科技部補助

大專學生研究計畫研究成果報告

計　畫
　　：　棒球數據自動記錄
名　稱

執行計畫學生：林祐任
學生計畫編號：MOST 109-2813-C-035-036-E
研 究 期 間 ：109年07月01日至110年02月28日止，計8個月
指 導 教 授 ：許懷中

處 理 方 式 ：本計畫可公開查詢

執 行 單 位 ：逢甲大學資訊工程學系（所）

中 華 民 國　110年03月17日

摘要

棒球比賽中，安打上壘的串聯是造成很大的得分關鍵，然而在許多比賽中，不論學生或是職棒比賽，常常留下許多的殘壘，其中有個重大的因素，就是對投手了解的數據不足，這造成賽前需要從有限的影片中研究投手的攻略，或賽中的前幾局需要適應投手及蒐集投手可能投的球種，才能讓擊球者有更好的攻擊成效。

棒球數據自動分析系統，主要想要做到即時辨認球種，建立投手的投球資訊，以及自動化紀錄要比人工紀錄還低的出錯率。期盼透過此系統能夠提升投手資訊蒐集及呈現的效果，讓比賽更為精彩。並從分析投手的資訊來對棒球有更深入的了解，也對未來的台灣棒球賽事能有所貢獻。

這個系統先以中職為研究對象，尋找近期中職比賽完整的影片資料，以及跟台體大合作，獲取人工紀錄的資料，利用資料前處理，做出需要的影片內容及資料文件，透過不同的建模做出影像辨識的結果，結合 OCR 抓取有用的資訊，其中影像辨識抓出的球座標，透過非監督式學習來進行分群分析，抓出是軌跡上的球座標後，透過回歸分析來建立所需參數，將可能影響球種的特徵，用不同機器學習模型來做出球種辨識，透過模型評估來看結果，最後將數據紀錄下來，完成數據自動記錄的系統。

關鍵詞: 資料前處理、影像辨識、OCR、非監督式學習、分群分析、回歸分析、模型評估

目錄

圖目錄

第一章 緒論

1.1 研究背景

在棒球的比賽裡，為了要贏下一場比賽，會對敵方進行研究，不外乎從過去的對戰紙本記錄，影片分析，從中來模擬投手的球種來進行攻略，

圖 1 紙本比賽紀錄表示意圖

圖 2 CPBL 職棒比賽示意圖

指本紀錄上仍需要人工統計數據，以及較有可能遺失。從影片分析來看，其實需要的數據在影片上大致上都能掌握，但是需要紙本記錄的輔助，或是再透過人工的方式來統計數據。兩種方式都各有優缺點，但也是相輔相成。

1.2 研究動機

我們有組員曾在高中加入棒球隊，三年的參賽經驗中，能參考對手的資料不多，因為缺乏專業器具，所以資訊來源僅限於過去的紙本記錄，然而紙本記錄可能會因為人為疏失而有所缺漏，甚至是遺失，這可能造成對其他隊伍的數據無法有效地評估，進而只能在場上熟悉對手的球路後，才能有更好的攻擊策略，這會造成得分不彰。而上了大學，加入球隊後仍然遇到同樣的問題，因而有了這次想要製作專題的想法。

在國外有許多研究棒球的各種數據，大多都是以大聯盟為研究對象，但這些數據與台灣職業球員與大學球隊球員會有些差異，因此我們決定先研究國內賽事為主。然而資料最齊全，甚至能有影片來輔助的研究對象就是中華職棒，所以我們拿中華職棒的比賽影片為研究對象，希望在有限的資源下，透過影片來分析不同投手的特徵，探討投手從投球的過程中，自動產生的資料，像是球速、軌跡等等，從資料中讓我們更方便的研究出投手的特性，進而達到攻克投手的效果。

1.3 研究目的

數據自動產生及建檔，能夠取代傳統的紙本記錄，並降低了錯誤率及減少遺失的可能性。要了解投手可能投的球種，透過影像辨識及 OCR 將可能為影響辨識球種的資料抓取出來，在訓練完模型後，能夠對辨識球種有很大的幫助。

。

1.4 研究範圍

本研究以數據分析為主，一方面探討數據取得與處理的流程，另一方面則是研究不同的建模方式，讓造成的結果更佳的優化。

主要研究範疇：

- 資料處理
- 影像辨識
- OCR
- 分群分析
- 回歸分析
- 機器學習
- 模型評估

1.5 研究規劃

1.5.1 研究規劃時程規畫安排

圖 3 研究規劃時程規畫甘特圖

1.6 研究流程

1.6.1 資料前處理

我們選用的研究對象為 CPBL，很榮幸的能夠跟台體大合作，獲取職棒的整場比賽影片及資料檔，我們將獲取的影片及資料檔處理及標記標籤，讓後續能夠進行訓練。

1.6.2 影像辨識

在眾多影像辨識模型當中，我們嘗試了 Keras yoloV3 及近期最新發布的 Yolo V4，我們最終選擇以 Python 為基底的 Yolo V4 模型來使用，除了 Python 好上手以及在影像識別的領域資源較多之外，Yolo V4 相較 Yolo V3 更加穩定。

1.6.3 OCR

它是光學字元辨識，透過影像辨識抓出所需資訊的座標，將其切成照片，並轉成文字，而我們嘗試用原始 python 套件、google 提供的串接，以及自行訓練 OCR 模型，因為原始 python 套件辨識度不太好，而 google 的 OCR 雖然有準又快，但是過了一定的數量就會開始收費，因此最終我們使用自行訓練好的 OCR 來使用。

1.6.4 分群分析

在機器學習中，有許多演算法是能針對點進行分群的，因此透過分群分析，我們分類出每個打席中所有可能為投球軌跡的球。

1.6.5 回歸分析

透過分群後的點，我們試著用回歸演算法來進行回歸，當中嘗試了不同次方回歸來比較、最後我們選用二次曲線，並且我們提取回歸線的係數當特徵。

1.6.6 機器學習球種辨識

將前面所獲得可能為影響球種的特徵資訊，我們透過半監督式的機器學習來進行球種辨識，其中我們嘗試了不同的機器學習模型，最終我們選擇成效最好的 Xgboost 來進行球種辨識。

1.6.7 模型評估

將產生出來的辨識結果，我們透過許多可驗證的方式來評估模型的優劣，並且將錯誤值抓取出來，找到錯誤源頭。

1.7 章節結構

本研究之重點，在於如何建立與棒球數據自動分析的實務運作。首先，在第二章中，會依序回顧系統的相關技術之簡介與概念，做為構建後續更了解系統的開發及整體規劃之基礎，在第三章中，將會提及最終系統是如何架構。在第四章中將會針對開發完成的系統進行測試，並透過實驗出來的結果，訂定後續的改善目標，以及可優化的設計。最後，在第五章中對本次專題所開發的系統做結論，以及給與後續可能繼續研究的方向。

第二章 背景知識與相關技術介紹

2.1 球種

在棒球比賽中，投手只會投直球是很難生存的，因此從古至今研發出奇形怪樣的球種，而這次專題我們則是將球種分為三種簡單的分類，速球、變速球與變化球。

2.2 影像辨識 – Yolo V4

影像辨識是種物件偵測的技術，透過人工標籤的方式，讓模型能夠學習，並且能夠對不同影片中的相同物體進行辨認。而 yolo 系列是物件偵測的類神經網路演算法，它的優點就是 real time，其中 V4 更是最新又穩定的一個版本。

2.3 分群分析

將同屬性的資料，透過演算法把資料經不同的特徵及相似度較高的集群聚集成的群體，進而達到分群的效果。

2.4 回歸分析 – 二次回歸模型

主要是點轉換到線的過程，探討自變數與依變數之間的關聯性，透過模型建立，藉此預測出研究者有興趣的值，其中投球的軌跡與二次回歸最為相似。

2.5 OCR

是為光學字元辨識，是一種將圖片進行分析處理的技術，並取得文字及版面資訊的過程。

2.6 球種辨識 – 半監督式機器學習

在 python 的套件中，已有許多的機器學習模型，而這次系統選用的模型則為 Xgboost。半監督式學習的原理是標註部分答案，透過這些辨識的依據在對相關無答案的資料進行辨識。

第三章 系統分析

3.1 流程規劃

本次專題將會以下列流程進行

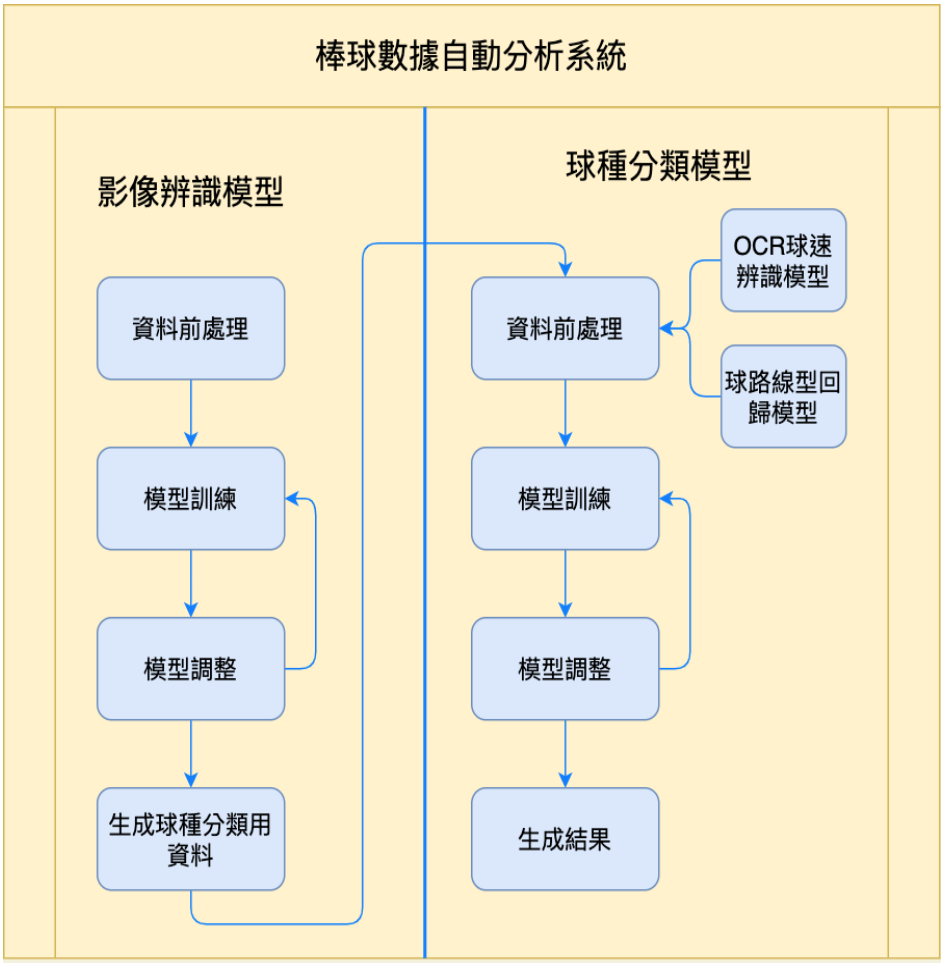

圖 4 系統流程圖

3.2 系統設計

3.2.1 影像辨識模型

1. 前處理：我們要尋找實驗的影片，並將其處理。
2. 模型訓練：我們決定實作不同的視覺辨識模型來做這次的實驗。
3. 模型調整：檢查是否有缺漏的數據或是解決問題。
4. 生成球種分類用資料：輸出結果

3.1.1 球種分類模型

1. 資料前處理

   甲、OCR 球速辨識模型：我們將使用 python 的內建套件來對圖片轉文字，這樣就能抓取圖片中的球速了。

   乙、球路線性回歸模型：我們將抓取的球座標進行線性回歸模型，這樣就能看出球路的大概軌跡。

2. 模型訓練：我們選用分類模型來做球種分類，會多嘗試不同的模型來觀察好壞。
3. 模型調整：檢查是否有缺漏的數據或是解決問題。
4. 生成結果：最後產生出球種分類的情形。

第四章 實證分析

4.1 實驗設計

4.1.1 影像辨識模型

1. 前處理：將影片下載後，使用模型進行每個投球的影片分割，再將每場比賽的紀錄檔跟我們處理過的每個投球的影片進行匹配，如果有一筆紀錄無法匹配到任何一顆球就留空白；再將每個投球的影片切割成圖片進行物件的標記，我們要標記的物件類別有：棒球、球棒、投手、捕手、裁判、主隊分數、客隊分數、壘包、局數、總球數、好壞球、出局數和球速，最後生成的訓練資料包含所有切割的圖片（.jpg 格式）跟標記檔案（.xml 格式），放置於兩個資料夾裡，圖片跟標記檔案名字要一樣。

2. 模型訓練：我們採用的模型是 darknet 架構下的 YOLOv4[1](https://github.com/AlexeyAB/darknet)，參考 GitHub 上的說明進行訓練，經過訓練產生權重檔案（.weights）用於預測。

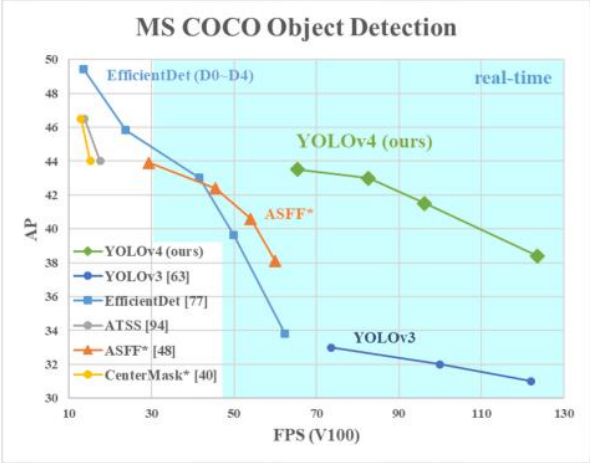

圖 5 模型比較(https://github.com/AlexeyAB/darknet)

3. 模型調整：我們嘗試過增加訓練的圖片數量來提高訓練的準確度；也有因應後續模型的需求補充標記新的類別
4. 生成球種分類用資料：經過影像辨識的模型，每個投球的影片會生成各個類別在影片中出現的幀（frame）的位址與在每一幀上的位址座標（如果沒有辨識出該類別就沒有位址座標），每個辨識出的類別還會有辨識的信心值，我們利用每個辨識出類別的門檻值（threshold）進行篩選，篩選出信心值較高的類別用於後續處理，這些資料將在球種分類模型中進行前處理

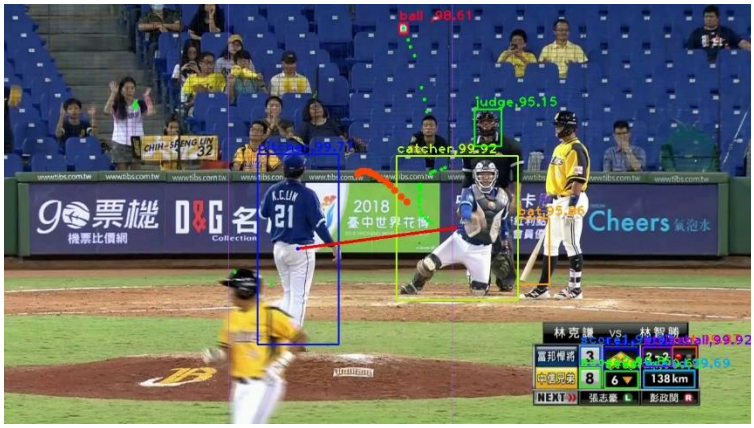

圖 6 影像辨識結果

4.1.2 球種分類模型

1. 資料前處理

甲、OCR 球速辨識模型：我們利用影像辨識模型得到的速度物件坐標進行圖片切割，可以得到只包含速度的圖片，接著，我們利用了

pytesseract[2](https://pypi.org/project/pytesseract/)套件加入我們從速度圖片中選取部分的訓練圖片建構成模型進行訓練，最後生成的模型用於進行球速辨識，辨識出的球速作為一個特徵

乙、球路線性回歸模型：同樣地，也利用影像辨識模型得到的棒球物件坐標進行處理，將所有點進行分群，得到最有可能的每個投球過程中所有球的坐標使用二次曲線進行回歸，得到的二次回歸線系數 a, b, c 作為這次投球的另一組特徵

2. 模型訓練：我們選用隨機森林模型（Random forest）進行球種預測，輸入模型的特徵包含球速辨識模型辨識出的球速、球路線性回歸模型的曲線參數 a, b, c 及投手捕手連線的長度跟斜率
3. 模型調整：由於每種球種的數量不一樣，我們根據模型的表現變化嘗試過使用升採樣（upsampling）將數量較少的球種數量提升到比較多；我們也進行過模型參數調整來提升模型的準確度
4. 生成結果：根據這個模型的預測結果可以得到最終每個投球影片預測的球種分類結果

REPORT:

	precision	recall	f1-score	support
變化球	0.79	0.79	0.79	39
變速球	0.70	0.44	0.54	16
速球	0.92	0.97	0.95	111
accuracy			0.88	166
macro avg	0.81	0.74	0.76	166
weighted avg	0.87	0.88	0.87	166

圖 7 評估結果

```
print("F1 score macro:", f1score_macro)
print("Confusion matrix:\n", confusion_matrix_plot(cm))
```

```
Accuracy: 0.8795180722891566
F1 score macro: 0.7602339181286549
```

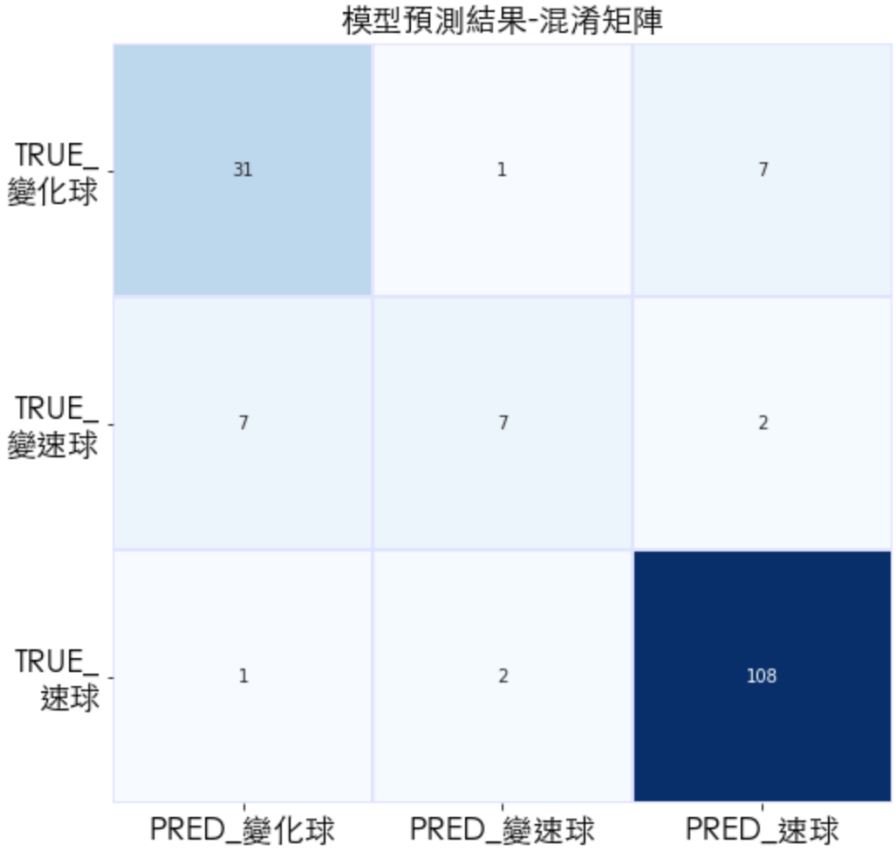

圖 8 混淆矩陣

第五章 結論與建議

5.1 結論

在視覺辨識模型上，因為在實驗階段，所以先從剪輯過後的影片來做實驗，而在模型的訓練上因為不確定有甚麼參數可以添加，所以只從我們覺得可能影響的因素來當作參數，而目前只有取回歸線的系數、球速、投手、日期、投捕間的斜率及投捕間的長度，在模型的選用我們選擇 yolo v4，這個模型實作上效果不錯，搭配訓練過後的 ocr，基本上在參數視覺辨識上不是太大的問題。在球種分類模型上，我們先將球種分好群以後，因為三種球種的數量以速球為最多，在抽樣上可能會有所不均，因此我們決定將球種作 up sampling，於是我們探討

了使用 smote 有無的好壞，我們發現資料量小時：使用起來的效果比較好；反之效果則普普，所以我們最後沒有使用，在資料處理上，最常聽見的處理莫過於用 normalization 跟 standardization，然而出來的結果卻是出乎我們意料之外，反而用原始資料還好一些，因此就沒使用以上兩者技術。而在分類的模型選用上，我們嘗試了隨機森林樹、XGBoost 以及 SVG……等等，而最後我們發現隨機森林樹的結果不論在正確率、f1 score 以及混淆矩陣都好一些。

5.2 研究心得

從這次的系統發想，我們認為整份專案只要投入足夠的時間，提早完成應該不會有太大的問題，畢竟組上可是有接觸過這類相關比賽的組員。當然一開始在資料處裡的時候，真的是實扎實打，花多少時間做多少事，像是在比對影片跟記錄檔，還有當時聽到要標記 10 個標籤，有幾千張要標記完成，真的是嚇到了，但我們還是硬著頭皮去做，結果出乎我們意料之外的，我們很快就標記完了，這也很順利的進入到下一個階段。

接下來最複雜的就是在建模上面，首先遇到很棘手的問題就是環境，因為組內所使用的環境不太一樣，所以一開始採到很多洞，後來跟老師申請了 ubuntu 的主機以後，我們就解決大部分環境的問題了，接著是在選用模型的使用問題，因為一開始我們想說要做到即時辨識，那最耳熟能詳的就是 Yolo 系列的模型了，所以我們找了幾個 YoloV3 來測試，結果我們研究完，也使用的很上手時，YoloV4 就被發表了，那時候抱著想再多研究的心情，我們決定重頭再來，開始轉為使用 YoloV4，當然有了前面的底子，這部分就沒有卡太久了，接下來就是要塞選出需要用到的特徵，這也是我們組最開心的時候，看著 CPBL 在研究到底甚麼可以被當作特徵，這時候就像是腦洞大開，一直在挖掘新事物。

最後當然是辨識球種，這時測試完模型過後，雖然挑選了評比最好的模型，但是，它的評比看起來還是不太行，所以我們就從頭開始追根，找尋著到底是甚麼影響著我們的結果，一直反覆試練後，我們終於找到合適的方式，完成的喜悅讓我們很有成就。

5.3 後續方向

未來我們希望能夠將球路篩選機制變得更加的完善，這除了需要提升模型的精確度之外，還有究竟要加哪些參數未來能夠探討的更加深入。現在還只是透過結束後的影片來實作這系統，等到 2D 的影像做到預期的結果以後，希望能夠挑戰 3D 影像，並且達成 real time 的實作。

計畫成果自評

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值(簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性)、是否適合在學術期刊發表或申請專利、主要發現(簡要敘述成果是否具有政策應用參考價值及具影響公共利益之重大發現)或其他有關價值等，做一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況做一綜合評估
   - ■ 達成目標
   - □ 未達成目標(請說明，以 100 字為限)
     - □ 實驗失敗
     - □ 因故實驗中斷
     - □ 其他原因

   說明：

2. 研究成過在學術期刊發表或申請專利等情況(請於其他欄註明專利及技轉之證號、合約、申請及洽談等詳細資訊)

   論文：□已發表 □未發表之文稿 □撰寫中 ■無

   專利：□已獲得 □申請中 ■無

   技轉：□已技轉 □洽談中
   ■無

   其他：(以 200 字為限)

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值(簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性，以 500 字為限)。

   此次的研究將影響未來的運動產業，不論是現階段所研究的棒球項目上，在其他領域仍有機會應用此套系統，從影像辨識到文字自動記錄，這都是節省人力資源的,相信在未來會有更充分的研究成果,因此在這份研究上,可以再向更高更遠的地方邁進。

4. 主要發現

   本研究具有政策應用參考價值：■否 □是，建議提供機關_____

   (勾選「是」者，請列舉建議可提供施政參考之業務主管機關)

   本研究具影響公共利益之重大發現：■否 □是

   說明：(以 150 字為限)