Project Themis (泰美斯)

基於 VLM 與 RAG 的自動化數位鑑識與法律事實分析系統
(AI Digital Witness: Automated Forensic Analysis System)

姓名： 程煒倫
報告日期： 2025年12月

Project Themis is named after the Greek goddess of law and order, symbolizing an AI system that does not judge, but enables fair and evidence-based judgment.

摘要 (Abstract)
隨著監控設備在現代社會的普及，影像數據的產量呈指數級增長，但在法律實務與數位鑑識領域中，這些海量的監視器畫面多屬於非結構化的像素資訊，導致證據檢索面臨極大的挑戰。律師、檢察官或執法人員為了尋找特定的嫌疑犯或釐清過失責任，往往必須耗費數百小時進行人工肉眼過濾，這不僅效率低落，且極易因人為疲勞而產生疏漏。本專案「Project Themis」旨在開發一套 AI 數位證人系統，透過整合最先進的電腦視覺技術 YOLOv8 與 ByteTrack 進行物件偵測與追蹤，並引入視覺語言模型 Gemma 3 賦予系統「視覺理解」能力，最終結合 Milvus 向量資料庫與檢索增強生成（RAG）技術，建立一套可透過自然語言進行互動的證據檢索系統。本系統能自動將影像中的人物特徵轉化為結構化的文字描述，讓使用者僅需輸入如「尋找穿黃褲子的人」等自然語言指令，即可在秒級時間內調閱相關證據影片。本報告將詳細論述系統的架構設計、程式實作邏輯、模擬場景的實驗結果，並從法律人的視角，深入探討大型語言模型在法律科技領域的未來應用與倫理挑戰。

1. 創作動機與背景 (Motivation)
在法律實務工作中，無論是刑案偵查或是民事侵權糾紛的處理，核心證據往往來自於監視器畫面。然而，現有的監控系統雖然解決了「看見」的問題，卻無法解決「看懂」與「檢索」的難題。數據的盲目性成為了司法效率的瓶頸，例如在處理賣場滑倒意外或居家照護過失致死案件時，法律從業人員面臨的首要痛點便是時間成本過高。尋找一個關鍵的時間點、一個特定的證人，或是一個異常的行為片段，往往需要線性的回放海量影片，這在分秒必爭的訴訟程序中是極大的資源浪費。

此外，現有的影像證據缺乏結構化數據的支援。影片本質上是由像素組成的矩陣，缺乏語意標籤，這使得我們無法像使用搜尋引擎查找文字資料那樣，輸入關鍵字來查找影像內容。最後，在責任釐清的過程中，客觀的時序與行為紀錄至關重要。例如在長照案件中，看護究竟離開了受照顧者多久？長輩倒地後滯留了多少時間？這些關於「不作為」的證據，往往因為缺乏精確的量化數據而難以舉證。基於上述痛點，本專案提出了一種跨模態的解決方案，利用人工智慧將視覺訊號轉譯為法律證據，透過引入視覺語言模型讓電腦具備看圖說話的能力，自動生成類似警詢筆錄的人物特徵描述，並結合語意檢索技術，旨在為法律人打造一位不知疲倦、客觀中立的 AI 數位證人。

2. 系統架構 (System Architecture)

本系統的設計邏輯模擬了人類鑑識專家的認知流程，透過串聯不同的 AI 模型，完成了從感知、邏輯判斷、理解到歸檔的完整證據鏈處理。首先在視覺感知層，系統採用了 YOLOv8 物件偵測模型搭配 ByteTrack 多目標追蹤演算法。這個組合構成了系統的「眼睛」，負責從即時的影像流中精確地偵測出人體目標，並賦予每個目標唯一的追蹤 ID。ByteTrack 的優勢在於其強大的關聯能力，能夠有效處理目標短暫被遮擋或交錯的情況，確保了證據連續性，這在法律上至關重要，因為它能證明畫面中不同時間點出現的人物確實為同一對象，避免身份誤判。

在感知基礎之上，系統進入行為邏輯層，利用滯留演算法（Dwell Time Algorithm）對目標行為進行初步篩選。系統允許使用者定義特定的多邊形區域，並計算目標在該區域內的停留時間。這一層的設計具有深刻的法律意義，透過設定特定的時間閥值，例如五秒或更長，系統能夠自動過濾掉單純路過的無關人員，將運算資源集中在那些具有預謀、徘徊或異常滯留特徵的高價值目標上，從而大幅提升蒐證的精確度。

當目標被判定為需要關注的對象後，系統便進入視覺理解層，這是本專案的「大腦」。系統會觸發本地部署的 Gemma 3 視覺語言模型，對目標進行自動化特徵描述。在此階段，我們實作了一項關鍵的技術創新，即上下文感知裁切（Context-aware Cropping）。我們發現在傳統做法中，若僅裁切物件偵測的邊界框，往往會導致人物特徵不全，例如只截取到腹部而無法辨識衣著全貌。因此，我們在程式中加入了擴邊策略，保留人物周圍的環境上下文，這顯著提升了 VLM 對於性別、年齡估計以及衣著顏色辨識的準確度，降低模型產生錯誤推論（hallucination）的可能性可能性。

最後，所有的分析結果匯總至證據檢索層。系統利用 Milvus 向量資料庫結合 LangChain 框架實作檢索增強生成（RAG）。VLM 生成的文字描述被轉化為高維向量存儲，當使用者輸入自然語言查詢時，系統透過相似度搜尋召回最相關的證據片段，並由大型語言模型整理成結構化的鑑識報告，實現了從非結構化影像到結構化法律證據的轉化。

3. 程式實作解析 (Implementation Details)

本專案的程式實作分為前端的感知日誌生成與後端的 RAG 檢索兩大核心部分。在感知與描述系統中，程式維護了一個複雜的記憶體結構，用於記錄每個追蹤 ID 在不同區域的狀態，包括進入時間、滯留時長以及即時更新的最佳畫面。為了確保視覺語言模型能獲得品質最好的輸入，系統並非將每一幀畫面都送入模型，而是實作了一套最佳畫面選取機制，動態計算邊界框的面積與信心度，確保送出的是該人物在畫面中佔比最大、最清晰的時刻。此外，為了優化效能，系統設定了滯留時間門檻，只有當目標停留超過預設秒數時，才會調用本地的 Ollama API 進行特徵描述，這種設計不僅節省算力，更符合法律蒐證中「抓大放小」、聚焦異常行為的原則。

在數位鑑識檢索系統方面，程式負責將非結構化的日誌轉化為可對話的介面。我們選用了 nomic-embed-text-v1.5 模型來進行向量化，這是一個專為檢索任務優化的輕量級模型，能夠將如「穿黃褲子的男子」這類自然語言描述精確地轉換為數學向量。在建立索引時，我們採取了元數據感知的策略，不僅存入特徵描述的文本向量，還將追蹤 ID、原始影片檔名、滯留時間以及案發時間等關鍵資訊一併存入 Metadata。這使得後端的語言模型在回答查詢時，能夠精確地指出證據來源於哪一支影片檔案以及具體發生的時間點，滿足了法律證據對於「原始性」與「可驗證性」的要求。在 RAG 的生成階段，我們將 Llama 3.1 模型的溫度參數設為零，並透過精細的提示工程要求其扮演專業的數位鑑識專家，確保 AI 不會憑空捏造不存在的特徵，而是嚴格基於資料庫檢索到的內容進行客觀摘要。

4. 實驗結果：模擬案發現場 (Case Study)

為了驗證系統在複雜環境下的穩定性與效能，本專案選擇了高人流的零售賣場影片作為壓力測試的場景，模擬真實的蒐證情境。在第一個特徵檢索的實驗案例中，我們模擬警方接獲通報，需尋找一名穿著顯眼特徵的關鍵證人，輸入了「幫我找穿黃褲子、紅色拖鞋的人」作為查詢指令。系統在毫秒級的時間內，從數十筆複雜的人物紀錄中精準鎖定了追蹤 ID 為 297 的目標，並回傳了該目

標穿著亮黃色短褲與紅色拖鞋的準確描述，同時提供了對應的證據影片檔路徑。這一結果證明了 VLM 對於色彩與細微物件如拖鞋的辨識能力，已足以應用於實務上的特徵搜索，大幅縮短了偵查所需的時間。

在第二個關於行為分析與排除的實驗中，系統展現了其在邏輯判斷上的價值。數據顯示 ID 297 在結帳區域滯留長達 55 秒以上，在法律鑑識的語境下，這可被視為「可疑徘徊」或「潛在糾紛」的重要篩選指標，提示調查人員應優先調閱該片段。同時，系統也成功識別出 ID 1 穿著白襯衫與黑領帶，並透過語意理解推測其為工作人員或經理。這種自動化的身份識別功能，允許系統在尋找外部嫌疑人時自動排除內部員工，進一步提升了調查的效率與精確度，展示了 AI 在輔助法律判斷上的巨大潛力。

5. 深度討論：大型語言模型與多模態 AI 的未來展望

本專案透過 Project Themis 展示了 AI 在法律科技領域的初步應用與潛力。然而，隨著大型語言模型技術的飛速發展，未來的數位鑑識與法律實務將迎來更深刻的變革。首先是從視覺語言模型向原生多模態模型的演進。目前的系統採用的是串接式架構，即先透過物件偵測模型截圖，再送入語言模型分析，這中間不可避免地存在資訊折損，特別是時間連續性的動作特徵容易丟失。未來的原生多模態模型，如 Google Gemini Pro 或 GPT-4o 的後繼者，將具備直接觀看與理解長影片的能力。這意味著未來的 AI 數位證人將不再依賴靜態截圖，而是能理解動作的連續性與因果關係，例如它能區分「意外跌倒」與「遭人推擠」，或識別「竊取動作」與「整理貨架」的細微差異。這將使法律對於「行為意圖」（Mens Rea）的判斷更加精準，為司法提供更深層次的證據支持。

其次，代理人工作流（Agentic Workflow）的崛起將改變監控系統的被動屬性。目前的 RAG 系統仍是被動地等待使用者查詢，但結合了 Agent 技術後，AI 將具備主動調查的能力。在長照或高風險場景中，當 AI 偵測到長輩異常滯留或跌倒時，Agent 可以自主執行一連串的應變流程，包括調用 VLM 進行二次確認、判斷緊急程度，甚至主動控制攝影機轉向以取得更佳視角，並即時發送通知給相關人員。這將使監控系統從單純的紀錄者進化為具備決策能力的守護者。

然而，作為法律人，我們必須審慎看待 AI 證據的倫理與效力問題。本專案在實驗中也發現，模糊影像可能導致模型產生幻覺，例如將陰影誤判為衣物顏色。因此，未來的法律 AI 發展必須走向可解釋性 AI（XAI）。法庭需要的不只是結論，而是推論過程。未來的系統必須能提供「視覺引用」，即 AI 需在畫

面上框出它依據哪些像素特徵做出了判斷，而非僅給出文字描述。此外，必須建立嚴格的證據力分級制度，將 AI 報告視為偵查線索而非直接證據，並堅持「Human-in-the-loop」機制，由人類鑑識人員進行最終核實，以確保司法正義不受演算法偏見或錯誤的影響。最後，隨著邊緣運算技術的成熟，未來監視器本身將能運行輕量級模型，影像分析將在本地完成，僅傳輸結構化數據，這將徹底解決隱私法規的合規問題，讓智慧監控能更安全地應用於私密場景。

6. 結論 (Conclusion)

Project Themis 成功實作了一個端到端的自動化蒐證系統原型，證明了透過結合電腦視覺的「眼」、邏輯演算法的「腦」與 RAG 技術的「記憶」，我們能將傳統的監視器升級為具備語意理解能力的數位證人。從零售場景的客流分析壓力測試，到居家照護的責任釐清願景，這套系統展現了高度的泛化能力與實務價值。對於法律人而言，這不僅僅是效率的提升，將數百小時的搜尋工作縮短為數秒，更是一種實現客觀正義的科技手段。儘管目前的 VLM 技術仍有其局限性，但隨著技術的迭代與法律規範的完善，AI 必將成為司法體系中不可或缺的輔助力量，協助我們在海量的數據中，更快速、更精準地還原事實真相。