**盲人即時避障輔助系統設計與實作**

**一、研究動機**

　　隨著現代社會人口逐漸老化，世界上需要導航協助的人口也持續增加。其中，視障人士因視力受損，在日常生活中經常會遇到行動上的困難。目前，大部分視障人士仍主要依靠傳統的導盲杖或導盲犬來協助行走。然而，導盲杖雖然方便且便宜，但只能偵測腳下較近距離的障礙物，無法提早察覺前方遠處或動態的障礙物；而導盲犬雖然能夠幫助視障者更靈活地避開障礙物，但**養成成本極高，且需要大量資源投入**，並非所有視障者都能負擔得起。

　　近年來，深度學習技術在計算機視覺領域取得顯著進展，提供了更好的障礙物識別和空間理解能力，如YOLO物件偵測、DeepLab語義分割及MiDaS單目深度估計等。同時，視覺語言模型（Visual-Language Models, VLM），如CLIP和BLIP，展現出強大的場景理解與語義表達能力，促使AI輔助視障行動逐漸成為可能。

　　然而，現有的視障輔助軟體（如Be My Eyes）雖然能提供語言化的描述，但通常面臨回答冗長、推論延遲較大、無法即時提示的問題。因此，本研究希望融合深度學習計算機視覺（DLCV）與視覺語言模型（VLM）的優點，發展一個可即時運行於行動裝置的高效能視障輔助系統。



Figure 1. 視覺語言模型對場景的描述相當冗長

**二、相關研究**

**1. 基於 DLCV 的研究**

　　在深度學習電腦視覺（DLCV）領域中，目前主流的輔助視障者行走方法通常包含三個步驟：首先利用物件偵測模型（如YOLOv8）在影像中定位行人、自行車或其他潛在障礙物；接著透過語義分割模型（如DeepLabv3+）將場景分成可通行的道路區域與不可通行的障礙區域；最後，借助單目深度估計模型（如Depth‑Anything‑V3）為每個像素賦予相對深度資訊，並將深度圖與分割結果融合，以確定障礙物與使用者之間的實際距離。

一張含有 腳踏車, 陸上交通工具, 拼貼畫, 腳踏車輪組 的圖片

AI 產生的內容可能不正確。

Figure 2. 物件偵測（左上），深度估計（右上），物件分割（下）

　　Dubey等人2023年提出了結合YOLOv3及DepthNet的方法。他們首先使用YOLOv3偵測障礙物，接著用光流法估計物體的運動方向和速度，並透過DepthNet對畫面進行深度預測。相機畫面劃分為左、中、右三個區域，並依據障礙物出現在哪個區域輸出的提示。

一張含有 文字, 螢幕擷取畫面, 設計 的圖片

AI 產生的內容可能不正確。

Figure 3. Dubey 2023

　　這種方法的優點在於它能同時獲取物體的類別、運動資訊與深度估計，並透過畫面分區的方式直接對應到使用者的左右轉向。然而，系統依賴固定的三區切割規則，對於物體跨越多個方向或在複雜背景中運動的場景，**泛化能力有限**。

**2. 基於 VLM 的研究**

　　近年來，視覺語言模型（Visual-Language Models，簡稱VLM）逐漸成為視障輔助領域的新興方向。VLM結合圖像、語言編碼器與語言解碼器，使模型能夠同時理解畫面內容與生成自然語言描述。一張含有 文字, 螢幕擷取畫面 的圖片

AI 產生的內容可能不正確。

Figure 4. SpatialVLM資料集範例

　　在空間推論的領域上，SpatialVLM將VLM訓練在大量關於空間推論的資料集上，以圖像和文字共同作為輸入，生成空間關係的自然語言描述。研究結果顯示，模型的空間推論能力明顯提升，能非常準確的描述物體間的相對位置（如上下、左右），也能大致估算物體間的距離。一張含有 文字, 螢幕擷取畫面 的圖片

AI 產生的內容可能不正確。

Figure 5. WAD資料集

　　WalkVLM進一步將VLM空間推論的能力應用在盲人行走輔助上，建立了名為「走路覺察資料集（Walking Awareness Dataset，WAD）」的資料集。13小時的第一人稱視角胸前攝影機拍攝影片，包含都市街道、人行道、路口等。依照視障人士可能面臨的環境要素進行分類與描述，包括天氣狀況、交通流量、障礙物種類與位置等。

一張含有 文字, 螢幕擷取畫面, 網站, 網路廣告 的圖片

AI 產生的內容可能不正確。

Figure 6. WalkVLM模型架構

　　WalkVLM透過VLM的視覺編碼器（ViT）與語言解碼器（基於大語言模型，如PaLM-E）作為基礎。首先透過視覺模組感知並識別畫面中重要的元素，例如障礙物、街道地標及天候情況等；接著再將這些資訊整合；最後進行決策階段，輸出簡短、明確且適合視障人士即時採取行動的語音提醒。

此外，WalkVLM還引入「時間感知自適應控制（Temporal-Aware Adaptive Prediction, TAP）」，利用時間序列資訊動態判斷何時提供提醒，避免重複且不必要的警告來將低運算負擔和提升即時性。

　　儘管 WalkVLM 成功將深度學習電腦視覺（DLCV）與視覺–語言模型（VLM）結合，但在實際路測過程中，仍出現許多缺點。  
**1. 空間理解不足，難以掌握道路走向**

　　WalkVLM 能辨識畫面是否「空曠」，卻無法進一步理解人行道或道路的彎曲走向。例如在圖中，模型僅回報「前方通暢，請繼續前行」，卻未注意到道路右側正彎向建築內縮的走道。對視障者而言，這類彎道往往伴隨欄杆或凸起路緣，一旦直行未轉，就會偏離安全區，造成跌倒或碰撞風險。

一張含有 戶外, 天空, 樹狀, 棕櫚樹 的圖片

AI 產生的內容可能不正確。

Figure 7. 道路彎曲 / 認為前方空曠

**2. 無法區分最危險的障礙物，提醒方向常常不精準**

　　當前場景同時出現多種障礙物時，WalkVLM 往往只會生成「避開某一物件」的提示，而沒有能力判斷哪個目標最可能發生碰撞。譬如在一輛公車與人、欄杆同時出現的情況下，模型將公車列為首要避開對象。

一張含有 文字, 陸上交通工具, 戶外, 車輛 的圖片

AI 產生的內容可能不正確。

Figure 8. 提醒避開公車而不是更可能碰撞的障礙物

1. **幻覺，誤報不存在的物體**

　　例如在人行道上，WalkVLM 回報「前方有車輛」，這種幻覺不只干擾使用者判斷，也降低了整體系統的可信度。

一張含有 文字, 戶外, 足部穿著, 樹狀 的圖片

AI 產生的內容可能不正確。

Figure 9. 認為前方有車輛

**三、研究目的**

　　比較 DLCV 和 VLM 這兩種方法的優缺點。Dubey 方法的優點是對幾何資訊具有較高的準確性，但在處理複雜場景時容易失效。相較之下，WalkVLM 則能夠更好地應對更多樣且複雜的場景，然而其在空間推理方面的表現仍不夠精細。

　　綜合前面的觀察，我們認為要同時擁有 DLCV 的幾何準確性和 VLM 的語意理解能力，就需要把兩者的優點結合起來，**把更完整的空間資訊（例如道路的彎曲程度、障礙物的距離和移動方向）加入到提示內容與觸發機制中，讓 VLM 能在可靠的空間基礎上進行判斷。**

　　基於這個想法，本研究有三個主要目標：

1. 讓更多 DLCV 的資訊（如物件邊界、深度分布和路徑預測）可以和 VLM 的圖文理解進行融合。

2. 比較不同的觸發方式在真實場景中的表現，找出最適合行走避障的方式。

3. 利用胸前放置的單目手機攝影機作為唯一的感測器，研發一套能夠即時進行障礙偵測與避障的實用系統。

一張含有 寫生, 圖畫, 線條藝術, 白色 的圖片

AI 產生的內容可能不正確。

Figure 10. 示意圖

## **研究方法**

1. **資料蒐集與標註**

　使用 WalkVLM 提供的資料，用現有的模型取得空間上的資訊：

* **物件偵測（YOLO）**：障礙物位置
* **語義分割（DeepLabv3+）**：可行走區域
* **深度估計（Depth-Anything）**：得到物體與使用者之間的距離

1. **比較不同輸入下WalkVLM的表現**

　　將前一步驟得到的 DLCV 結果整理、設計成簡短的文字提示，輸入原本的WalkVLM，提供更多的視覺資訊，和原始方法的表現比較。

1. **微調模型**

　　使用 LoRA（Low-Rank Adaptation）嘗試微調 WalkVLM，讓模型適應運用更多視覺資訊的狀況，以較表現。

1. **比較不同的提醒觸發模式**

* 固定頻率觸發：每隔幾秒就執行一次完整的偵測與提醒。
* 事件驅動觸發：當 DLCV 偵測到重要事件（障礙物距離幾公尺內）觸發。
* 其他觸發邏輯：設計其他動態或自適應的觸發邏輯。

1. **主觀評估**

## 邀請受試者，在學校校園、市區街道等多個真實場景中進行系統實際測試與錄影。記錄整個操作過程，並進行案例分析，以檢驗系統在真實環境中是否能夠順利地幫助使用者避開障礙物。而後蒐集使用者的主觀回饋，並透過訪談或問卷方式，評估系統提醒的清晰度、實用性以及整體滿意度。

## **五、參考資料**

1. **Redmon, J. et al.** “You Only Look Once: Unified, Real-Time Object Detection.” CVPR, 2016.
2. **Chen, L. et al.** “Depth-Anything: Universal Monocular Depth Estimation with Sequence Modeling.” arXiv:2406.09414, 2024.
3. **Ranftl, R. et al.** “Vision Transformers for Dense Prediction.” CVPR, 2024.
4. **Liu, Z. et al.** “YOLOv8 Documentation.” Ultralytics. <https://docs.ultralytics.com/>
5. **Dubey, A. et al.** “Guidance System for Visually Impaired Persons using Deep Learning and Optical Flow.” arXiv:2310.14239, 2023.
6. **Li, X. et al.** “SpatialVLM: Learning Spatial Reasoning in Vision-Language Models.” CVPR, 2024.
7. **WalkVLM Team.** “WalkVLM: Aid Visually Impaired People Walking by Vision Language Model.” <https://walkvlm2024.github.io/>
8. **He, K. et al.** “Deep Residual Learning for Image Recognition.” CVPR, 2016.
9. **Kuo, S. et al.** “DeepLabv3+: Encoder-Decoder with Atrous Separable Convolution.” TPAMI, 2018.
10. **Kendall, A. et al.** “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?” NIPS, 2017.