的連結，並在不同層中穿插靜態與動態特徵的融合層。我們所提出的架構在大型數據集 NTU RGB+D 中的單資料評估為 95.9%的準確率，多資料評估為 96.8%的準確率。實驗結果證實了，我們所提出的方法達到更好的結果。

關鍵字：動作辨識、圖卷積網路、特徵融合

Integration of Slow-Fast Network and Attention Adaptive Graph Convolutional Network for Skeleton-Based Action Recognition

Student：Min-Yan Tsai Advisor：Dr. Cheng-Hung Lin

Department of Electrical Engineering
National Taiwan Normal University

ABSTRACT

This paper discusses RGB-based action recognition and Skeleton-based action recognition tasks. In recent years, skeleton action recognition tasks have been rapidly developed, and a way of expressing human body structure through graph convolutional neural networks combined with adjacency matrices has been developed, with particular emphasis on the cross-distance connection ability in the graph convolutional neural network, and learn different types of skeleton information to achieve higher accuracy in large data sets. We believe that it is equally important to focus on the analysis of actions instead of learning various data types. Therefore, we introduce a two-stream method for RGB action recognition, using high frequency and low frequency to analyze a single

type of skeleton sequence, so as to extract different static and dynamic Action information. At the same time, the two streams are used as two connection strategies for joint points, respectively, focusing on the connection between inter-lattice time and adjacent time, and interspersed with fusion layers of static and dynamic features in different layers. The accuracy of our proposed architecture is 95.9% in single-data evaluation and 96.8% in multi-data evaluation in the large dataset NTU RGB+D. The experimental results confirm that our proposed method achieves better results.

Keywords：Action Recognition, Graph Convolutional Network, Feature Fusion

目　　錄

頁次

圖 目 錄

頁次

表 目 錄

頁次

第一章　緒論

1.1 研究背景與動機

人體動作辨識(Human Action Recognition) 正被廣泛的使用在多項應用領域上，包括人機互動、自駕車、影像監視與機器人自動化，在當前深度學習領域的重要性不斷上升。動作辨識最早使用RGB-Based的方法並被快速的擴展[1-8]，彩色影像(RGB Image)是由色彩資訊以及其形狀所組成，而動作辨識的關鍵在於有效連結不同時間中的空間訊息，通常藉由 3D卷積神經網路或是 2D卷積神經網路結合 1D卷積神經網路來萃取空間維度與時間維度的特徵。然而堆疊靜態圖像的影片所擁有的動態特徵是不足的，為此雙流架構被提出，其中[4]提出使用光流(Optical flow)強調動作的軌跡變化，使模型的一流專注在人體動作變化的特徵。此種雙流架構雖然補足了動態特徵，卻使得模型的運算量更加龐大。近年因應深度感測器的應用與發展，尤其延伸出基於人體骨架(Skeleton-based)的動作辨識[9-23]。優點是比起RGB影像骨架資訊更加的輕量。不易受複雜的背景資訊以及環境的光源所影響，如圖 1-1 所示。缺點是骨架樣本能使用的資訊量有限，需要經過預處理或是使用深度攝影機取得樣本資訊。在最近幾年的發展中這種感測器的發展與推廣使得這項缺點正逐漸消失，並被提出公認的大型數據集NTU RGB+D，使骨架動作辨識的發展追上影像的動作辨識。

圖 1-1 RGB 影像中擁有大量與人體動作無關的圖像資訊，像是背景物件與背景的紋理，以及會影響辨識效果的事件，如拍攝光源的亮暗程度以及遮擋關係。

而最新的研究中大多依照 STGCN(Spatial Temporal Graph Convolutional Network)[26]所提出的基於時空間的圖卷積網路架構，藉由人體關節作為 graph vertices，將人體結構在不同時間的自然連結關係作為 graph edges，如圖 1-2 所示。先以圖卷積神經網路(Graph Convolutional Network，GCN)萃取人體空間特徵，再透過時間卷積網路(Temporal Convolutional Network，TCN)對時間軸進行 1D 的卷積，藉此萃取時間軸上的特徵。此種作法在大型數據集表現良好，並成為現在大多數 Skeleton-Based 方法的比較對象或基準(baseline)。在後續的研究中發現，首先被提出的 graph structure 為了因應不同的動作，以鄰接矩陣(adjacency matrix)定義三種固定的拓樸圖(Topology)策略，如圖 1-3 所示。但仍可能因為不能被學習的關係，無法表達不自然的關節連結關係，進而導致 GCN 產生侷限性。例如: 原先的鄰接矩陣策略中，左手與右手的關聯程度為 0，會導致無法表達兩隻手並用的動作，如圖 1-4 中使用平板電腦的動作。儘管雙手位置的變化有緊密的關係，都將因為關聯程度的關係被忽略不計，因此使分類結果錯誤。

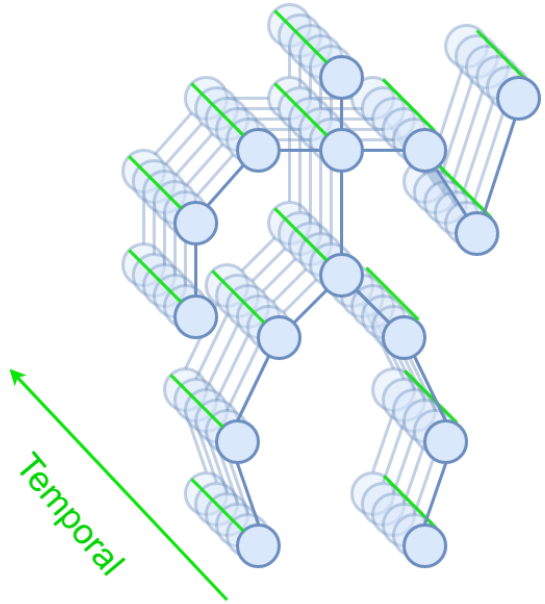


圖 1-2 STGCN 依照兩種 Edge 所構成: Spatial edges(藍色)依照人體在空間中的連結所構成，Temporal edges(綠色)為在空間中相同的關節在不同時間上的連結

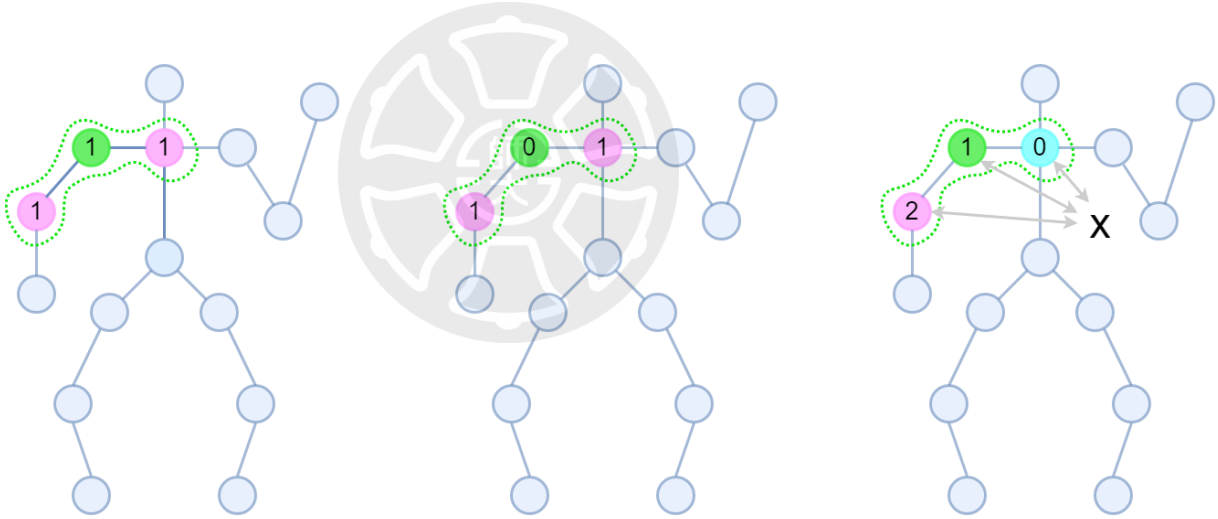
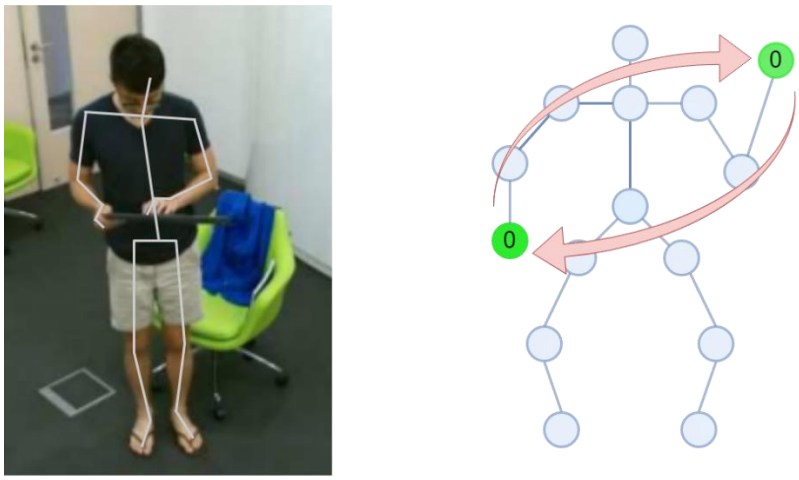


圖 1-3 STGCN 中所使用的三種拓樸圖策略: (a)策略 1 將對中心關節與其鄰近關節採用等價的權重計算。(b)策略 2 則以中心關節為根節點，權重為 0，到鄰近節點做距離加成，權重為 1。(c)策略 3 依照空間上的相對關係分為 3 群的動作，3 群分別是根節點、向心群與離心群，節點將根據它們到骨架重心與根節點的距離為依據，劃分為三種不同的權重。



(a) (b)

圖 1-4 (a)圖為 RGB 影像中骨架資訊的範例圖。(b)圖為拓樸圖中雙手的關係，不論左手對右手或是右手對左手，彼此的關係都將因為所定義的距離過遠而被設置為 0。

因此包括：2s-AGCN[28]、SGN[30]與 Dynamic gcn[31]，提出將 GCN 中所使用的鄰接矩陣中所有關節點之間的關係為可學習，例如左手跟右手可以藉由學習產生關聯，藉此提升 GCN 的辨識效果。

MS-AAGCN(Multi-Stream Attention Adaptive Graph Convolutional Network)[29]在原先 2s-AGCN 中加入了 STC Attention ，STC Attention 分別對不同的維度添加注意力機制:從空間的角度來看，不同的動作通常與特定的關節有關連性。從時間的角度來看，一個動作的構成可能包含多個階段，不同的階段發生的時間(幀)對於最終辨識的評估具有不同程度的重要性。從特徵的角度來看，卷積特徵圖的多個 channel 包含多個語意的級別，每個的 channel 對於不同的動作樣本產生不同的解釋。該注意力模組以自適性的方式重新校正了不同樣本的關節、時間和 channel 的激活權重，使模型更能夠專注在重要的特徵上。在 MS-AAGCN 中完整描述了四種資料型態，包括原始樣本的關節以及後續延伸的骨骼、關節動量與骨骼動量，如圖 1-5 所示，補充了單一資料型態資訊可能不充足的情況，使得模型的評估有了更多的指標。

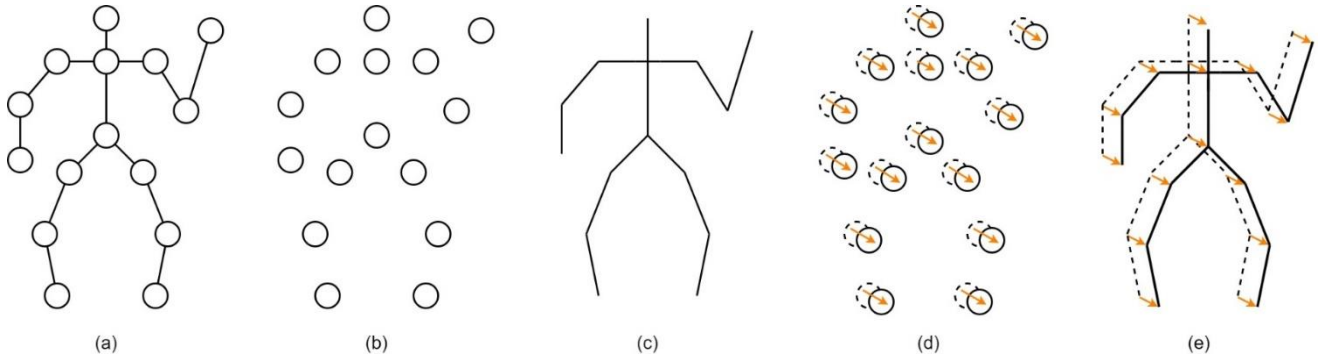


圖 1-5 各種資料型態之表示。(a)圖為一般人體軀幹之示例圖，包含關節(joint)與骨骼(bone)。(b)關節為初始的骨架關節點資料型態。(c)骨骼為鄰接關節點之座標相減得出。(d)關節動量(joint-motion)與(e)骨骼動量(bone-motion)分別為相應型態的資料，在相鄰前後畫面中的座標相減得出。

在加入以上四種資料型態後，後續的研究更專注在資料的特徵萃取上，尤其增強關節點在跨時空上的關聯性，如圖 1-6 所示。MS-G3D[32]提出 G3D module 結合 Multi-Scale TCN ，其中 G3D module 一次計算空間與時間的三維卷積，Multi-Scale TCN 則以多個尺度的 TCN 進行時間軸的解析，構成專為人體架構而生的 Inception 架構。Shift GCN[33]提出 Shift GCN module 與 Shift TCN module，分別在 GCN 與 TCN 中將特徵進行位移，增強不同特徵之間的連結性。因為圖卷積具有結構性的架構，因此增加跨時空的機制是困難的，這使得後續以圖卷積為基礎的設計更加複雜，更多的參數需要被調整，同時依賴四種型態的特徵來提升整體準確度。

我們認為除了在圖卷積架構中有效學習跨時空的關聯性外，在基於骨架的方法中探討頻率變化在動作的可解釋性也是同等重要的，而非仰賴多型態的資料強化來彌補動作在時間變化上的資訊。

假設動作可以被分為兩類，分別為高頻率的動作與低頻率的動作。高頻率的動作通常在一段時間內不斷重複相似的變化，像是刷牙、鼓掌、讀書、敲打鍵盤、

使用平板電腦。低頻率的動作在一段時間內進行不同階段的變化直到動作完成，如坐下、跳起來、穿外套、跌倒、握手、擁抱。上述所提出的例子皆為大型數據集 NTU RGB+D 中的實例。若可以依據兩種不同的動作給予不同時間長度的關聯性，則可以使模型對動作在時間關聯性的解釋變得更多樣。進而提升模型在大型數據集的準確率。

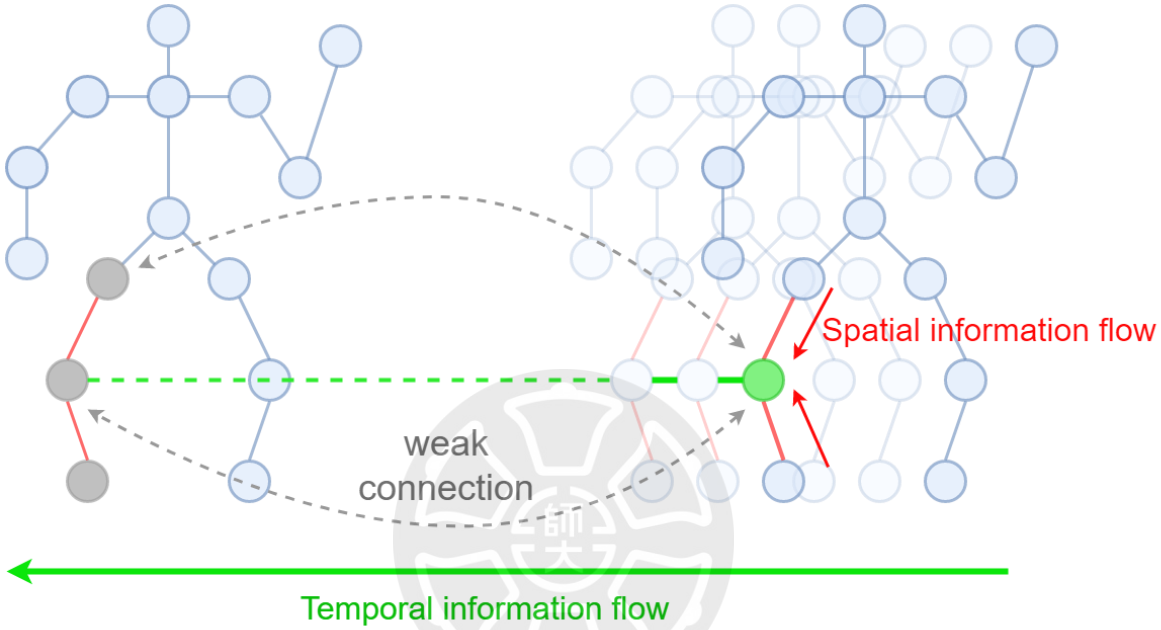


圖 1-6 綠色結點為當前的中心關節，紅色線表示關節間有較強的空間連接關係，綠色線為時間資訊流的前進方向，灰色節點為連接性較弱的節點，儘管在空間中為同一個關節，間隔較遠的時間也會使關聯性變弱。

1.2 研究目的

本研究所欲達成之目的如下條列:

1. 我們期望以 RGB-Based 的 SlowFast 架構結合 Skeleton-Based 時空間的圖卷積方法，使模型有能力專注在對人體動作的頻率變化上。使辨識模型在大型數據集的準確率可以得到提升。
2. 我們希望以外部簡潔的架構降低內部圖卷積的複雜程度，藉由雙流架構的設計提供強化拓樸圖連結的方式。使用更少的資料形態下萃取更多潛在的時空間資訊，降低對多資料型態的依賴。

1.3 研究方法概述

本研究所提出之方法可分為兩大部分:

1. 基於研究目的,本論文提出一種 SlowFast AAGCN 架構,將過去 RGB-Based 所盛行的雙流架構結合現今 Skeleton-Based 圖卷積的方法。雙流分別為快速流(Fast Stream)與慢速流(Slow Stream),藉由在輸入資料時跨間隔提取,使雙流的輸入為不同速率的動作序列,快速流專注在需要全時段關注的動態動作,慢速流則專注緩慢變化的靜態動作,如圖 1-7 所示。在不同速率採樣下,圖卷積中的關節連結性同樣得到兩種策略的增益,最後在兩流之間加入特徵融合層,有效連結不同速率採樣下的特徵資訊。
2. 藉由將所設計的架構實驗在各式大型數據集中,找出最合適的快速流與慢速流的架構配置。探討快速流與慢速流共同分析的比重。將所設計的架構與原先的 backbone 進行比較。

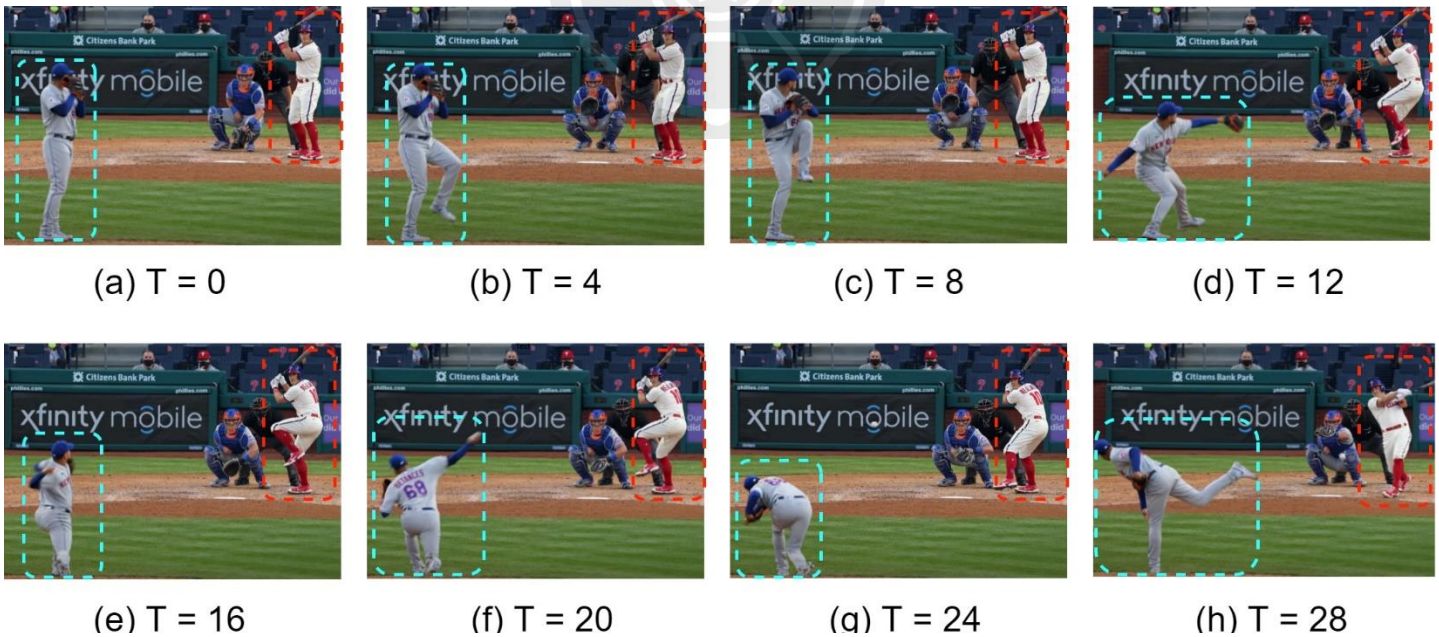


圖 1-7 將一段序列的 RGB 影像以間隔 4 幀進行採樣,藍色框為動作快速變化的投手,紅色框為動作緩慢變化的打擊者。快速變化的動作需要頻繁的採樣動作變化量,緩慢變化的動作跨間隔採樣後可以使變化量放大。

1.4 研究貢獻

本論文提出的基於骨架的 SlowFast 架構易於嵌入於各式特徵萃取 Backbone 上，使得後續的研究可以降低對多型態資料的依賴以及降低跨時空連結能力所需的複雜度。我們的實驗結果證明了過去 RGB-Based 探討動作原理的方式在 Skeleton-Based 上同樣有效，實現了我們所提出的目標以單資料型態的準確度達到接近其他方法中多資料型態的準確度，雙資料型態的準確度則超越許多使用四種資料型態的方法。將我們所提出的架構實驗在三個不同的大型數據集中，我們所提出的 SlowFast+AAGCN 在大型數據集 NTU RGB+D 中的 Xview 指標中，採用單資料型態中達到 95.9%的準確率，雙資料評估為 96.6%準確率，相較於原先的 MS-AAGCN 提升了約 1%的準確率，以上兩個結果中比目前 state-of-the-art 的方法更佳。另外我們的方法在 NTU RGB+D 120 中的 Xset 多資料型態達到 88.9%，Xsub 多資料型態達到 86.3%，在 Northwestern-UCLA 中的多資料型態達到 97.2%，這個結果也比目前 state-of-the-art 的方法更佳。

1.5 論文架構

本論文分為五個章節：緒論、文獻探討、研究方法、實驗結果以及結論與未來展望，以下條列各章節內容簡介。

第一章 緒論：概述圖像與人體骨架之動作辨識任務的背景，選擇人體骨架動作辨識任務結合圖像任務作為研究之動機，條列研究目的與簡述研究方法，最後闡述研究貢獻與此論文架構。

第二章 文獻探討：針對與本研究方法相關之文獻討論介紹，主要從過去圖像動作辨識任務、人體骨架動作辨識模型、基於人體骨架之圖卷積神經網路三大方向做探討。

第三章 研究方法：說明所提出的人體骨架動作模型結合圖像任務之雙流架構、以及對人體動作解析之設計所使用的研究方法。

第四章 實驗結果：介紹實驗配置之設備、實驗參數設計、實驗流程以及實驗結果之分析。

第五章 結論與未來展望：最後對此研究提出的方法與實驗結果總結，並對未來研究做延伸與改善方向。

第二章　文獻探討

基於深度學習之動作辨識方法於 2013 年後被大量提出，於本章節將整理過往之動作辨識方法。2.1 探討動作辨識任務於深度學習領域之演進。2.2 介紹基於人體骨架圖之神經網路架構。2.3 介紹近年基於圖卷積之人體骨架動作辨識方法。

2.1 動作辨識

動作辨識的經典演算法 iDT[1] 在 2013 年被提出，作為深度學習踏入動作辨識領域之最好的演算法，並影響後續的研究方向。其所提出的構想包括:密集採樣特徵點、特徵軌跡的追蹤與根據軌跡進行特徵提取。

而後，卷積神經網路(Convolutional neural network, CNN)在圖像領域大放異彩，基於 CNN 的方法不需手動提取特徵，而其性能亦超過大多數的手工提取方法。在動作辨識任務當中，其輸入資料相較一般圖像任務多出時間維度以表達人體動作隨時間之變化，為此於過往研究當中以 3D 卷積神經網路(3D CNN)[2][5]與循環神經網路(Recurrent neural network, RNN)等結合 RGB 影像的空間資訊(RGB-Based)以及時間序列資訊之架構為主。其中，為了產生連續靜態圖像隨時間變化的特徵，首先提出基於空間與時間資訊之 CNN 架構[4]，其保留原始基於 RGB 影像的空間資訊，並引入基於時間資訊的光流架構(Optical flow)使樣本的時間變化可被有效學習；[6]在後續實驗進行時間與空間資訊之雙流特徵融合，證實了雙流特徵融合的方法對雙流之架構有極大的影響性；[7]基於預訓練的 2D CNN 模型並拓展時間維度，提出 3D 卷積結合雙流架構達到更好的效果；SlowFast networks[8]根據不同的資料特徵提取速率以進行雙流架構的設計，與過往架構不同，其架構可藉由不同頻率採樣下的動作變化資訊以進行動作解析。

於上述研究當中，基於 RGB 圖像的樣本(RGB-Based)對動作的解析容易受背景影響，使分類結果產生誤差，如光源明亮程度造成的像素資訊變化、衣服與背

景的紋理的近似度、人物與背景雜物的遮擋關係等等。除此之外，背景資訊可能造成模型當中存在大量冗餘運算，為此部分研究轉而探討基於人體骨架(Skeleton-Based)的方法。

Skeleton-Based 的方法聚焦於人體關節的空間關係，使大量冗餘運算的問題被解決。然而，若直接將人體骨架資訊視做偽圖像與序列向量，並套用過去之 RGB-Based 的方法，其模型之準確率將會十分低落。其主要原因為不完整數據集導致的骨架資訊缺失，從而使得骨架的表達能力被限制進而影響模型之泛化能力。為此，當代研究方向開始朝向提取人體關節點之特徵並解析不同關節點之間的關係，如 HBRNN-L[11]將人體分作四肢與軀幹，並以雙向循環網路(Bidirectional recurrent neural networks, BRNN)得到各個部位的特徵，最後將各個部位的特徵進行連接以取得關聯性；[13]設計了空間與時間維度的二維長短期記憶模型(Long Short-Term Memory, LSTM)結合 skeleton-based Tree Traversal 以表達關節點的關係。上述方法的提出儘管提升了動作辨識準確率，其效果仍不如 RGB-Based 的方法。

2.2 圖卷積神經網路

在卷積神經網路(CNN)的設計當中，其針對局部區域中所有資訊進行線性變換並取得平均值以產生特徵，亦即所有線性變換後之資訊其重要程度相等，使得 CNN 通常應用於定義在歐幾里得空間之樣本當中，相關樣本通常具有規律的資料排列方式，如一維的聲音資訊、二維的圖像資訊與三維的影片資訊。然而，在圖(graph)的架構下各頂點(vertex)之間的關係可能隨著輸入資料而有所不同，例如社交網路、生物網路，如圖 2-1 所示，使得此種定義於非歐基里德空間的數據，較難採用 CNN 進行特徵提取，因而於當代研究當中發展出圖卷積網路(GCN)。在動作辨識任務當中，人體骨架中的各個關節點之間具有獨特的連結架構，將各關節點連接起來即可形成基於人體之骨架圖(skeleton graph)，基於圖的特性使其亦可採用圖卷積網路進行特徵提取，如 STGCN[26]在 2018 年建構了基於人體骨架

的時空間圖卷積網路(Spatial Temporal Graph Convolutional Network，STGCN)，其以人的關節作為圖的頂點(vertex)，以前後相鄰的時間做為邊(edge)，並以鄰接矩陣表示不同關節點之間的關聯性。除此之外，考量到人體的運動可以分為近重心運動與偏重心運動，STGCN 定義了三種基於人體骨架物理連結之鄰接矩陣以輔助不同關節點的特徵聚合。除了空間特徵的提取之外，如何有效連結時間與空間的特徵亦是一大課題，為此於該文章中採用時間卷積網路(Temporal Convolutional Network，TCN)以聚合不同時間之相同關節點資訊，實驗結果表明 STGCN 之架構在大型動作辨識數據集 NTU RGB+D[16]中得到最好的結果，並證明圖卷積的架構可更完整表達人體關節點之間的關聯性，從而使得後續的人體骨架動作辨識通常以 STGCN 為範本並進行深入探討。

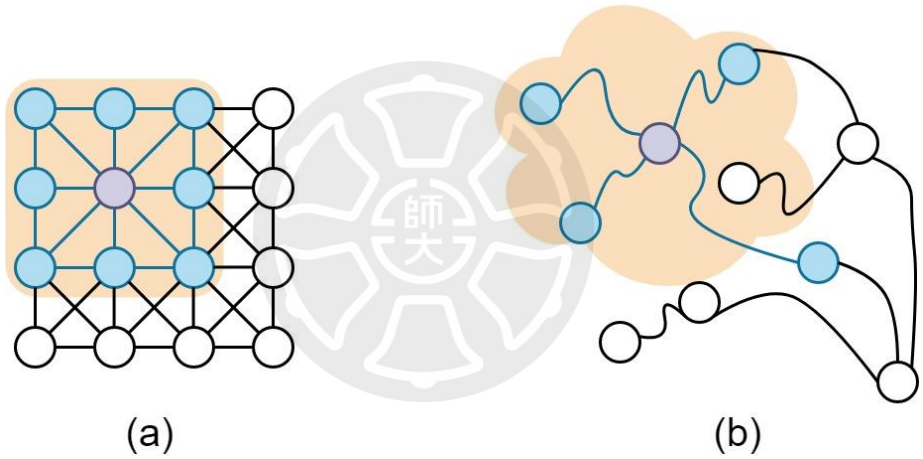


圖 2-1(a)為具有規律的資料排序，可以藉由 CNN 進行採樣。(b)為具有不規則性結構的資料，不適合直接以 CNN 進行採樣。

2.3 基於圖卷積之骨架動作辨識

ST-GCN 的提出使得後續針對人體骨架之動作辨識任務的研究通常以該架構為基準並針對其特性進行深入探討。舉例而言，STGCN 所定義的三種固定鄰接矩陣之策略不一定能滿足所有的動作，如拍手此動作當中雙手之關節點的關聯性即無法透過固定鄰接矩陣進行表示。除此之外，圖卷積之特徵聚合能力可能隨著關節點彼此距離過遠而受到侷限。若是該動作所使用的重點關節距離過遠將導致

圖策略的學習受到局限性，為解決上述問題，2s-AGCN[28]引入自注意力機制以產生自適應的鄰接矩陣，該鄰接矩陣之資訊將隨著輸入資料而有所不同。SGN[30]同樣使用注意力機制以進行鄰接矩陣之修正，並以階層式架構分析關節點與時序的語意資訊以增強時空資訊的融合。 Dynamic gcn[31]提出結合了所有關節的上下文資訊以學習跨時序之關節之間的關係。研究數據表明固定鄰接矩陣結合自適應鄰接矩陣生成之拓樸結構具有更強的泛化能力。

在資料型態方面，MS-AAGCN[29]將兩關節點之特徵向量相減以產生的關節點之間的骨骼資訊，並計算關節點與骨骼隨時間之位置變化量以產生兩種動量(motion)資料，如圖 2-2 所示。實驗結果顯示上述四種資料經分別訓練產生之結果具有互補關係，亦即多型態資料之輸出結果的綜合評估可大幅提升準確率，從而使得多型態資料的準確度成為評估標準之一。

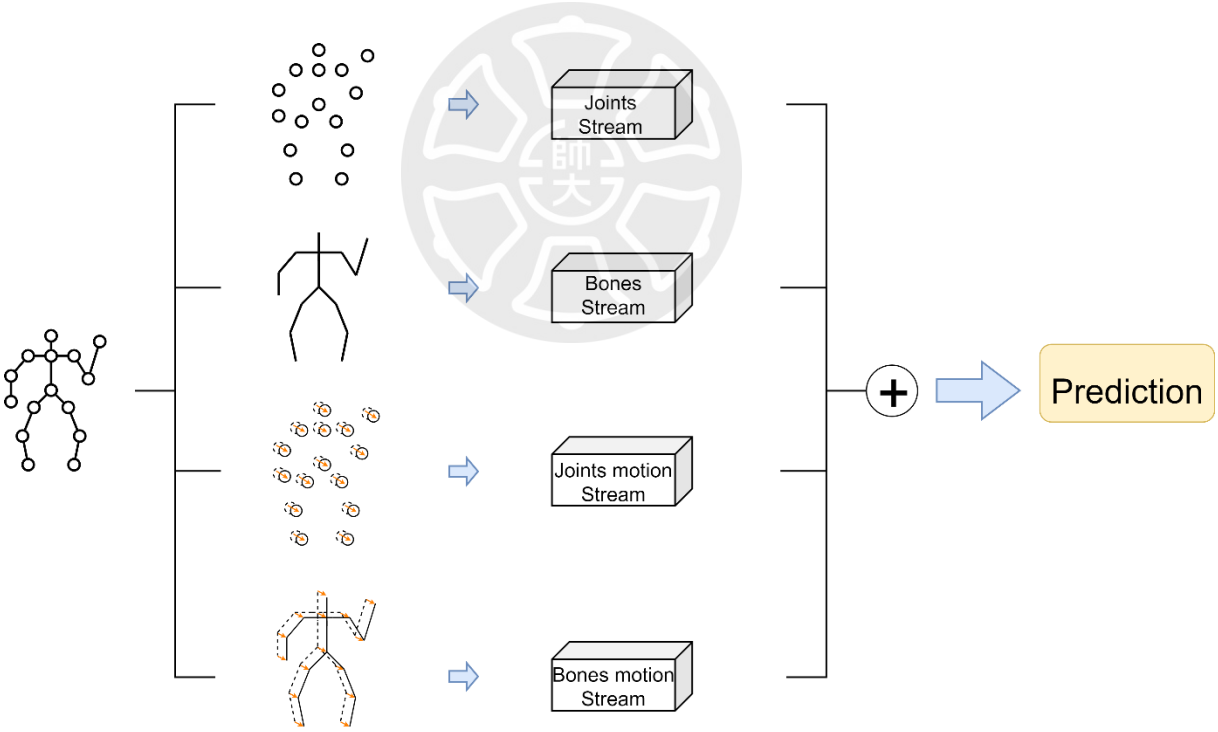


圖 2-2 將不同的資料型態分別經過不共享權重的模型訓練，而後將個別輸出以權重加總進行預測。

在後續的研究當中，通常專注在圖卷積網路的特徵提取能力上，其中跨時間與空間(Cross-spacetime)的特徵提取受到大量關注，如圖 1-6 所示。MS-G3D[32]

根據鄰接矩陣內容的不同以設計多組並行之 GCN 進行特徵提取，並採用使用多尺度(Multi scale)的 TCN 以增強跨時間的連結能力。[33]提出 Shift GCN 與 Shift TCN，透過偏移空間特徵與時間特徵使較遠距離之關節的相連資訊可被有效聚合。綜上所述我們可以發現後續研究通常依賴學習多資料型態的擴充以取得更好的準確度，並修改 GCN 與 TCN 之架構以提升其在跨距離連結的能力使其可更好的學習資料特徵。

第三章　研究方法

本章節將介紹本論文提出的基於人體骨架動作辨識的 SlowFast AAGCN 網路架構之設計與研究方法，為了在不增加圖卷積內部的複雜程度為前提下，提升圖卷積對於動作的解析能力，降低多資料型態的依賴性，我們結合過去 RGB-Based 的雙流架構[8]於現今 Skeleton-Based 的動作辨識方法上，降低對 motion 資料的需求，只需要關節與骨骼即可以藉由此種解析頻率的雙流模型獲得更多潛在的動作變化資訊。人類視覺對動作的直接理解除了包含形狀、顏色、深度外還有速率，相較於 RGB 影像，骨架資訊對於形狀、深度以及其變化速率擁有更多的訊息，並且其輕量的特性可以改善雙流架構運算量大的缺點。若是將一序列的動作切成影格，間格性的選取片段進行觀看，人類還是有很大機率能夠分出來動作，因為這時候速率的表達更加強健。我們的想法是將骨架的動作特徵分成時間特徵與空間特徵，其中時間特徵透過快速流萃取，而空間特徵透過慢速流萃取。時間特徵為快速變化的動作特徵，必須注重鄰近時間點的變化，因此需要高頻率採樣才能夠分辨動作。空間特徵為較長時間的動作變化，注重骨架在長時間的關聯性，每個影格中的空間特徵。透過雙流解析動作的方法能夠有效限縮時間長度，專注在重要的時間特徵與空間特徵上。

整體架構如圖 3-1 所示。此架構方法主要分為三個部份，細節將個別於以下章節中說明：3.1 介紹架構的特徵萃取單元選擇；3.2 說明快速流(Fast Stream)與慢速流(Slow Stream)之設計與原理；3.3 講解快速流與慢速流之特徵融合。

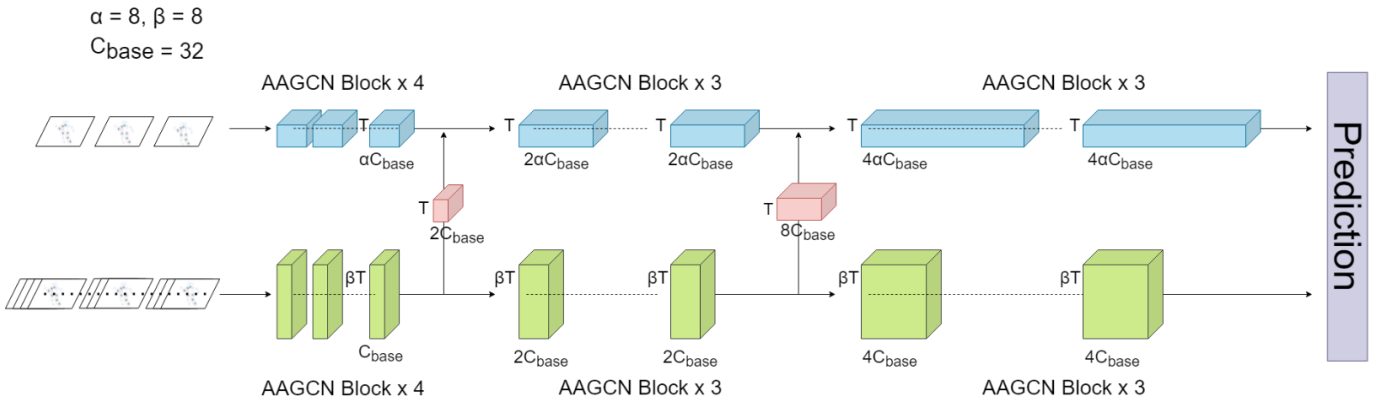


圖 3-1 SlowFast AAGCN Networks 架構圖。藍色的部分為慢速流的特徵萃取單元，綠色的部分為快速流的特徵萃取單元，紅色的部分為特徵融合層。在快速流與慢速流的最後一層將特徵合併進行分類預測。T 與 C_{base} 分別表示了各層特徵在時間維度與 channel 維度的比例。

3.1 特徵萃取單元

現今的方法通常由圖卷積神經網路搭配時間卷積網路以構成特徵萃取單元(Block)，並以多個特徵萃取單元構成辨識模型，於當代研究中通常會複雜化特徵萃取單元以期可提取更好的特徵。在 SlowFast AAGCN 架構當中，特徵萃取單元是可以被替換的。我們採用 AAGCN 之特徵萃取單元(AAGCN Block)並結合 SlowFast 之技巧以期望在不增加特徵萃取單元之內部複雜度的前提下，增強外部之整體架構以提升關節點在跨時空間的連結能力。AAGCN Block 中包含 AAGCN(Attention-adaptive graph convolution)與 TCN，在兩層的中間穿插 BN 層(Batch Normalization layer)與激勵函式(Activation function)，如圖 3-2 所示。其中 GCN 的部分針對輸入資料產生鄰接矩陣並結合全局之可學習的鄰接矩陣以進行特徵提取，如圖 3-3 所示。除此之外，該單元亦引入注意力機制以針對空間、時間與特徵維度進行特徵強化，如圖 3-4 所示。

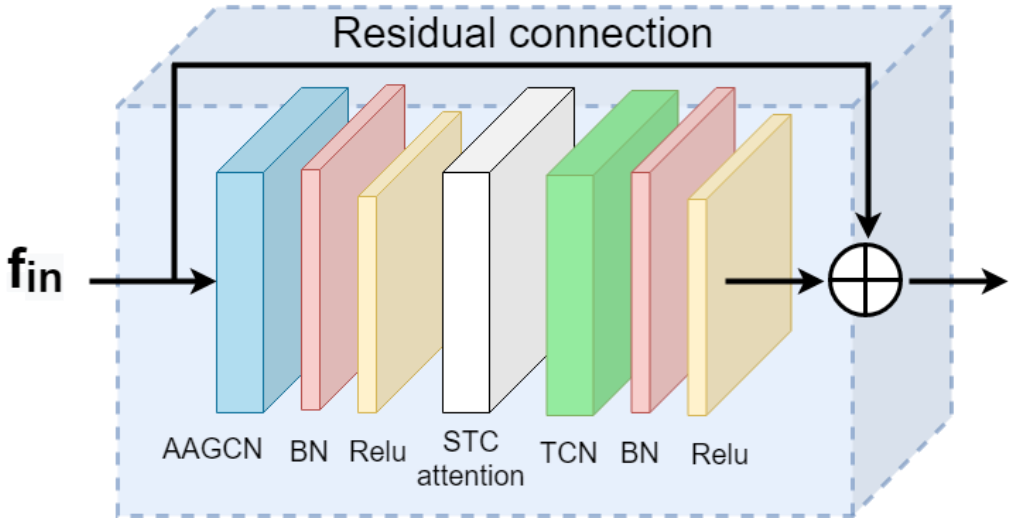


圖 3-2 AAGCN Block 內部結構圖。

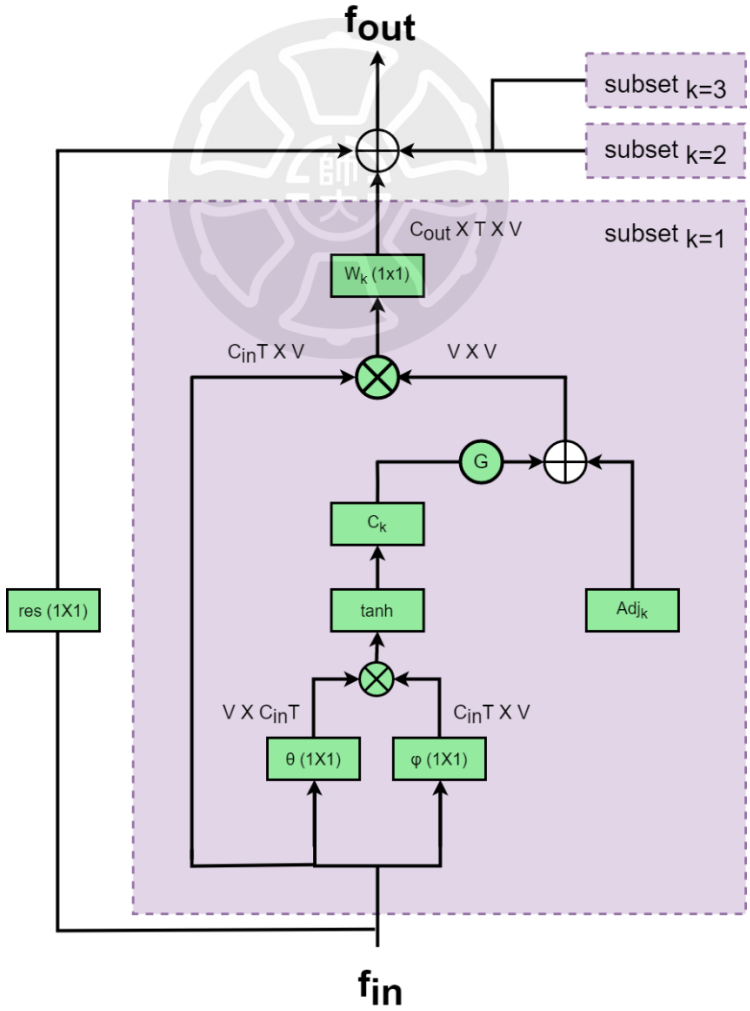


圖 3-3 AAGCN 層的實例化，$subset_k$為當前所使用的拓撲圖策略子集合。

圖卷積神經網路(GCN)用於聚合骨架的關節點特徵，其關鍵在於該網路當中存在鄰接矩陣使關節與關節之間的關係可被有效表達。在 AAGCN Block 當中其針對每個輸入資料X_{in}皆會產生一個鄰接矩陣C_k^{ℓ}(data-dependent graph)，其中$f \in \mathbb{R}^{C\times T\times V}$為連續的骨架動作，$C_k^{\ell}$之計算如式 3-1 所示：

$$C_k^{\ell} = tanh({f_{in}}^T W_{\theta k}^{(\ell)^T} W_{\varphi k}^{(\ell)} f_{in}) \quad (3\text{-}1)$$

其中，輸入資料經過φ與θ以進行特徵變換，而後採用點積(dot product)以計算經特徵變換的資訊於 Embedding Space 之中的相似程度，而後採用雙曲函數(tanh)對其進行非線性轉換，經上述過程即可取得與輸入資料相依之鄰接矩陣C_k^{ℓ}。

除此之外，AAGCN Block 當中存在全局之可學習的鄰接矩陣，其根據三種物理關節連接策略進行初始化並可被訓練樣本修正，以 $Adj_k^{(\ell)}$表示為此層的第 k 個策略的鄰接矩陣，在 data-dependent graph 與 global graph 進行融合時可以藉由可以學習的 G 做為 Gate 控制兩圖評估時的權重，特徵萃取單元中圖卷積最後的輸出 f_{out} 表達為式(3- 2)：

$$f_{out} = \sum_{k=0}^{K} W_k^{(\ell)} f_{in}(G^{(\ell)} C_k^{(\ell)} + Adj_k^{(\ell)}) \quad (3\text{-}2)$$

有效的融合了同一個時間中的空間特徵後，以時間維度的卷積對各時間點的關節做連結，為 TCN 的部分，並且根據快速流與慢速流做不同的 TCN 配置，如下式(3- 3)與(3- 4)所示：

$$f_{Fast}^{\ell} = Conv_{1,1}(f_{Fast}^{\ell}, W_{t_{Ft\times 1}}^{(\ell)}) \quad (3\text{-}3)$$

$$f_{Slow}^{\ell} = Conv_{1,1}(f_{Slow}^{\ell}, W_{t_{St\times 1}}^{(\ell)}) \quad (3\text{-}4)$$

其中 $Conv_{n,m}$ 表示 Stride 為(n, m)的卷積運算， $W_{n\times m}^{(\ell)}$則表示此次 Convolution 所使用的 kernel size 為n × m，第一個維度是對時間維度，第二個維度對應的則是對空間維度(關節資訊)，而快速流與慢速流在 kernel size 的時間感受野配置不同，將跟隨採樣速率做變動。

將式(3- 2)與分別與式(3- 3)、(3- 4)進行傳遞後得到完整的時空間特徵的連結關係。而後將時空間特徵以三個維度的注意力機制(STC-attention)精煉，如圖 3- 4 所示，空間維度如式(3-5)，時間維度如式(3-6)，Channel 維度如式(3-7)。

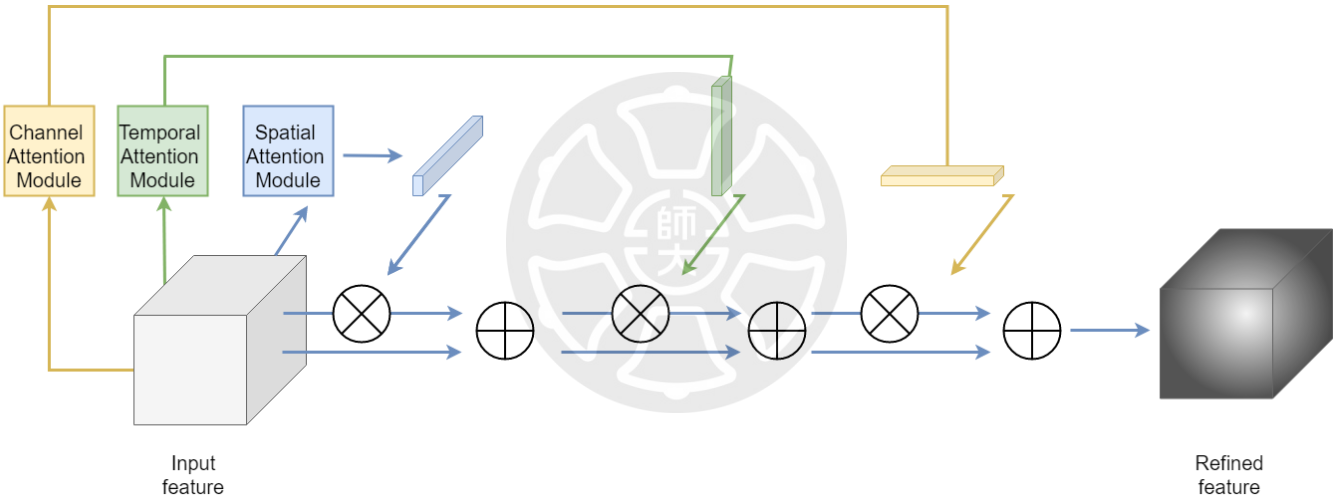


圖 3-4 分別以 Spatial Attention Module、Temporal Attention Module 與 Channel Attention Module 進行三個維度的精煉。

$$M_s = \sigma(g_s(Avgpool(f_{in})) \tag{3- 5}$$

$$M_t = \sigma(g_t(Avgpool(f_{in})) \tag{3- 6}$$

$$M_c = \sigma(W_2\left(\delta(W_1\left(Avgpool(f_{in})\right)\right)) \tag{3- 7}$$

其中 σ 為激勵函式 $Sigmoid$，δ 為激勵函式 $Relu$， g_s 與 g_t 代表 1D Convolution，W_1 與 W_2分別為兩個全連結層的權重，$W_1 \in \mathbb{R}^{(C\times\frac{C}{r})}$，$W_2 \in \mathbb{R}^{(\frac{C}{r}\times C)}$， $M_s \in \mathbb{R}^{1\times1\times V}$，$M_t \in \mathbb{R}^{1\times T\times1}$，$M_c \in \mathbb{R}^{C\times1\times1}$，以殘差的方式(residual manner)乘上原先的輸入，實現了專注在各自維度的特徵。

3.2 Slow Fast Structure

我們提出基於人體骨架的 SlowFast AAGCN 為一種雙流模型，其可被套用於任何單流模型中。將原先的單流模型視作快速流(Fast Stream)，加入另一相同模型為基底的慢速流(Slow Stream)，兩流將分別以不同的時間間隔 τ_s 與 τ_s 對輸入樣本進行提取，如圖 3-5 所示。輸入資料的不同使得兩流的拓樸圖學習策略有所區分。其中快速流之輸入資料採樣全時段的特徵樣本，使其專注於快速變化的動態動作。強調鄰近時間點的資料關聯。而慢速流之輸入資料的採樣頻率較低，使其更專注於緩慢變化的靜態動作。從而使得在時序上相距較遠的空間資訊得以被拉近學習。

然而，基於慢速流的輸入資料採樣頻率較低，其產生之特徵圖參數量與快速流有明顯的落差，為此我們選擇在慢速流的特徵維度進行補償。其目的除了平衡參數量以避免兩流資訊比重失衡之外，亦可使慢速流在少量的時間資料中挖掘更多潛在的空間特徵。除此之外，為了避免跨時間的採樣資訊被 TCN 層再度壓縮，我們選擇保持慢速流與快速流在各層的時間採樣步伐(stride)，亦即原始模型之 TCN 的 stride 將會設置為 1。

$$\beta = \frac{slow\ \ stream\ \ frame\ \ rate}{fast\ \ stream\ \ frame\ \ rtae} \tag{3-8}$$

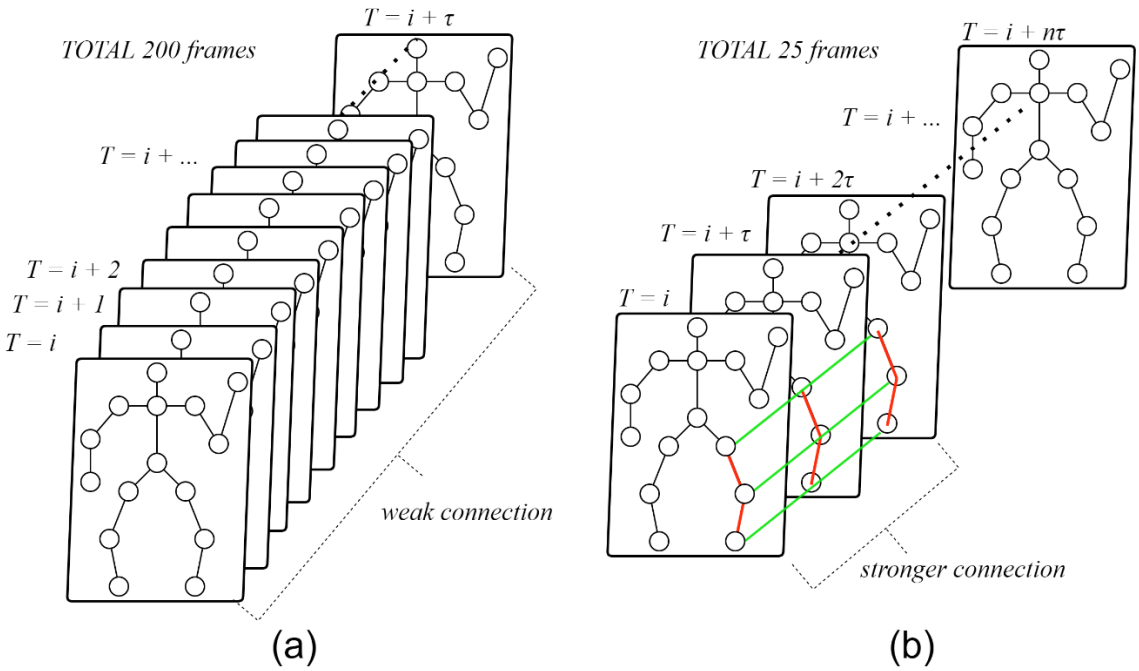


圖 3-5 綠色實線為鄰近時間點的連結，紅色實線為空間中相鄰的資訊連結(a)圖為快速流的特徵連結策略。強調鄰近時間點的連結，而在相隔 τ 幀的特徵連結變得脆弱。(b)圖為慢速流的特徵連結策略。強調長時間距離的連結，藉由間隔提取拉近原先相鄰 τ 幀的資訊，使相隔 τ 的連結強度被拉近為相隔 1。

3.3 快速流與慢速流的特徵融合

由過去的研究已經證實不同流之間的特徵融合將會影響雙流模型的性能[6]，而前述章節中的設置皆為了雙流之特徵資訊的有效融合。如在時間採樣步伐的調整與慢速流的特徵維度補償。上述設置在網路當中僅存在單流的情況下可能會降低模型之泛化能力，但在雙流架構的狀況下合理的設計則可達成有效的特徵融合。

不同於多資料型態之輸入資料之共同評估(co-analysis)方法，我們選擇在網路架構之特定層中將快速流的特徵資訊進行時間維度的聚合並融入到慢速流當中。而非採用針對慢速流資料之時間維度的上採樣融合方法，以避免上採樣過程產生的雜訊。而上述側向連接被我們稱為特徵融合層(Fusion layer)，其計算方法如式 3-8 所示，而整體示意圖如圖 3-6。在快速流資訊融入慢速流資訊的做法當中，快速流的特徵聚合主要強調相鄰畫面之間的關節點關係。而慢速流則強調跨越時間

尺度之關節點連結。兩者資訊的融合使慢速流可取得部分快速流之相鄰區域資訊，而慢速流的特徵維度補償亦將平衡雙流之特徵圖的比重。使特徵融合層更專注在快速流的特徵聚合。為使聚合的特徵盡可能被保留，我們僅使用含有權重的卷積層並不加入偏移資訊，並採用基於時間的感受野 κ 與步伐 τ 進行特徵聚合。在此設置化需要被調整的參數更少，其特性也容易被直觀理解。

$$f_{Slow}^{\mathscr{b}} = Cat_{Channel}\left(f_{Slow}^{(\mathscr{b}-1)}, Conv_{\beta,1}\left(f_{Fast}^{(\mathscr{b}-1)}\right)\right) \quad (3\text{-}9)$$

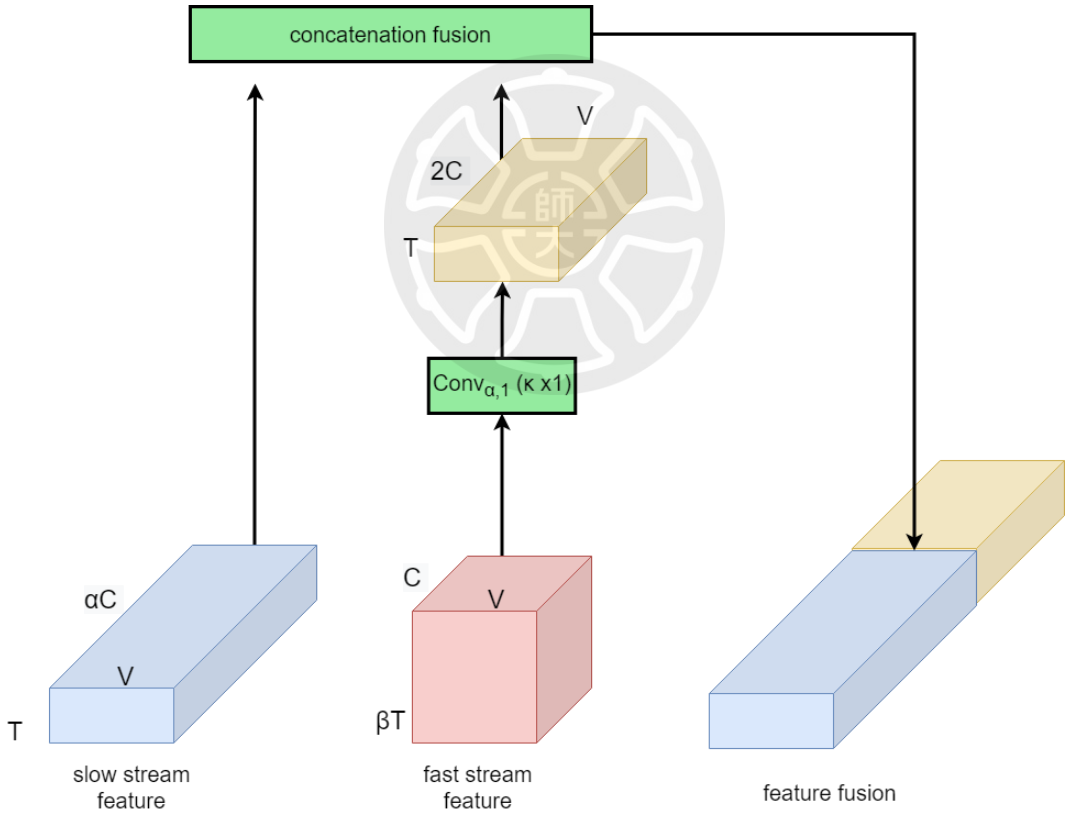


圖 3-6 慢速流與快速流特徵融合示意圖。

3.4 實驗設置

對於我們所提出 SlowFast AAGCN 架構的實驗，將分為兩個部分，第一個部分為消融實驗(ablation studies)，顯示我們所提出架構的各個模塊設置對整體準確度的影響。第二個部分則是將我們所提出的方法與過去的方法在三個大型數據集進行比較。

在消融實驗中，首先需要探討的即是快速流與慢速流共同分析的比重，有無平衡兩流的參數量對此種雙流架構的影響。第二步要探討的為不同速率下對動作進行分析所產生的影響，比較設置不同速率下的單流模型的準確率。可以觀察出在數據集中不同速率可以提供的準確率，進而以不同速率組合而成的雙流模型提供最好的互補性。第三步為針對特徵融合層的設置，包含特徵融合層設置的數目，以及不同時間長度的聚合對特徵融合層的影響。

在大型數據集之實驗中，我們選擇了三個骨架動作辨識任務當前公認的大型數據集，NTU RGB+D、NTU RGB+D 120 與 Northwestern-UCLA，並分別在以下進行詳細的介紹。

NTU RGB+D 數據集包含 56578 部影片分為 60 種動作類別，共有兩種評估基準，Cross-subject (Xsub)與 Cross-view (Xview)。Xsub 中共有 40 位受試者，在訓練集與測試集個包含 20 位受試者，受試者介於 10 歲至 35 歲之間，每個受試者被分配一個 ID 編號。由於不同人會有不同的動作習慣，人體的骨架大小與骨骼長度也不相同，以測試模型是否有能力藉由學習不同受試者的動作進行正確的動作預測。Xsub 中共有 40091 筆樣本，測試集中共有 16487 筆樣本。Xview 中同時設置了三個攝影機進行拍攝，三個攝影機分別以相同的高度，不同的水平視角設置：+ 45 度、0 度、- 45 度。每個受試者被要求執行動作兩次，一次朝向左側的攝影機，一次朝向右側的攝影機。其中 1 號攝影機始終觀察 45 度視角，而 2 號與 3 號攝影機分別觀前方視角與側視角。以 1 號攝影機的拍攝樣本作為測試集，2 號與 3 號的拍攝資樣本為訓練集，以不同的視角的動作評估模型的能力。Xview

中訓練集包含 37646 筆樣本，測試集包含 18932 筆樣本。

NTU RGB+D 120 數據集以 NTU RGB+D 數據集做為基礎擴增 57367 部影片，包含 113945 部影片，並分為 120 種動作類別。共有兩種評估標準，Cross-subject (Xsub) 與 Cross-setup (Xset)。Xsub 中共有 106 位受試者，受試者介於 10 歲到 57 歲之間，身高由 1.3m 至 1.9m，每個受試者被分配一個 ID 編號。以不同的受試者的動作評估模型的能力，訓練集中含有 63026 筆樣本，測試集含有 50919 筆樣本。Xset 中使用 32 種集合設置，不同的設置中改變了位置和背景。如同原先 NTU RGB+D 中的 Cross-View 一樣，設置三台攝影機，三個攝影機分別以相同的高度，不同的水平視角設置(+ 45 度、0 度、- 45 度)進行拍攝。每個受試者被要求執行動作兩次，為了進一步增加攝影機視角數量，在不同的集合設置中改變了攝影機的垂直高度，與拍攝對象的拍攝距離。

Northwestern-UCLA 數據集使用三個 Kinect 相機從多個視角同時拍攝，包含 1494 個影片剪輯，涵蓋 10 種動作類別，每個動作由 10 個不同的受試者進行動作，我們遵循[10]中使用相同的評估基準，前兩個攝影機的樣本作為訓練集，最後的攝影機樣本做為測試集。

3.5 骨架資料的處理

人體骨架的資訊大多是採納關節點於空間中的三軸座標(x, y, z)，因而其數據容易受拍攝位置以及角度所影響。為了方便與其他方法比較，對於 NTU RGB+D 與 NTU RGB+D 120 兩個大型數據集的資料，我們參考 2s-AGCN[28]所使用的資料前處理方式：若輸入資料為多人骨架，則首先使用熵(Entropy)值進行排列，並選擇熵值最高的兩個人物作為輸入樣本。換句話說輸入資料主要選擇骨架動作變化最大的兩個目標。由於數據集中每個樣本執行動作的時間長度不一，而最長時間的樣本擁有 300 個幀的時間長度，因此若樣本不足 300 個幀，則不斷重複相同

的骨架序列直至滿足 300 個幀的長度。在骨架校正的部分，將每個人的骨架座標之關節點座標值減去的 21 號關節(頸椎下端)的座標值，亦即以各自骨架的頸椎下端為中心，並將旋轉整體骨架使得兩肩膀關節點相連的直線與 X 座標軸平行，而頸椎下端之關節點與尾椎相連的直線 Z 座標軸平行，以進行骨架資訊的校正。對於 Northwestern-UCLA 的資料，統一將所有樣本的時間長度縮放至 64 個幀，對不足 64 個幀的樣本以內插法進行上採樣，而對於超過 64 個幀的樣本進行下採樣。在骨架校正的部分，我們參考 Shift GCN[33]，將每個人體骨架之關節點座標值減去的 2 號關節(脊柱中間)的座標值，以進行骨架資訊的校正。

第四章　實驗結果

此章節將說明本論文嘗試之所有實驗。將在以下章節進行詳細說明：4.1 為 SlowFast AAGCN 架構設計所做之消融實驗；4.2 為大型數據集之實驗，包含：以 MS-AAGCN 作為基準(baseline)，在大型數據集 NTU RGB+D 中比較各式指標的準確率；與過去所有提出的方法在 NTU RGB+D[16]、 NTU RGB+D 120[34]、Northwestern-UCLA[10]等大型數據集中進行比較，比較指標包含單資料型態(Joint)，雙資料型態(Joint & Bone)，多資料型態，並說明各項指標所代表的意義；4.3 將詳細說明我們的實驗設備以及針對不同數據集所設置的訓練細節。

4.1 消融實驗

本次研究對架構之探討共分為三個部分依序說明，**4.1.1 快速流與慢速流之參數量平衡，4.1.2 快慢流之採樣速率比，4.1.3 特徵融合層之設置，**實驗皆以 NTU RGB+D X-View 作為基準。

4.1.1 快速流與慢速流之參數量平衡

由於快速流與慢速流之採樣間隔 τ 不同，導致特徵萃取單元之輸出在時間維度產生 α 倍$(\frac{\tau_S}{\tau_F})$之差距，進而影響到雙流架構最後在特徵融合時快速流資訊佔有明顯的比重。表 4-1 中探討了雙流架構對於慢速流之 channel 維度放大 β 倍之補償效果，以及在雙流架構中將快速流所帶有的資訊量降低，並以慢速流進行輔助所能得到的提升。我們以 AAGCN 單流模型(Fast only) $\tau_F = 1$作為基準，而 ABCDEF 組為雙流實驗組，所設置之快速流採樣間格$\tau_F = 1$，慢速流採樣間格$\tau_S =$

8，其中 AB 組與 CD 組分別在不同 base channel 情況下進行參數量平衡與否之實驗。BDEF 組則是在兩流的參數量平衡下以不同 base channel 進行設置。

AB 組與 CD 組之結果顯示，在不同的 base channel 情況下，對慢速流之補償確實起了作用。若是不對慢速流進行補償(β = 8)，以此平衡融合時之總資訊量，將如一開始所設想的慢速流所能提出的資訊差異無法被顯示出來，兩流所提供的資訊量會在最後合併時以快速流佔比較多，雙流架構甚至會使辨識效果降低。在 BDEF 組的實驗中，我們認為 B 組在 base channel 為 64 的情況下設置平衡的雙流架構，使模型參數量變得過大而導致雙流架構的準確率提早達到飽和。這個問題也曾在[46]被發現與進行實現。而在 DE 組顯示出，在降低模型的參數量後比起原先的單流模型有明顯的辨識率提升，表示可以降低快速流的資訊量(將 base channel 由 64 降至 32)，並以慢速流合奏得到更好的效果。而若是在僅有快速流的情況下降低資訊量將直接影響到模型準確度。其中 EF 組以更少的 channel 數實驗雙流的互補性，以 F 組與原先 base channel 為 64 的單流模型進行比較，可以發現我們以更少的運算量仍得到了不錯的準確率。

表 4-1 快速流與慢速流參數量平衡對準確度之影響

Method	Base Channel	α,β	Parameter	Acc(%)	Improvement
Fast only	64	-	3.78M	95.4	-
Fast only	32	-	0.98M	94.6	- 0.8
A	64	8, 1	6.8M	94.8	- 0.6
B	64	8, 8	145M	95.4	+ 0.0
C	32	8, 1	1.75M	95.0	- 0.4
D	32	8, 8	36.1M	95.9	+ 0.5
E	16	8, 8	9.24M	95.6	+ 0.2
F	8	8, 8	2.37M	95.2	- 0.2

4.1.2 快慢流採樣速率比

表 4-2 將探討單流情況下採樣間格(τ)之設置對快速流與慢速流之影響，以及雙流情況下快速流與慢速流分別所設置之τ_F、τ_s與雙流採樣速率比率(α)之比較。表中的 EFGH 實驗組將顯示雙流分別設置之速率(τ_F,τ_S)與彼此之速率比率在何種情況下是最佳的。此實驗之所有設置皆以 half-channel 進行，快速流的輸入時間長度固定為 200 幀，而慢速流的輸入時間長度則依據其採樣速率(τ_s)得出$\frac{200}{\tau_S}$。

在單流快速流(Fasr only)$\tau_F = 1$與$\tau_F = 2$的情況下可以顯示出快速流是否需要完全連續的骨架序列資訊。而單流慢速流(Slow only)不同τ_s的設置可以顯示出，慢速流以不同間隔採樣的差異性。在擁有獨立的單流模型的結果後，就能以 EFGH 組觀察雙流架構下兩流在不同採樣間格設置下的互補性。

首先 Fast only τ_F= 2 與 Slow only τ_s= 4 分別為單流時效果最好的兩組，然而雙流的結果下 F 組得到更好的結果。我們認為 Slow only $\tau_S = 4$ 在單流時可以更泛化的提取快動作與慢動作的特徵，但卻缺少提取更長時間特徵的特性。因此可以提取更長時間特徵的快速流($\tau_S = 8$)跟快速流($\tau_F = 1$)配合後有更多的互補性。而若是將 G 組與 F 組進行比較，可以顯示出在同樣有慢速流$\tau_s = 8$的輔助下，儘管 Fast only τ_F= 2 有更好的表現，雙流時快速流以$\tau_F = 1$進行配置表現出更強的互補性。在 Slow only $\tau_S = 16$中可以發現過長的採樣間格導致準確率有明顯的衰退，在 G 組與 H 組擁有相同快速流設置的情況下，這個衰退也無法在擁有快速流的情況下獲得互補，也可以說是$\tau_s = 8$的設置下擁有更好的泛化能力，$\tau_s = 16$已經可以被認定是極端的採樣間格。

結果顯示出最好的設置為$\tau_F = 1$與$\tau_s = 8$的情況下。特別的是基於骨架的 SlowFast 架構中顯示出與基於影像的 SlowFast 架構中都是以這個數值產生最好的結果。最佳的動作速率比在不同資料的情況下是吻合的，可能因為動作辨識對這兩種速率很敏感。

表 4-2 不同速率下快速流與慢速流之比較與速率對雙流之影響

Method	τ_F	τ_S	α	Acc(%)
Fast only	1	-	-	94.6
Fast only	2	-	-	95.0
Slow only	-	4	-	95.5
Slow only	-	8	-	94.9
Slow only	-	16	-	93.0
E	1	4	4	95.8
F	1	8	8	95.9
G	2	8	4	95.7
H	2	16	8	95.3

4.1.3 特徵融合層設置

在快速流與慢速流確實能進行互補的前提下，此章節將探討作為快速流與慢速流溝通渠道之特徵融合層。特徵融合層將對快速流之時間維度進行壓縮，使輸出特徵之時間維度對齊慢速流之時間維度。而 SlowFast AAGCN 整體架構嚴謹的依據時間速率的採樣而建構，因此我們同樣必須細膩的處理特徵融合層的時間感受野κ。並且實驗了三種融合方式: None Fusion layer 為僅在最後層之輸出做串接，N=2 為在快速流與慢速流的第四層與第七層輸出做連結，N=3 在第一層、第五層與第八層輸出做連結。在 stride = 8 的情況下，不同的 κ 可以解釋對快速流的時間維度特徵進行不同比例的聚合。κ = 3 在聚合時捨棄了最多的快速流資訊而導致準確率下降最多。在 κ = 5 與 κ = 7 的情況下，捨棄部分快速流的資訊比起全採樣(κ = 9)有更好的表現。

N=2 與 N=3 的比較可以發現，僅在中層與高層語意的情況下進行融合有些微的提升，我們認為可能是在淺層保有慢速流的獨立性使結果有更好的表現。而若是沒有任何融合層作為渠道，結果甚至比單流模型低落，證實了融合層的必要性。

表 4-3 特徵融合層感受野與融合次數之影響

SlowFast($\tau_F = 1, \tau_S = 8$)	Acc(%)
Fusion layer κ = 3	95.4
Fusion layer κ = 5	95.9
Fusion layer κ = 7	95.8
Fusion layer κ = 9	95.7
Fusion layer numbers N (κ = 5)	Acc(%)
None Fusion layer	95.3
N = 2	95.9
N = 3	95.8

4.2 大型數據集之實驗

首先將我們所提出之架構與原先的 AAGCN (baseline)在大型數據集 NTU RGB+D[16] 上進行評估，結果顯示在表 4-4。在套用我們的架構後原先被提出的四種資料型態的準確度都得到了大幅的提升。

我們比較了我們的方法與過去被提出的方法，以單資料型態(joint)與雙資料型態(joint & bone)為指標在 NTU RGB+D 中進行比較，結果顯示在表 4-5 與表 4-6。結果顯示我們所提出之方法在單資料型態與雙資料型態下皆達到了 state-of-the-art(SOTA)。並且我們的架構在僅使用單資料型態下的準確度，非常接近過去方法的雙資料型態準確度。而在雙資料形態下，我們的方法同樣接近過去方法的多資料型態的準確度。結果顯示出我們所提出之架構在僅有單資料型態與雙資料型態中能夠獲更多能有效分類動作的特徵。

表 4-4 SlowFast AAGCN 與 baseline(MS-AAGCN) 在四種資料型態下的評估

Methods	Xsub(%)	Xview(%)
baseline Joint	88.0	95.1
baseline Bone	88.4	94.7
baseline Joint Motion	85.9	93.0
baseline Bone Motion	86.0	93.1
SlowFast AAGCN Joint	89.3	95.9
SlowFast AAGCN Bone	90.0	95.2
SlowFast AAGCN Joint Motion	87.3	94.0
SlowFast AAGCN Bone Motion	87.6	93.7

表 4-5 SlowFast AAGCN 與過去所提出之方法在 NTU RGB+D Joint 單資料形態下之評估比較

Methods	Xsub(%)	Xview(%)
2s-AGCN [28]	-	93.7
MS-AAGCN[29]	88.0	95.1
Shift GCN[33]	87.8	95.1
DC-GCN+ADG[36]	88.2	95.2
MS-G3D[32]	89.4	95.0
CTR-GCN[41]	89.9	94.8
MST-GCN[42]	89.0	95.1
SlowFast AAGCN	89.3	95.9

表 4-6 SlowFast AAGCN 與 過去所提出之方法在 NTU RGB+D Joint&Bone 雙資料形態下之評估比較

Methods	Xsub(%)	Xview(%)
2-AGCN [28]	88.5	95.1
MS-AAGCN[29]	89.4	96.0
Shift GCN[33]	89.2	95.5
DC-GCN+ADG[36]	89.7	96.0
MS-G3D[32]	91.5	96.2
MST-GCN[42]	91.1	96.4
SlowFast AAGCN	91.2	96.6

最後將我們所提出之架構與過去達到 SOTA 的方法在三種大型數據集 NTU RGB+D、 NTU RGB+D 120 與 Northwestern-UCLA 進行比較，結果分別顯示在表 4-7、表 4-8 與表 4-9。在 NTU RGB+D 與 NTU RGB+D 120 中與近年所提出的方法十分接近。而我們在 Northwestern-UCLA 的準確率達到了 SOTA。Northwestern-UCLA 數據集比起 NTU RGB+D 與 NTU RGB+D 120，在訓練集與測試集規模上小了非常多，所需要分類的類別也僅有 10 類。結果顯示我們所提出的方法在此種規模下的數據集有非常好的效果，大幅超越過去所提出的方法。

表 4-7 SlowFast AAGCN 與 過去達到 state-of-the-art (SOTA)的方法在 NTU RGB+D 之比較

Methods	Xsub(%)	Xview(%)
Ind-RNN[25]	81.8	88.0
HCN[44]	86.5	91.1
STGCN[26]	81.5	88.3
SGN[30]	88.5	94.5
AGC-LSTM[43]	89.2	95.0
DGNN[27]	89.9	96.1
2s-AGCN[28]	88.5	95.1
MS-AAGCN[29]	90.0	96.2
DC-GCN+ADG[36]	90.8	96.6
PA-ResGCN-B19[35]	90.9	96.0
Shift GCN[33]	90.7	96.5
Dynamic GCN[31]	91.5	96.0
MS-G3D[32]	91.5	96.2
DDGCN[38]	91.1	97.1
CTR-GCN[41]	92.4	96.8
MST-GCN[42]	91.5	96.6
SlowFast AAGCN	91.7	96.8

表 4-8 SlowFast AAGCN 與 過去達到 state-of-the-art (SOTA)的方法在 NTU RGB+D 120 之比較

Methods	Xsub(%)	Xset(%)
ST-LSTM[13]	55.7	57.9
GCA-LSTM[15]	61.2	63.3
Rotclips+MTCNN[45]	62.2	61.8
SGN[30]	79.2	81.5
2s-AGCN[28]	82.9	84.9
DC-GCN+ADG[36]	86.5	88.1
PA-ResGCN-B19[35]	87.3	88.3
Shift GCN[33]	85.9	87.6
Dynamic GCN[31]	87.3	88.6
MS-G3D[32]	86.9	88.4
CTR-GCN[41]	88.9	90.1
MST-GCN[42]	87.5	88.8
SlowFast AAGCN	86.7	88.5

表 4-9 SlowFast AAGCN 與 過去達到 state-of-the-art (SOTA)的方法在 Northwestern-UCLA 之比較

Methods	Top1(%)
Lie Group[12]	74.2
Actionlet ensemble[9]	76.0
HBRNN-L[11]	78.5
Ensemble TS-LSTM[14]	89.2
AGC-LSTM[43]	93.3
Shift GCN[33]	94.6
DC-GCN+ADG[36]	95.3
CTR-GCN[41]	96.5
SlowFast AAGCN	97.2

4.3 訓練細節與實驗設備

所有實驗均在具有 PyTorch 深度學習框架的 Tesla V100-SXM2-32GB 上進行。採用 Nesterov 動量 (0.9) 的隨機梯度下降 (SGD) 作為優化策略。對於 NTU-RGB+D [16]與 NTU-RGB+D 120[34]數據集，進行[28]中所採用的預處理，而後在每次訓練與評估中隨機取 0~99 中的整數作為起始的幀，依序選取 200 幀作為輸入。每次訓練的批量大小為 40。選擇交叉熵作為損失函數來反向傳播梯度。權重衰減設置為 0.0002。訓練 epoch 設置為 120，並且在前 5 個 epoch 中使用了預熱策略以使訓練過程更加穩定。學習率設置為 0.2，並依序在 epoch [40, 50, 70, 80, 90, 110] 以 learning rate [0.2, 0.1, 0.02, 0.01, 0.002, 0.0004, 0.00016] 進行訓練。對於 Northwestern-UCLA[10]數據集，每次訓練的批量大小為 16。選擇交叉熵作為損失函數來反向傳播梯度。權重衰減設置為 0.0004。訓練 epoch 設置為 120，並且在前 5 個 epoch 中使用了預熱策略以使訓練過程更加穩定。學習率設置為 0.1，並依序在 epoch [60, 100] 以 learning rate [0.01, 0.001] 進行訓練。

第五章　結論與未來展望

5.1 結論

藉由實驗結果，證明了我們基於骨架的方法結合圖像任務的架構，所提出的 SlowFast AAGCN 確實能夠在大型數據集得到更好的表現。以不添加複雜模塊的情況下，提升了原先特徵萃取單元的準確度。以單資料型態與雙資料型態達到準確度的大幅提升，藉此降低多資料型態的需求。RGB-based 解析時間速率的雙流架構結合 Skeleton-based 的特徵萃取方法在骨架動作辨識任務上非常有效。驗證了我們對此實驗所提出的假設。

5.2 未來展望

經過所有實驗與討論，整理出以下數點作未來改進方向：

1. 為了在架構上實現平衡的兩流架構並與 baseline 比較使得參數量提升，未來可以依據數據集的不同，設置不同的模型參數量設置。
2. 此種 SlowFast 架構可以降低對輸入樣本時間長度的需求，但在訓練中會受樣本挑選所影響，尤其是數據集規模較小的情況，未來可以對不同的數據集的預處理做更多的討論。
3. 由於特徵萃取單元的不同，以 SlowFast 架構進行雙流設計，會有不同的模型參數量變化以及適應性問題。

參 考 文 獻

[1] Heng Wang, and Cordelia Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2013, pp. 3551-3558.

[2] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, "3D convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence, 35*(1), 2012, pp. 221-231.

[3] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1725-1732.

[4] Simonyan, Karen, and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *arXiv preprint, arXiv:1406.2199*, 2014.

[5] *Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri*, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2015, pp. 4489-4497.

[6] *Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman*, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 1933-1941.

[7] Joao Carreira, and Andrew Zisserman, "Quo vadis, action recognition? new models and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6299-6308.

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, Kaiming He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision* (*ICCV*), 2019, pp. 6202-6211.

[9] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan, "Learning actionlet ensemble for 3d human action recognition," *IEEE transactions on pattern analysis and machine intelligence, 36*(5), 2013, pp. 914-927.

[10] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu, "Cross-view action modeling, learning and recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2014, pp. 2649–2656.

[11] Yong Du, Wei Wang, and Liang Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015, pp. 1110–1118.

[12] Vivek Veeriah, Naifan Zhuang, and Guo-Jun Qi, "Differential recurrent neural networks for action recognition," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2015, pp. 4041–4049.

[13] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang, “Spatio-temporal lstm with trust gates for 3d human action recognition,” in *European Conference on Computer Vision (ECCV)*, 2016, pp. 816–833.

[14] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee, “Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks,” in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017, pp. 1012–1020.

[15] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C Kot, “Skeleton-based human action recognition with global context-aware attention LSTM networks,” in *IEEE Transactions on Image Processing*, *27*(4), 2017, pp.1586-1599.

[16] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang, “Ntu rgb+ d: A large scale dataset for 3d human activity analysis,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1010–1019.

[17] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu, “An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data,” in *Proceedings of the AAAI conference on artificial intelligence* , vol. 31, no. 1, 2017, pp 4263–4270.

[18] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng, “View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition From Skeleton Data,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2117–2126.

[19] Mengyuan Liu, Hong Liu, and Chen Chen, “Enhanced skeleton visualization for view invariant human action recognition,” *Pattern Recognition, 68*, 2017, pp.346–362.

[20] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Ahmed Sohel, and Farid Boussad, “A New Representation of Skeleton Sequences for 3d Action Recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4570–4579.

[21] Bo Li, Yuchao Dai, Xuelian Cheng, Huahui Chen, Yi Lin, and Mingyi He, “Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN,” in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* , 2017, pp. 601–604. IEEE.

[22] Hongsong Wang and Liang Wang, “Modeling temporal dynamics and spatial configurations of actions using two stream recurrent neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 499-508.

[23] Tae Soo Kim and Austin Reiter, “Interpretable 3d human action analysis with temporal convolutional networks,” in *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, 2017, pp. 1623–1631. IEEE.

[24] Congqi Cao, Cuiling Lan, Yifan Zhang, Wenjun Zeng, Hanqing Lu, and Yanning Zhang, "Skeleton-Based Action Recognition with Gated Convolutional Neural Networks," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018, pp. 1–1.

[25] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao, "Independently recurrent neural network: Building a longer and deeper rnn," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 5457–5466.

[26] Sijie Yan, Yuanjun Xiong, and Dahua Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[27] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7912–7921.

[28] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu, "Two stream adaptive graph convolutional networks for skeletonbased action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12026–12035.

[29] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu, "Skeleton-Based Action Recognition With Multi-Stream Adaptive Graph Convolutional Networks," *IEEE Transactions on Image Processing*, vol. 29, 2020, pp. 9532–9545.

[30] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1112–1121.

[31] Fanfan Ye, Shiliang Pu, Qiaoyong Zhong, Chao Li, Di Xie, and Huiming Tang, "Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp 55–63.

[32] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, ZhiyongWang, and Wanli Ouyang, "Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 143-152.

[33] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu, "Skeleton-Based Action Recognition with Shift Graph Convolutional Network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 183-192.

[34] Jun Liu, Amir Shahroudy, Mauricio Lisboa Perez, GangWang, Ling-Yu Duan, and Alex Kot Chichung, "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 42, no. 10, 2020, pp. 2684-2701.

[35] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang, "Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1625–1633.

[36] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu, "Dynamic convolution: Attention over convolution kernels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11030–11039.

[37] Ke Cheng, Yifan Zhang, Congqi Cao, Lei Shi, Jian Cheng, and Hanqing Lu, "Decoupling gcn with dropgraph module for skeleton-based action recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[38] Matthew Korban, and Xin Li, "Ddgcn: A dynamic directed graph convolutional network for action recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 761–776.

[39] Yuya Obinata, and Takuma Yamamoto, "Temporal Extension Module for Skeleton-Based Action Recognition," in *arXiv preprint, arXiv:2003.08951*, 2020.

[40] Ailing Zeng, Xiao Sun, Lei Yang, Nanxuan Zhao, Minhao Liu, and Qiang Xu, "Learning Skeletal Graph Neural Networks for Hard 3D Pose Estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11436-11445.

[41] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu, "Channel-Wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 13359-13368.

[42] Zhan Chen, Sicheng Li, Bing Yang, Qinghan Li, and Hong Liu, "Multi-Scale Spatial Temporal Graph Convolutional Network for Skeleton-Based Action Recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 1113–1122.

[43] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1227–1236.

[44] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *arXiv preprint*, arXiv:1804.06055, 2018.

[45] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid, "Learning clip representations for skeleton-based 3d action recognition," *IEEE Transactions on Image Processing*, 2018, pp. 2842–2855.

[46] Mingxing Tan, Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks. in *International Conference on Machine Learning*," PMLR, 2019, pp. 6105-6114.

自傳

蔡旻諺，1996 年出生於臺北市。

- 新北市昌隆國小
- 新北市頭前國中
- 臺北市大安高級工業職業學校
- 國立臺灣師範大學 電機工程學系
- 國立臺灣師範大學 電機工程學研究所

學術成就

- Min-Yen Tsai, Cheng-Hung Lin, Hsin-Ying Lin, and Jia-Hao Chang. “A Basketball Self-Training System Based on Artificial Intelligence and Machine Vision,” in *Taiwan Society of Biomechanics in Sports (TSBS)*, 2020
- Cheng-Hung Lin, Po-Yung Chou, C. -H. Lin, and Min-Yen Tsai, “SlowFast-GCN: A Novel Skeleton-Based Action Recognition Framework,” in *International Conference on Pervasive Artificial Intelligence (ICPAI)*, 2020, pp. 170-174. (**Best paper runner up award**)
- Cheng-Hung Lin, Min-Yen Tsai, and Po-Yung Chou, “A Lightweight Fine-Grained Action Recognition Network for Basketball Foul Detection,” in *IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, 2021.