多重智慧體推論預測之犯罪偵防通報警示系統

A Multi-Agent Reasoning and Prediction System for Crime Detection and Alert Notification

指導教授：李國川

學生：李兆晞、徐茂程、郭獻壬、黃彥承、吳奇翰

國立聯合大學 資訊工程學系

苗栗市南勢里聯大 2 號

gclee@nuu.edu.tw

摘要

本文實作結合 RAG 與多模態 LLM，整合 YOLO 影像處理與多人即時通報，萃取多來源犯罪特徵，透過 Line 群播發布緊急訊息，並結合 GPT-SoVITS 實現客製化語音生成、播報。本系統導入思考鏈（CoT）與多重虛擬角色智慧體（Multi-Agents）技術進行犯罪預測推論，並支援自然語言查詢。有效降低人工調閱成本，協助預防無差別傷害事件，提升案件應變效率。

關鍵詞：檢索增強生成、大型語言模型、思考鏈推論、犯罪預測分析、聲音克隆

Abstract

This paper implements a system combining RAG and multimodal LLM, integrating YOLO image processing and multi-user real-time reporting to extract multi-source crime features and broadcast emergency messages via LINE, incorporating GPT-SoVITS for customized voice synthesis and audio broadcasting . The system introduces Chain-of-Thought (CoT) and Multi-Agent technologies for crime prediction inference and supports natural language queries. This effectively reduces the cost of manual video retrieval, helps prevent indiscriminate harm incidents, and enhances case response efficiency.

Keywords: Retrieval-Augmented Generation, Large Language Model, Chain-of-Thought Reasoning, Crime Prediction Analysis, Voice Cloning

一、研究背景與動機

近年來，即時影像監視器已廣泛建置於大眾運輸、商場與校園等公共場所，成為維護治安的重要防線。然而，面對日益增加的捷運隨機傷人、校園或商場持刀等無差別攻擊事件，現有安防機制仍顯不足。傳統監控系統多屬「被動影像」，高度仰賴保全人員 24 小時肉眼盯視；人工監看不僅容易因視覺疲勞產生疏漏，更難以在高危事件發生之初做出即時預警。事後調查亦需耗費大量時間人工回放歷史影像，缺乏多維度輔助檢索工具。

為突破上述困境，一套理想的智慧安防系統須具備三項核心能力：即時識別危險行為、跨攝影機快速追蹤嫌疑人，以及支援自然語言語意查詢以加速事後調查。本研究旨在整合電腦視覺、自然語言處理、工具調用推理與多代理人協作等技術，填補現行安防系統的空缺。

二、系統架構

(一)系統架構圖（雙流程並行）

本系統以攝影機即時串流為單一輸入來源，採用多流程同步並行的處理架構，將語言模型任務與即時視覺偵測分流為兩條平行後端流程。

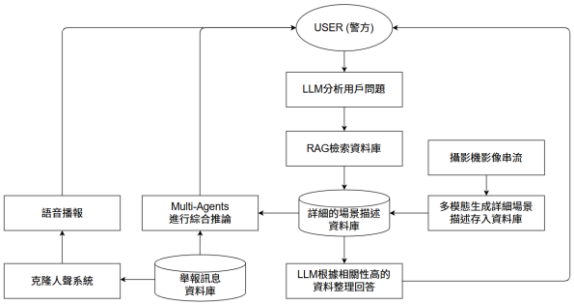

圖 1 系統架構圖(流程一)

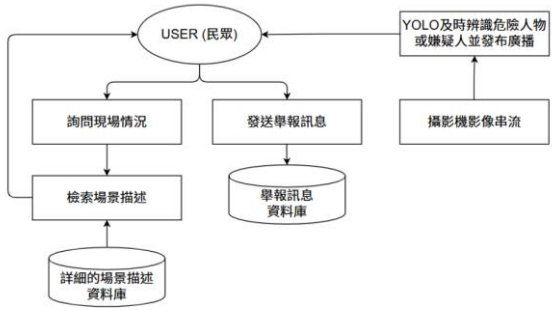

圖 2 系統架構圖(流程二)

(二)系統架構說明

1、流程一：多模態場景理解與 RAG 查詢

系統定期將監控畫面截圖送入本地端部署的多模態[4]大型語言模型（如：Gemma 3:27B），自動生成場景描述並存入向量資料庫。使用者介面整合檢索增強生成（RAG）[5]，並支援多重角色智慧體（Multi-Agents）交叉討論機制與 ReAct 框架，由輕量級模型負責意圖解析與任務路由，重量級模型生成高品質自然語言回覆。

2、流程二：即時危險偵測與警示通報

透過 YOLO 演算法[1]進行物件與人體姿態辨識，搭配 InsightFace 人臉辨識[2]與 DeepFashion 衣物分析[3]。當偵測到刀具等威脅時，系統即時觸發 LINE 警報推播給使用者；使用者亦可主動上傳場景資料，串接語音處理並，使系統能在網頁端提供即時語音互動，協助警方搜查並完善資料庫。

三、專題理論與實作

(一) 多特徵即時影像偵測與追蹤技術

系統採模組化多模型協同架構，分別運用 YOLO（空間定位）[1]、InsightFace（人臉特徵）[2]與 DeepFashion（外觀特徵）[3]進行分層處理。在即時危險偵測方面，透過提取人體手腕節點並計算其與危險物品中心的距離，當兩者距離在動態閾值內且持續一定時間，系統即判定該人物正持有危險物品，從而觸發危險事件通報與警報聲。

圖 3 後端陌生人持刀判斷畫面

在跨區追蹤方面，InsightFace[2]負責提取嫌疑人高維度人臉特徵向量，與資料庫名單進行相似度比對，若確認為新嫌疑人則自動生成唯一編號，供各區域監視器同步共享，實現跨鏡頭追蹤。

圖 4 嫌疑人不同區域辨識畫面

在輔助特徵提取方面，系統導入 DeepFashion 模型[3]對嫌疑人進行衣物區域分割，透過 K-Means 分群演算法萃取主色調，自動生成如「黑色/長袖上衣」、「紅色/長褲」等特徵描述，同步記錄於資料庫並即時推播給保全人員。

圖 5 DeepFashion 辨識結果

(二) 智能多特徵分析與 RAG 檢索

系統資料庫累積了具備豐富語意與結構化標籤（時間、地點、人物特徵）的歷史日誌，前端查詢機制導入 RAG 核心[5]架構。當接收到使用者查詢時，系統先透過意圖分析模組解析自然語言，萃取「時間範圍」、「地點關鍵字」及「查詢類型」，再動態決定向量資料庫的搜尋策略：時間型查詢加上時間區間 Metadata 過濾器；地點型查詢則進行地點屬性強過濾，最後結合語意相似度提取最符合的歷史監控紀錄。

(三) 多重角色智慧體（Multi-Agents）推論

系統定義三個具有不同身份的 LLM Agent（gpt oss:20B）——偵探、資深警官、年輕員警——使用序列式多代理人對話（Sequential Multi-Agent Dialogue）[16]，每輪依序發言並共享同一上下文。最後由第四個模型（Gemma 3:27B）擔任 。Aggregator Agent，整合所有發言產出最終結論。系統運用三種核心提示工程技術[19]：角色提示（Role Prompting）、情境學習（In-Context Learning）以及少樣本情境引導（Few-shot Context Priming）。

(四) 基於 ReAct 框架的推理與行動機制

系統導入 ReAct（Reasoning and Acting）推理架構[11]，由輕量級模型（ChatGLM3）[20]擔任任務路由器，結合短期對話記憶，進入「思考（Thought）→行動（Action）→觀察（Observation）」的迭代迴圈。思考階段採用 Few-Shot Prompting 與 In-Context Learning 分析使用者意圖[19]；行動階段輸出標準化 JSON 格式指令決定工具調用[8]；觀察階段接收後端回傳的結構化事實作為新上下文。資訊蒐集完整後，由重量級模型（Gemma3:27B）擔任生成器，將結構化事實轉譯為高品質自然語言回覆，實現運算資源最佳化配置。

(五)克隆人聲與自動播報系統

將 GPT-SoVITS 語音克隆技術整合至監控

即時通報系統，實現自動化語音播報功能，採用 GPT-SoVITS 雙階段管線。GPT-SoVITS 採用 Zero-Shot 方式進行語音克隆，僅需一段約 5 秒的參考音檔（ref_audio)即可複製目標音色，無需重新訓練模型。

在系統整合方面，本研究於伺服器部署 GPT-SoVITS API 服務 ，並開發人聲管理系統。

管理員可管理參考音檔，系統能更新音色， 當民眾透過通報 App 送出訊息後，新通報於前端即時播放，提醒警察、保全快速掌握現場情況 。

四、系統功能與呈現說明

(一) 多模態輸入與即時通報

左側面板整合語音與文字查詢介面，前端利用 Web Audio API 擷取音訊並透過 Whisper 語音辨識模型[7]解析。下半部透過輪詢（Polling）機制即時接收並顯示最新監控事件日誌，確保管理者隨時掌握現場狀況。

圖 6 用戶提問及收到的回覆

(二) 多功能展示標籤頁

右側面板採用標籤頁設計，提供三種檢視模式：

1、影像走廊檢視：系統回報的監控紀錄包含影像時，前端同步生成影像輪播器，支援全螢幕燈箱模式（Lightbox）放大檢視。

圖 7 攝影機點按放大觀看功能

2、地圖與座標標定工具：使用者可載入平面圖並點擊新增監視攝影機節點，系統自動換算相對座標（X, Y）並提供儲存或匯出功能。

圖 8 地圖標記的使用範例

3、Multi-Agents 討論區：使用者可設定討論回合數，介面以對話氣泡動態顯示不同角色的推理過程，三個角色輪流推理後由第四個角色整合生成綜合研判報告[13][14]。

圖 9 Multi-Agents 推理及總結的過程

(三) 隱藏式即時攝影機矩陣

系統支援多路影音串流即時預覽，點擊浮動按鈕可滑出 3×2 的攝影機矩陣面板。後端透過電腦視覺套件擷取 RTSP 影像幀，以 Multipart/x-mixed-replace 格式連續推播至前端，使用者可點擊單路攝影機放大查看即時影像細節。

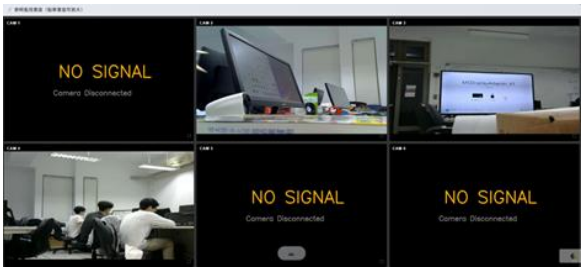

圖 10 即時串流攝影機的面板

(四) 克隆人聲管理系統與語音即時通報

克隆人聲管理系統提供管理者增加、刪除與選擇人聲模型，以及開關控制是否啟用人聲語音通報。

圖 11 人聲即時通報的操作畫面

五、專題作品影音平台網址

https://youtu.be/dHD2jTFtvec

六、參考文獻

[1] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[2] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4690–4699.

[3] Z. Ge, et al., "DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in Proceedings of CVPR, 2019.

[4] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

[5] Lewis, P., et al. "Retrieval-augmented generation for knowledge-intensive NLP tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.

[6] Reimers, N., and Iryna Gurevych. "Sentence-BERT: Sentence embeddings using Siamese BERT-networks." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.

[7] Radford, A., et al. "Robust speech recognition via large-scale weak supervision." International Conference on Machine Learning. PMLR, 2023.

[8] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom, “Toolformer: Language Models Can Teach Themselves to Use Tools,” arXiv preprint arXiv:2302.04761, 2023.

[9] A. Parisi, Y. Zhao, and N. Fiedel, “TALM: Tool Augmented Language Models,” arXiv preprint, 2022.

[10] E. Karpas, O. Abend, Y. Belinkov, B. Lenz, O. Lieber, N. Ratner, Y. Shoham, H. Bata, Y. Levine, K. Leyton-Brown, et al., “MRKL Systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning,” AI21 Labs, 2022.

[11] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, “ReAct: Synergizing reasoning and acting in language models,” International Conference on Learning Representations (ICLR), 2022.

[12] H. Liu, C. Sferrazza, and P. Abbeel, “Chain of Hindsight aligns Language Models with Feedback,” arXiv preprint, 2023.

[13] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework," *arXiv preprint arXiv:2308.08155*, 2023.

[14] S. Hong, X. Zheng, J. Chen, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, and C. Wu, "MetaGPT: Meta Programming for a Multi-Agent Collaborative Framework," *arXiv preprint arXiv:2308.00352*, 2023.

[15] T. Zhao, Z. Wang, Y. Li, Y. Liu, Y. Yang, J. Feng, and S. Chen, "Multi-Agent Collaboration Mechanisms: A Survey of LLMs," *arXiv preprint arXiv:2501.06322*, 2025.

[16] W. Y. Gu et al., "Explain-Analyze-Generate: A Sequential Multi-Agent Collaboration Framework," in

Proceedings of the 31st International Conference on Computational Linguistics (COLING), 2025, pp. 475.

[17] Xiao, M., Ye, M., Liu, B., Zong, X., Li, H., Huang, J., Xie, Q., & Peng, M. (2025). *A retrieval-augmented multi-agent framework for psychiatry diagnosis*. arXiv.

[18] Xiao, Mengxi, Mang Ye, Ben Liu, Xiaofen Zong, He Li, Jimin Huang, Qianqian Xie, and Min Peng. "A Retrieval-Augmented Multi-Agent Framework for Psychiatry Diagnosis." arXiv, June 4, 2025.

[19] Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications.arXiv.https://doi.org/10.48550/arXiv.2402.07927

[20] THUDM, "ChatGLM3," GitHub repository, 2023. [Online].Available:https://github.com/THUDM/ChatG

[21] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[22] RVC Boss, "GPT SoVITS: 1 min voice data can also be used to train a goodTTS model (few shot voice cloning),"GitHub repository,https://github.com/RVC-Boss/GPT- SoVITS. Accessed: Jun. 14, 2025.