FCU

ePaper

逢甲大學學生報告 ePaper

基於視覺辨識之機器人交互界面

Vision-Based Human-Robot Interaction Interface

作者：周竑宇

系級：資訊二甲

學號：D1209305

開課老師：葉春秀 老師

課程名稱：多媒體系統

開課系所：資訊工程學系

開課學年：113 學年度 第 2 學期

中文摘要

本專案致力於開發一套「基於視覺辨識之機器人交互介面」，旨在提升人機協作的直觀性與實用性 。系統核心目標是建構一個能夠即時感知、分析使用者行為，並觸發人形機器人交互功能的完整運行模式 。

本系統深度整合多項先進技術，實現精準的使用者行為捕捉與情緒感知：

多目標人像偵測與追蹤： 採用 YOLOv8-pose 進行人像偵測，並結合 DeepSORT 演算法實現多使用者穩定追蹤與身份識別，有效降低誤判並提升追蹤魯棒性 。

高精度手勢識別： 基於 MediaPipe Hand 提供的關鍵點座標序列，開發客製化揮手檢測算法 。該算法透過分析手部關鍵點軌跡的週期性運動，並設定像素位移閾值（如 50-100 像素）和 1-2 秒滑動視窗，結合多層過濾機制，確保高精度手勢識別 。

臉部情緒辨識： 針對現有模型（如 DeepFace）的限制 ，本專案選擇運用 MediaPipe Face Mesh 技術 ，並透過建立客製化臉部資料集顯著提升情緒辨識準確率，包含快樂、中性、驚訝等多種情緒，並經精確度、召回率和 F1-score 等指標嚴格評估 。

高效能數據流與機器人協同： 系統將影像處理後的 ROI 數據精確裁切並轉化為文字訊息傳輸，大幅降低延遲 。這些優化後的數據驅動機器人，使其能實現「骨架跟隨使用者移動」等交互功能，確保機器人動作與使用者行為的自然同步 。

關鍵字：

機器人交互介面 、視覺辨識 、人機交互 、YOLOv8-pose 、DeepSORT 、MediaPipe 、表情辨識 、物件追蹤 、骨架辨識 、關鍵點軌跡 、即時互動 。

Abstract

This project is dedicated to developing an "Vision-Based Human-Robot Interaction Interface," aiming to enhance the intuitiveness and practicality of human-robot collaboration. The system's core objective is to construct a fully operational mode capable of real-time perception and analysis of user behavior, thereby triggering interactive functions in humanoid robots.

This system deeply integrates multiple advanced technologies to achieve precise capture of user behavior and emotional perception:

- **Multi-Target Human Detection and Tracking:**The system employs the**YOLOv8-pose**model for high-efficiency human detection, and its detected bounding boxes are fed into the**DeepSORT**tracking algorithm. This combination ensures stable tracking and identity recognition of multiple users (track IDs) across continuous video frames, effectively reducing false positives and mitigating the impact of temporary occlusions, thereby enhancing tracking robustness.
- **High-Precision Gesture Recognition:**For interactive gestures such as waving, this system develops a customized waving detection algorithm based on the keypoint coordinate sequences provided by**MediaPipe Hand**. This algorithm analyzes the periodic motion of hand keypoint trajectories in the horizontal direction and sets pixel displacement thresholds (e.g., 50-100 pixels) and a 1-2 second sliding window. Combined with multi-layer filtering mechanisms, it ensures high-precision gesture recognition.
- **Facial Emotion Recognition:**Addressing the limitations of existing models (e.g., DeepFace) in specific application scenarios, this project opts to utilize**MediaPipe Face Mesh**technology. To this end, we specifically built and trained a customized facial dataset, significantly improving the accuracy of emotion recognition (including happy, neutral, surprised, etc.) in complex environments. This was rigorously evaluated using metrics such as precision, recall, and F1-score.
- **High-Efficiency Data Stream and Robot Collaboration:**The system design considers the demands of real-time interaction. After image processing, the analyzed ROI (Region of Interest) data is precisely cropped and converted into text messages for transmission, significantly reducing latency. This optimized data then drives the robot, enabling interactive functions such as "skeleton following user movement," ensuring natural synchronization between robot actions and user behavior.

Keywords:

Human-Robot Interaction Interface, Vision-Based Recognition, Human-Robot Interaction, YOLOv8-pose, DeepSORT, MediaPipe, Gesture Recognition, Emotion Recognition, Object Tracking, Skeleton Tracking, Keypoint Trajectory, Real-time Interaction.

目　次

一、專題動機

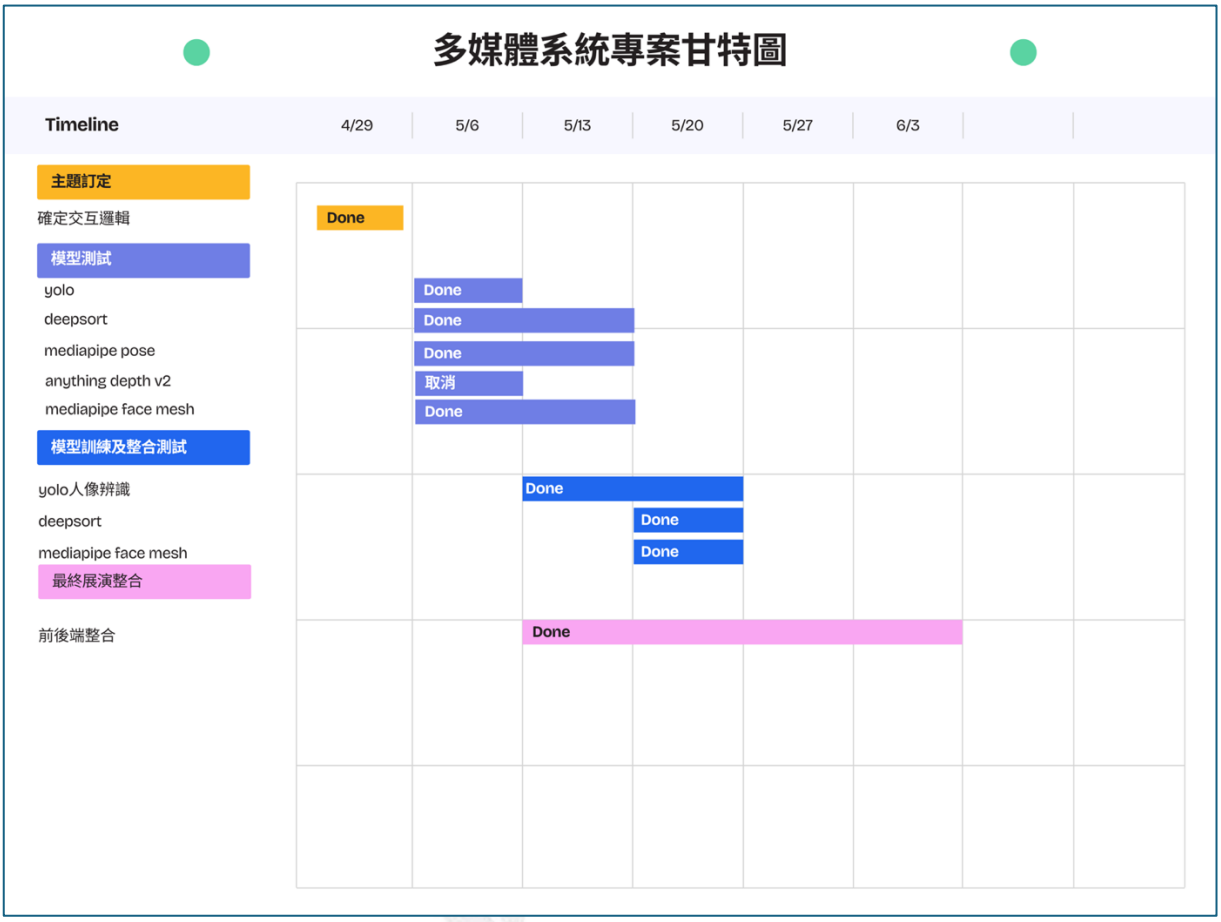
本專案的動機，是因為平時有在做機器人的專案，因此想做一個能夠觸發人形機器人交互功能的系統，需要是可以完整可以運行的模式。

二、專題目的（Project Objectives）

隨著人工智慧與機器人技術的快速發展，人機交互系統在各類應用場景中扮演著日益重要的角色。本專案的目標，正是針對此趨勢，打造一套具備以下特點的機器人交互系統：

1. **提升使用者體驗：** 透過直觀的操作介面與即時反饋機制，降低使用者與機器人互動的門檻，使其無需繁複設定即可輕鬆展開協作。
2. **促進創新與效率：** 透過視覺辨識技術實現對使用者行為與情緒的精準感知，使機器人能更智慧地響應，進而提升協作效率與應用創新潛力。
3. **確保系統完整性與易於部署：** 本系統特別強調其完整可運行性與易於部署的特性，旨在成為未來智能機器人交互應用的穩固基礎平台。
4. **實現多模態感知與交互：** 整合多項先進技術（如 YOLOv8-pose、DeepSORT、MediaPipe Hand/Face Mesh），實現對人像偵測、手勢識別、臉部情緒辨識等多模態資訊的即時處理與分析，並將這些感知結果有效地轉化為機器人的交互指令。

三、甘特圖


多媒體系統專案甘特圖
Timeline
4/29
5/6
5/13
5/20
5/27
6/3
主題訂定
確定交互邏輯
模型測試
yolo
deepsort
mediapipe pose
anything depth v2
mediapipe face mesh
模型訓練及整合測試
yolo人像辨識
deepsort
mediapipe face mesh
最終展演整合
前後端整合
Done
Done
Done
Done
取消
Done
Done
Done
Done
Done

四、使用之套件與程式

本專案為實現「基於視覺辨識之機器人交互介面」的核心功能，深度整合並運用了多項先進的機器學習模型與視覺處理套件。以下將詳細解釋各主要技術及其在系統中的應用：

1.YOLOv8-pose (You Only Look Once v8 - Pose Estimation)

用途： 在本專案中，YOLOv8-pose 主要用於高效能地偵測影像中的「人像」並獲取其邊界框，作為多目標人像偵測的第一階段，提供精確的人體位置資訊。

2.DeepSORT (Simple Online and Realtime Tracking with a Deep Association Metric)

用途： DeepSORT 與 YOLOv8-pose 結合，用於實現對多個使用者在複雜場景下的穩定、持續追蹤與身份識別，並有效降低誤判。

3.MediaPipe Hand

用途： MediaPipe Hand 提供即時、高精度手部 21 個 3D 關鍵點資訊，為本專案的「高精度手勢識別」提供底層的關鍵點座標序列，用以識別如揮手等特定交互動作。

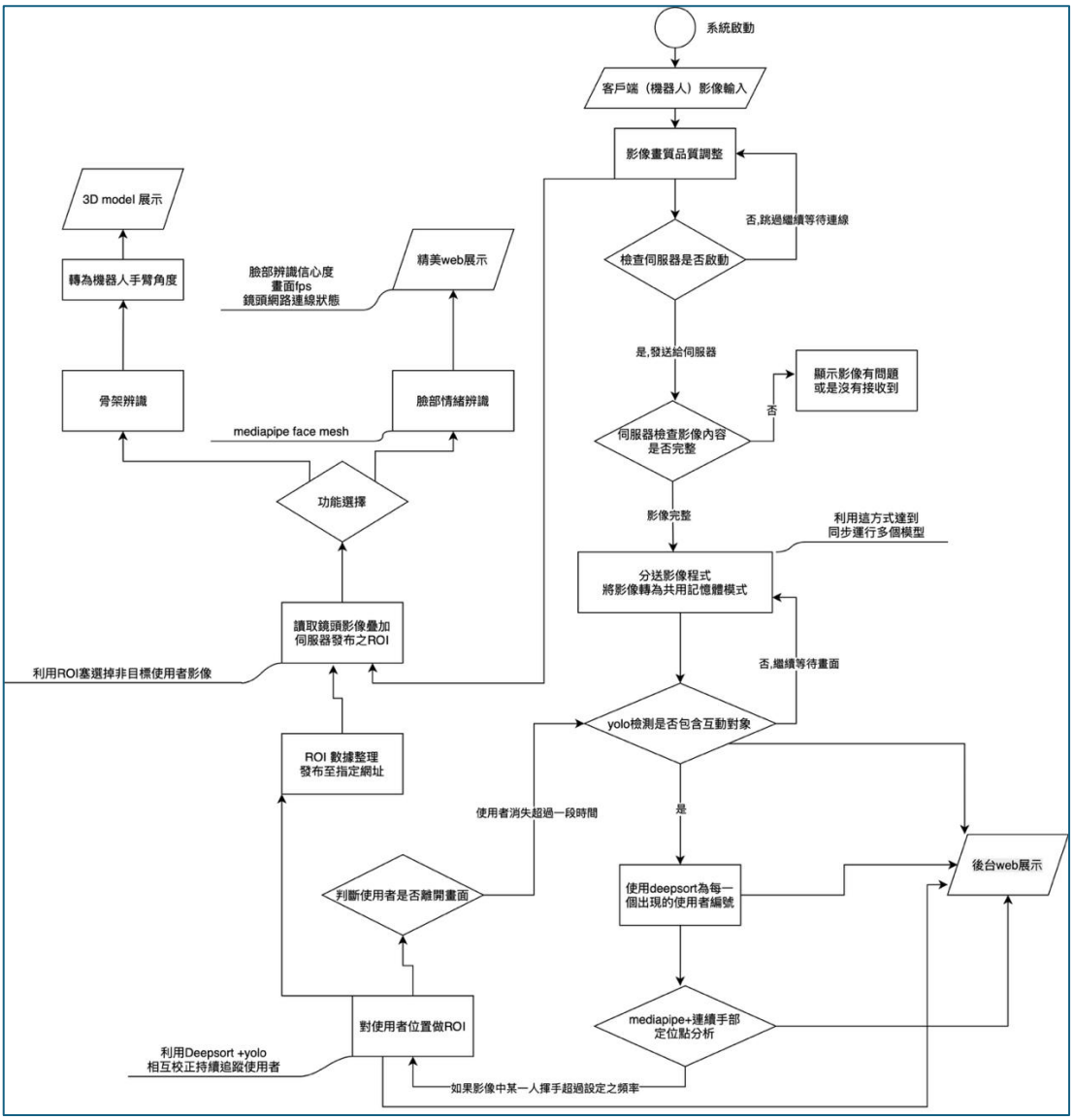
4.MediaPipe Face Mesh

用途： MediaPipe Face Mesh 能夠即時偵測人臉並估計出 468 個 3D 臉部關鍵點，用於本專案的「臉部情緒辨識」，透過分析這些關鍵點判斷使用者的情緒狀態。

5.Python（程式語言）

用途： Python 作為整個系統的開發基礎語言，用於整合各機器學習模組、處理數據流、實現演算法邏輯及構建後台應用。

五、功能架構圖


系統啟動
客戶端（機器人）影像輸入
影像畫質品質調整
檢查伺服器是否啟動
否,跳過繼續等待連線
是,發送給伺服器
伺服器檢查影像內容
是否完整
否
顯示影像有問題
或是沒有接收到
影像完整
利用這方式達到
同步運行多個模型
分送影像程式
將影像轉為共用記憶體模式
否,繼續等待畫面
yolo檢測是否包含互動對象
是
使用deepsort為每一
個出現的使用者編號
mediapipe+連續手部
定位點分析
如果影像中某一人揮手超過設定之頻率
後台web展示
使用者消失超過一段時間
判斷使用者是否離開畫面
對使用者位置做ROI
利用Deepsort +yolo
相互校正持續追蹤使用者
ROI 數據整理
發布至指定網址
讀取鏡頭影像疊加
伺服器發布之ROI
利用ROI塞選掉非目標使用者影像
功能選擇
骨架辨識
臉部情緒辨識
mediapipe face mesh
轉為機器人手臂角度
3D model 展示
臉部辨識信心度
畫面fps
鏡頭網路連線狀態
精美web展示

六、影像辨識模型的運用與嘗試

1. 揮手辨識

- 基於關鍵點軌跡的運動分析

客製化揮手檢測算法基於 MediaPipe Hand 提供的關鍵點座標序列進行運動分析。算法追蹤特定手部關鍵點（主要是手腕和指尖）在時間序列中的位置變化。揮手動作的特徵是手部在水平方向的週期性運動，伴隨適度的垂直位移。

運動軌跡分析使用滑動視窗方法，在固定時間間隔內收集關鍵點位置資料。算法計算位移向量的大小和方向，識別符合揮手模式的運動序列。為了提高檢測的魯棒性，系統設定了最小位移閾值和運動頻率範圍。

- 像素位移閾值與參數調整

揮手檢測的核心參數是像素位移閾值，該閾值決定了識別為有效揮手動作的最小運動幅度。系統根據攝影機解析度和預期交互距離進行閾值標定。典型設定下，手部關鍵點需要在水平方向產生至少 50-100 像素的位移才被識別為揮手動作。

時間視窗參數控制檢測的靈敏度和延遲。較短的視窗提高了響應速度但可能增加誤檢，較長的視窗提供更穩定的檢測但會增加系統延遲。專案中採用 1-2 秒的滑動視窗，在檢測準確性和響應速

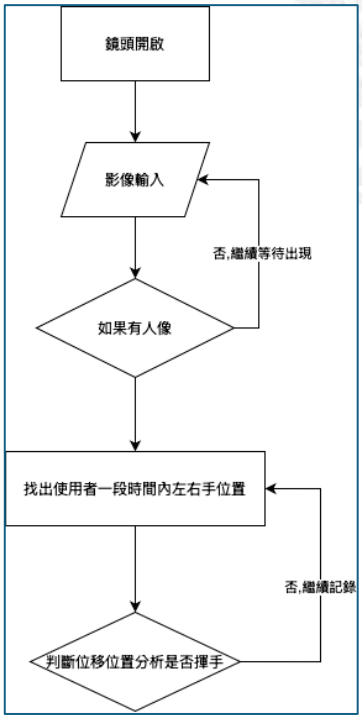
度之間達到平衡。

- 誤檢過濾與穩定性增強

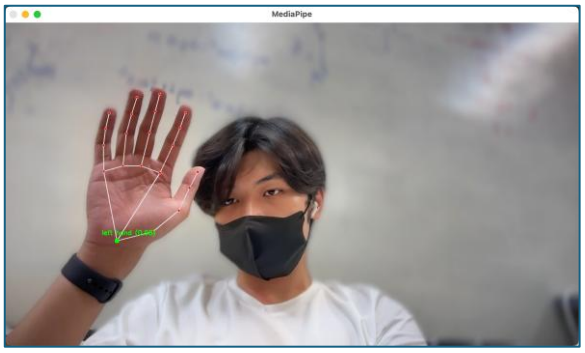
為了減少環境干擾和無意手部動作造成的誤檢，算法實現了多層過濾機制。首先，系統驗證檢測到的運動是否來自於有效的手部區域。其次，算法分析運動的連續性和週期性，排除隨機或非結構化的動作。

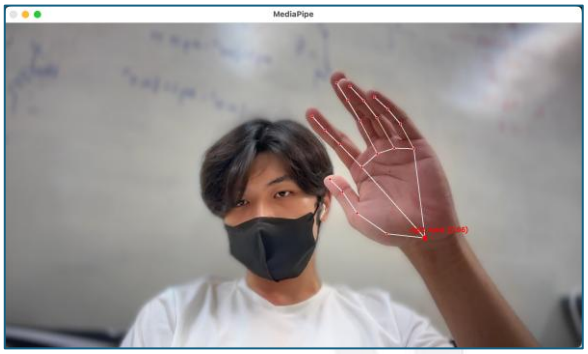
穩定性增強包括信心度評分和連續檢測確認。每個揮手候選動作都會獲得一個信心度分數，基於運動幅度、週期性和持續時間。只有連續多幀達到信心度閾值的動作才會被最終確認為有效的揮手指令。

- 辨識運行流程圖：

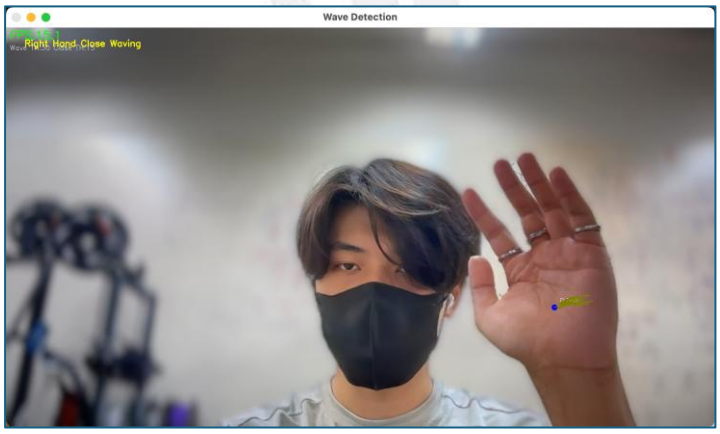


- 抓出手的位置

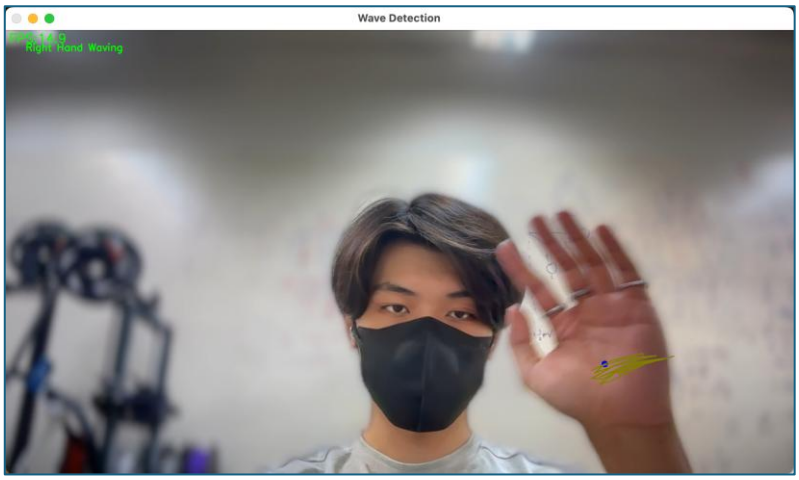




- 追蹤使用者左右手軌跡，手部位移像素點未能達到揮手標準



- 當位移達到目標像素點，即可輸出首部正在揮動。

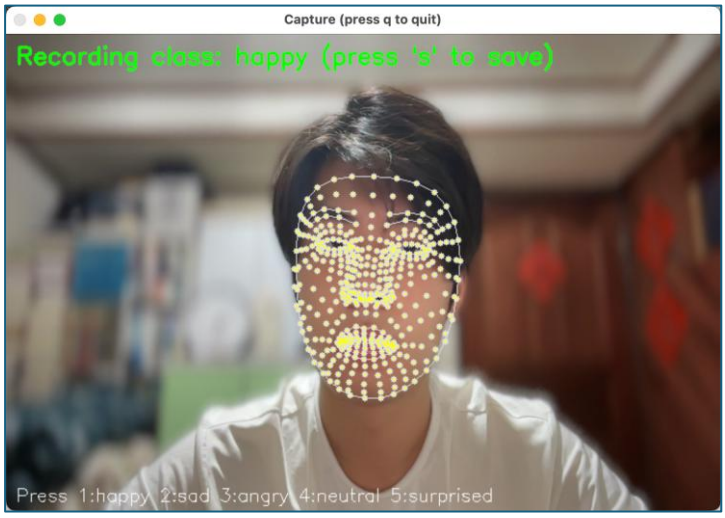


2. 表情辨識

專案初期，我評估了如 DeepFace 等現有模型，但測試後發現其辨識準確率未達預期標準。為解決此問題，我們最終選擇採用 MediaPipe Face Mesh 技術，並為其建立了一個客製化的臉部資料集。此方法在我們的特定應用中，成功實現了更為精準的表情辨識效果。

- 使用折線圖展示五在表情沒有變化下使用 DeepFace 辨識的結果，他的輸出極為不穩定

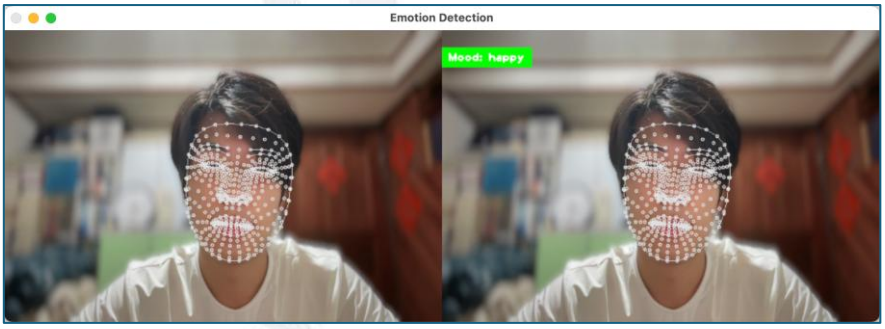
- 展示的是我修改後的臉部表情數據採集程式，能夠快速設定不同的表情類別，並將數據新增至 CSV 檔案，而不會覆蓋原有資料。

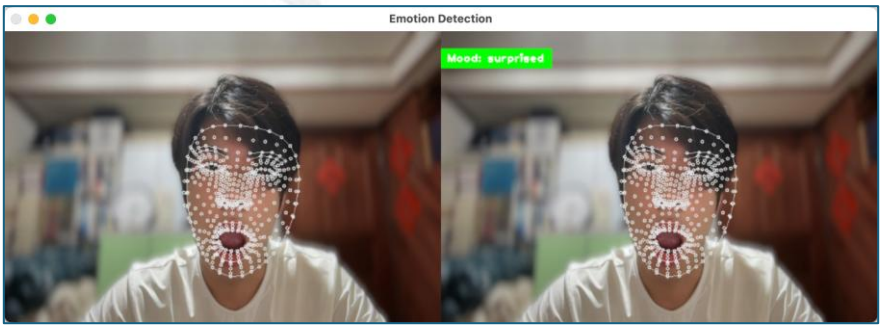


- 展示了利用採集完成的 CSV 數據進行模型訓練的過程。

```
[8 rows x 936 columns]
       x1   y1   x2   y2   x3   y3   x4   y4   x5   y5   x6  ...  y463  x464  y464  x465  y465  x466  y466  x467  y467  x468  y468
0      381  241  383  225  382  229  379  208  384  220  384  ...   228   398   194   395   196   393   197   422   194   425   192
1      383  241  386  226  384  230  381  208  386  221  386  ...   228   399   194   396   196   395   197   423   194   425   192
2      384  241  387  226  385  230  383  208  387  221  388  ...   228   400   194   397   196   395   197   423   194   426   192
3      385  241  388  226  386  230  384  208  389  221  389  ...   228   401   194   398   195   396   197   424   194   427   192
4      386  241  389  226  387  230  384  208  389  221  389  ...   228   401   194   398   195   397   196   424   193   427   192
...    ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...   ...   ...   ...   ...   ...   ...   ...   ...   ...   ...   ...
16594  415  269  414  236  412  248  402  215  413  228  411  ...   243   417   195   414   197   413   199   444   188   448   182
16595  415  269  414  237  412  249  403  215  413  229  411  ...   244   417   196   415   198   414   199   445   189   449   183
16596  416  270  415  238  412  250  403  216  414  230  412  ...   245   417   196   415   199   414   200   445   190   448   184
16597  415  270  415  238  412  249  403  216  414  230  412  ...   245   417   196   415   199   414   200   445   191   448   185
16598  416  270  416  239  413  250  404  217  415  231  413  ...   245   418   197   416   200   415   201   446   192   450   186

[16599 rows x 936 columns]
0           happy
1           happy
2           happy
3           happy
4           happy
          ...
16594    surprised
16595    surprised
16596    surprised
16597    surprised
16598    surprised
Name: Class, Length: 16599, dtype: object
/opt/anaconda3/envs/MediaPipe/lib/python3.10/site-packages/sklearn/linear_model/_logistic.py:470: ConvergenceWarning: lbfgs failed
to converge after 100 iteration(s) (status=1):
STOP: TOTAL NO. OF ITERATIONS REACHED LIMIT

Increase the number of iterations to improve the convergence (max_iter=100).
You might also want to scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
['neutral' 'surprised' 'neutral' ... 'happy' 'neutral' 'neutral']
Model Report:              precision    recall  f1-score   support

       happy       0.88      0.88      0.88       682
     neutral       0.94      0.96      0.95      2059
   surprised       0.98      0.92      0.95       579

    accuracy                           0.94      3320
   macro avg       0.93      0.92      0.93      3320
weighted avg       0.94      0.94      0.94      3320
```

- 模型實際測試



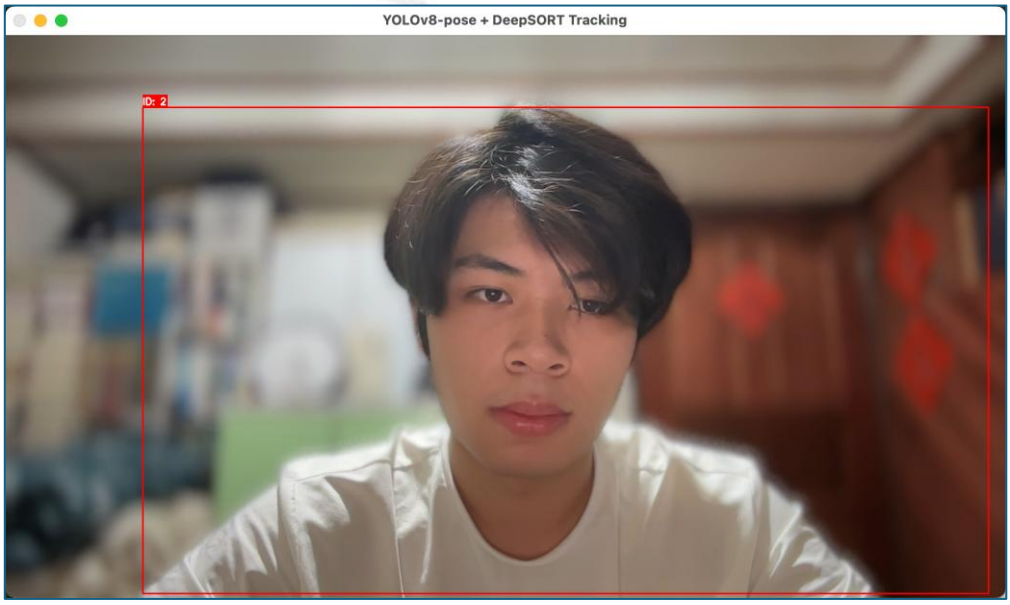


3. 物件追蹤

在專案中，我選用 YOLOv8-pose 進行物件追蹤。經過多種模型的測試與比較，確認 YOLOv8-pose 在人像辨識方面表現最為準確。由於專案需求僅需標註出人像的位置，並不需要輸出骨架節點數據，因此我們僅使用 YOLOv8-pose 所偵測到的人像標註結果作為最終輸出。

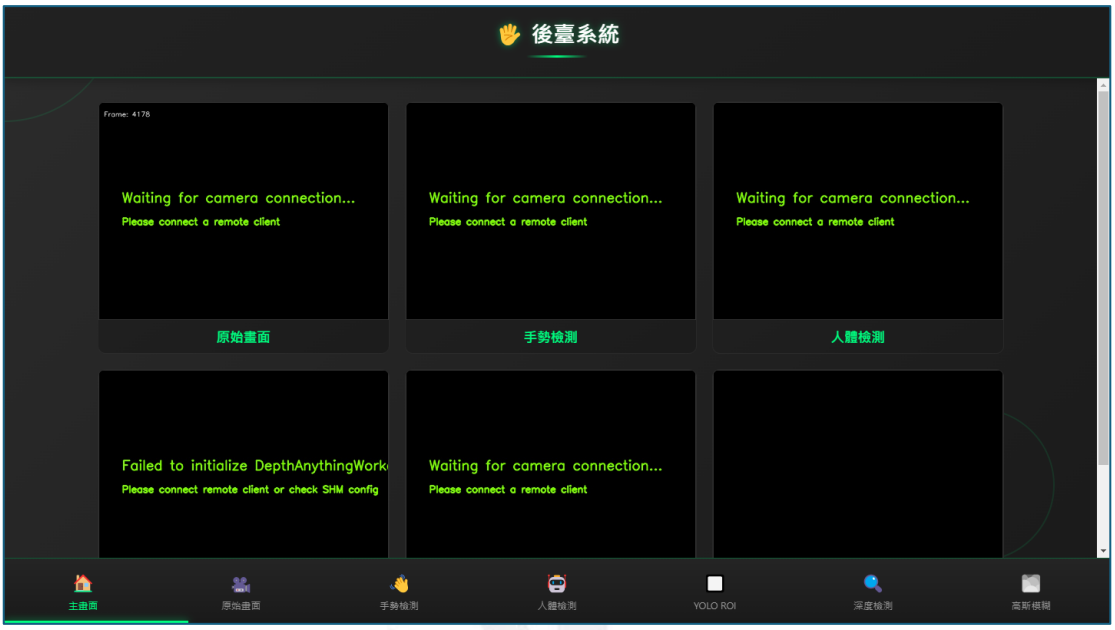
為了進一步提升追蹤效果，我將 YOLOv8-pose 所偵測到的人像位置（bounding boxes）輸入 DeepSORT 追蹤演算法，為每個偵測到的人像分配獨特標籤（track ID），並在連續畫面中穩定追蹤同一人物。透過 DeepSORT 的數據校正機制，能夠有效降低誤判與短暫遮蔽的影響，進而提升辨識與追蹤的準確性

- 展示出模型對人像的檢測還有編號

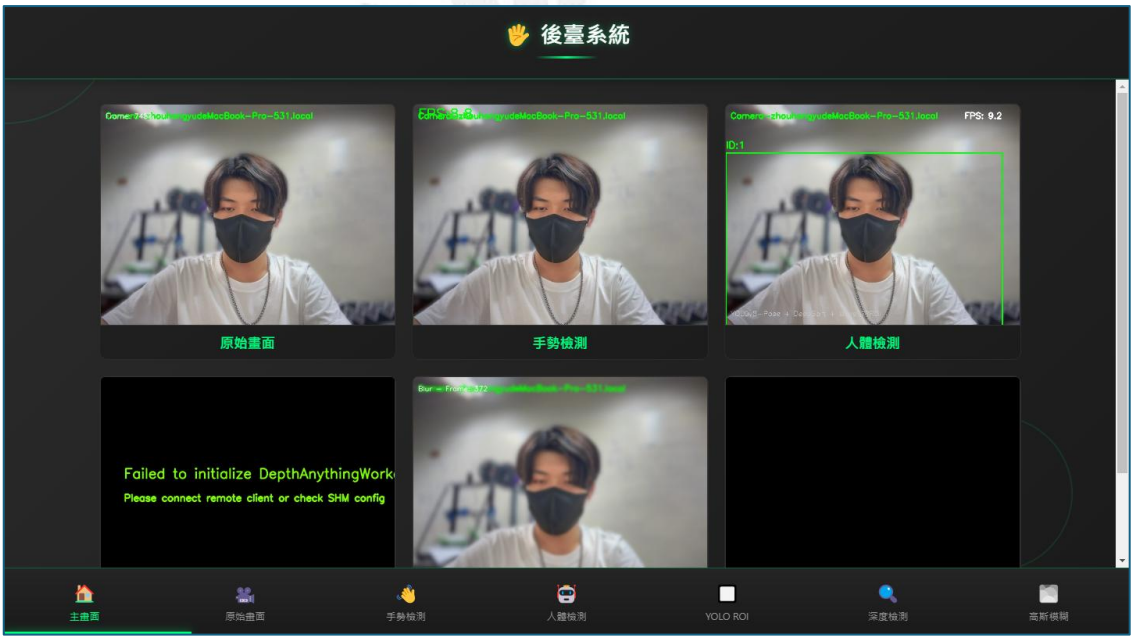


七、執行畫面

1. 後台無畫面輸入



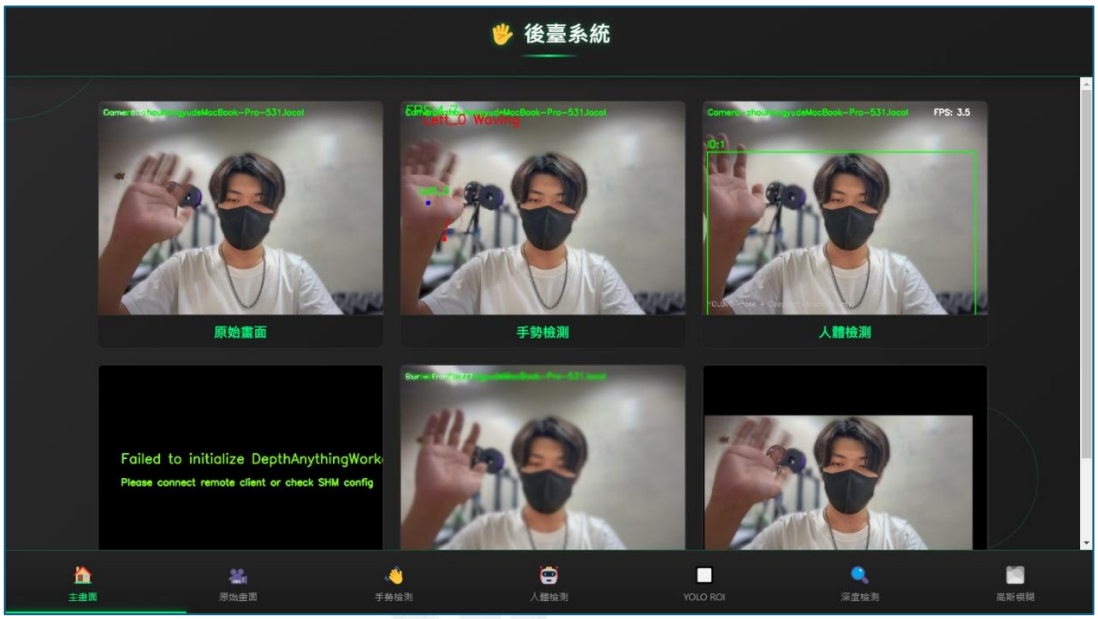
2. 人像 yolo 抓取 deepsort 追蹤編號



3. 手部骨架辨識揮手識別對操作者做追蹤 ROI

利用位的分析判斷揮手 不需要達成特定的姿勢就可以辨識出來

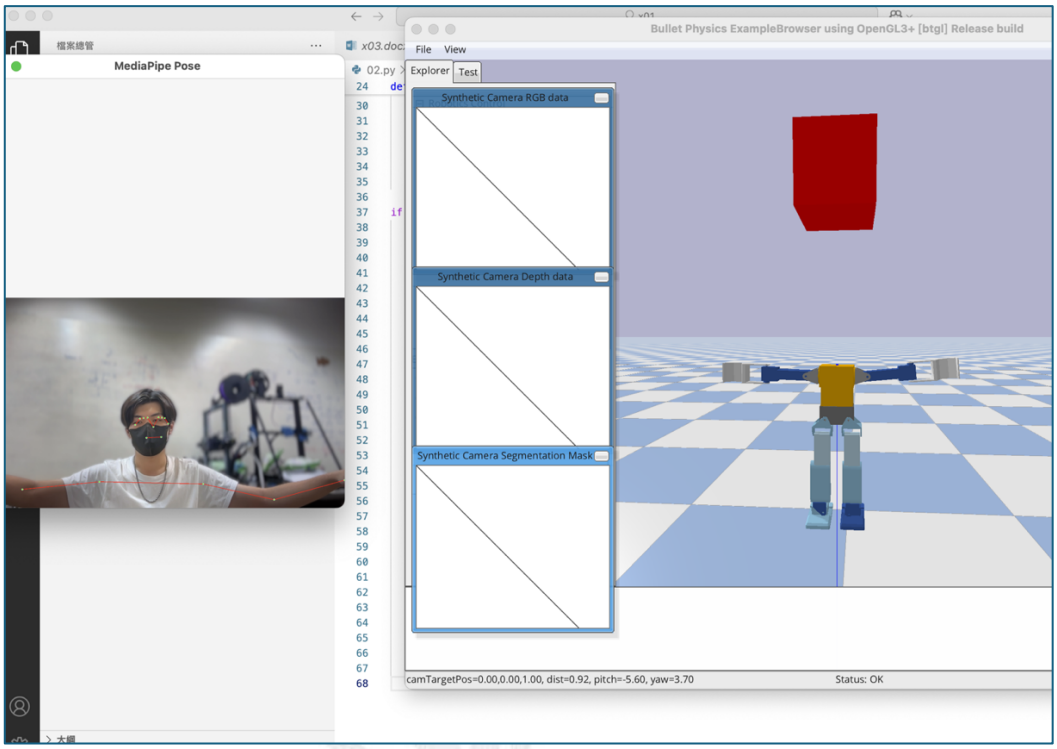
利用 yolo 以及 deepsoort 相互校正 每一個畫面都取得正確的 ROI 數值

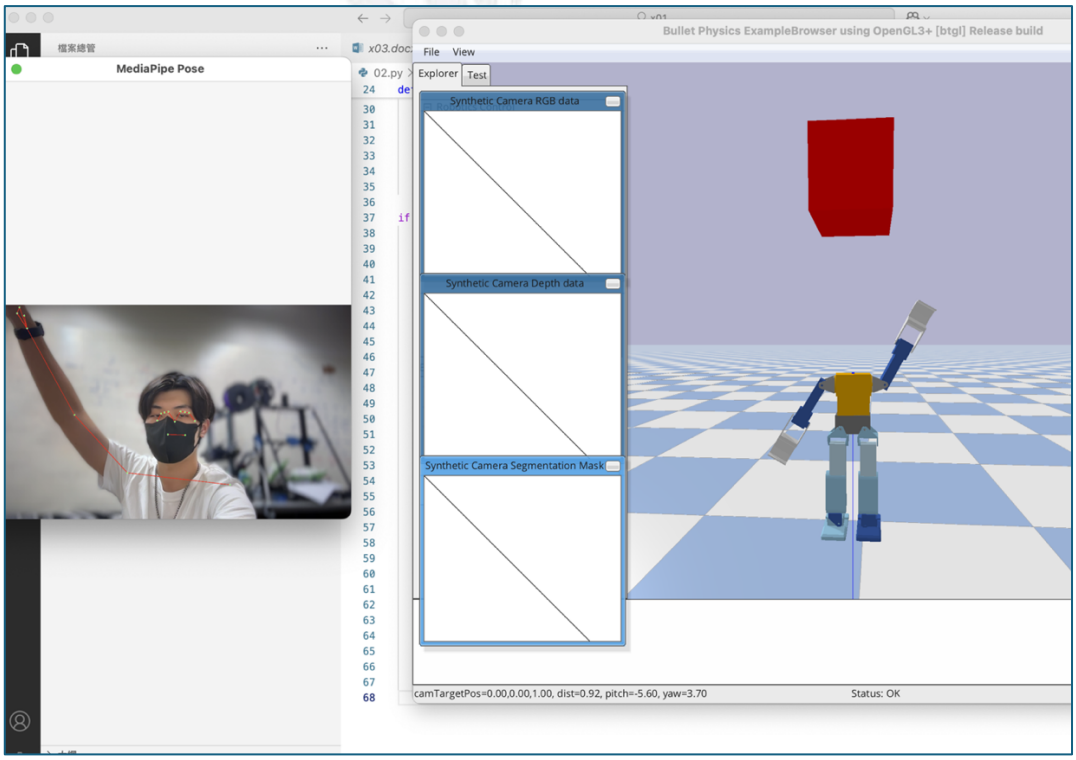


4. 客戶端接收 ROI 數據(文字訊息)裁切畫面，轉為文字傳輸降低延遲

影像處理結束後分析使用者表情，可搭配未來應用

5. 另一個互動功能 機器人骨架跟隨使用者移動，一樣是使用經過 ROI 處理之影像





八、未來展望

本專案所開發的「基於視覺辨識之機器人交互介面」已完成基礎的交互功能，未來仍有廣闊的發展潛力。我希望以下幾個方向以進一步提升系統的性能與應用範圍：

1. **擴展交互模式與複雜手勢識別：** 目前系統已能識別揮手等基本手勢，未來可研究導入更複雜、多樣化的手勢語義識別，例如手語、精細操作手勢等，以豐富人機交互的表達能力。
2. **多模態情感理解與響應：** 除了臉部情緒辨識，未來可整合語音情感分析、肢體語言解讀等模態，實現更全面、更細緻的使用者情感理解，使機器人能做出更具同理心和適應性的響應。
3. **跨平台與邊緣部署優化：** 探索將現有模型進一步優化，使其能在更多樣化的硬體平台（如嵌入式系統、低功耗設備）上高效運行，實現真正的邊緣計算，降低部署成本並提升即時性。
4. **個性化與自適應交互：** 引入機器學習的自適應能力，讓系統能根據不同使用者的習慣、偏好和學習曲線，動態調整交互模式和響應策略，提供更加個性化的人機協作體驗。
5. **結合語義理解與任務導向對話：** 將視覺感知與自然語言處理技術結合，使機器人不僅能理解使用者的動作和情緒，還能理解其語義意圖，進而執行更複雜、任務導向的對話與協作。
6. **擴展至多機器人協作場景：** 探索將單一機器人交互介面擴展至多個機器人之間的協作場景，實現更複雜的集體任務執行和人機群體協作。

九、結論

本專案成功開發了一套「基於視覺辨識之機器人交互介面」，實現了對使用者行為的即時感知與分析，並能有效地驅動人形機器人進行交互。透過整合 YOLOv8-pose 進行人像偵測與 DeepSORT 進行穩定追蹤，以及利用 MediaPipe Hand 和 MediaPipe Face Mesh 實現高精度手勢識別與臉部情緒辨識，本系統展現了其在複雜環境下捕捉細微人機互動資訊的強大能力。所有模組的協同運作，確保了數據流的高效處理與機器人動作的自然同步，大幅提升了人機協作的直觀性與實用性。本專案不僅驗證了多模態視覺辨識技術在人機交互領域的巨大潛力，也為未來智能機器人應用奠定了堅實的技術基礎，開啟了更智慧、更人性化互動模式的可能性。

十、參考資料

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, ChuoLing Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. “MediaPipe: A Framework for Building Perception Pipelines” arXiv preprint arXiv:1906.08172, 2019.

Jocher, G., Chaurasia, A., & Qiu, J. (2023). *YOLO by Ultralytics*. GitHub repository. Available at:https://github.com/ultralytics/ultralytics(Note: Specific YOLOv8-pose documentation might be within this repository or associated academic papers).

Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In*2017 IEEE International Conference on Image Processing (ICIP)*(pp. 3645-3649). IEEE.

Bazarevsky, V.; Grishchenko, I. On-Device, Real-Time Body Pose Tracking with MediaPipe BlazePose, Google Research. Available online: https://ai.googleblog.com/2020/08/on-device-real-time-body-pose-tracking.html (accessed on 10 August 2021). (Note: This reference is for MediaPipe BlazePose, but MediaPipe Hand/Face Mesh are part of the broader MediaPipe framework and share similar underlying principles and development context within Google Research.)