BÀI BÁO CÁO CUỐI KỲ NHẬP MÔN XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Lê Đại Hòa - 22120108

Nguyễn Tường Bách Hỷ - 22120455

Liêu Hải Lưu Danh - 22120459

Lê Hoàng Vũ - 22120461

Ngày 25 tháng 12 năm 2024

Mục lục

I

TÓM TẮT NỘI DUNG

1 Tóm tắt nội dung báo cáo

Kính gửi Quý Thầy/Cô,
Nhóm em xin phép được nộp bài báo cáo đồ án cuối kỳ Nhập môn Xử lý ngôn ngữ tự nhiên. Nội dung của bài báo cáo này phản ánh quá trình và kết quả tìm hiểu của chúng em về nội dung: **Hỗ trợ người khiếm thị đi lại**. Trong quá trình thực hiện, nhóm em đã cố gắng hoàn thành các yêu cầu đề ra một cách tốt nhất có thể.

Tóm tắt báo cáo

Báo cáo này trước hết trình bày bảng phân công công việc giữa các thành viên trong nhóm. Tiếp đến sẽ đi sâu vào phần nội dung cốt lõi của đồ án, mà ở đó nhóm làm rõ những giải pháp của mình với từng vấn đề gặp phải, cũng như đưa ra những vấn đề còn tồn đọng, những khó khăn trong quá trình thực hiện. Cuối cùng là những thông tin hữu ích để bổ sung cho Quý Thầy/Cô khi chấm điểm, chẳng hạn như `video` hay `slide`. Quý Thầy/Cô có thể theo dõi `source code` ở đây.

Quy trình	**Nội dung**	**Tự đánh giá mức độ hoàn thành**
Dữ liệu `training` và `test`		
`Thu thập dữ liệu`	`[Lê Đại Hòa]` Tập trung vào dữ liệu ở loại: hình ảnh tình huống giao thông đường bộ có vật cản và có biển báo	Tốt (100%)
Xử lý dữ liệu	`[Liêu Hải Lưu Danh]` Huấn luyện mô hình `CNN` để phân loại bộ dữ liệu và lọc ra những hình ảnh chất lượng đủ tốt để `server` xử lý về sau	Tốt (100%)
Công nghệ dùng cho `Server`		
Mô hình học sâu	`[Lê Hoàng Vũ]` Cụ thể là áp dụng 2 mô hình học sâu của `Google Cloud`: `Speech-to-text` để chuyển đổi giọng nói thành văn bản, và `Text-to-Speech` để chuyển văn bản thành giọng nói tự nhiên.	Tốt (100%)
`Gemini-1.5` kết hợp `FastAPI Framework`	`[Liêu Hải Lưu Danh]` Áp dụng `FastAPI` là `Framework Python` hiện đại để xây dựng `API`, ở đó dùng `Gemini-1.5` là mô hình đa phương thức mới nhất của `Google` để trả lời những câu hỏi đến từ `Client`, đặc biệt là kết quả nhận diện hình ảnh	Tốt (100%)
Xây dựng ứng dụng Mobile		

Client và Server	[Nguyễn Tường Bách Hỷ] Xây dựng được nền tảng cho ứng dụng Mobile với các tính năng cơ bản cho Client	Tốt (100%)
Thiết kế GUI	[Lê Hoàng Vũ] Thiết kế model cho ứng dụng bằng Figma	Tốt (100%)
Resource		
Video	[Lê Đại Hòa] Edit video giới thiệu và demo sản phẩm	Tốt (100%)
Slide	[Lê Đại Hòa] Thiết kế Slide giới thiệu sản phẩm	Tốt (100%)
Báo cáo	[Lê Hoàng Vũ, Nguyễn Tường Bách Hỷ] Viết báo cáo trình bày lại những nội dung cơ bản của đồ án	Tốt (100%)
Những nội dung đang trong quá trình phát triển thêm		
Siri Wave	[Nguyễn Tường Bách Hỷ] Ứng dụng công nghệ sóng của Siri khi giọng nói AI được sử dụng	Đang phát triển (80%)
Map Tracking	[Nguyễn Tường Bách Hỷ] Định hướng người dùng tới nơi được chỉ định, tích hợp audio	Đang phát triển (80%)

Lời kết

Nhóm em xin chân thành cảm ơn Thầy/Cô đã hướng dẫn và hỗ trợ trong suốt quá trình học tập. Nhóm đã cố gắng trình bày báo cáo một cách ngắn gọn, súc tích và đi thẳng vào trọng tâm của vấn đề. Nếu có thiếu sót, nhóm rất mong nhận được góp ý từ Thầy/Cô để có thể hoàn thiện tốt hơn ở những lần sau.

2 Tài liệu tham khảo

[Ah16] Ahmetovic, D., Gleason, C., Kitani, K. M., Takagi, H., & Asakawa, C. NavCog: turn-by-turn smartphone navigation assistant for people with visual impairments or blindness. In *Web for All Conference (W4A'16)*. ACM, Article 9, 1-2. URL: https://doi.org/10.1145/2899475.2899509 (cited pp. 1-2).

[Wi15] Wiegand, K., Schmitz, B., & Kurschl, W. BlindSquare: A location-based application to support visually impaired people. In *International Conference on Computers Helping People with Special Needs* (pp. 165-172). Springer. URL: https://doi.org/10.1007/978-3-319-20678-3_17 (cited pp. 165-172).

[Fa13] Fallah, N., Apostolopoulos, I., Bekris, K., & Folmer, E. Indoor human navigation systems: A survey. *Interacting with Computers*, 25(1), 21-33. URL: https://doi.org/10.1093/iwc/iws010 (cited pp. 21-33).

[Re16] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779-788). URL: https://doi.org/10.1109/CVPR.2016.91 (cited pp. 779-788).

[Pu20] Pustokhina, I. V., Pustokhin, D. A., Gupta, D., Khanna, A., Shankar, K., & Nguyen, G. N. An effective training scheme for deep neural networks in edge computing-enabled Internet of Medical Things (IoMT) for smart healthcare system. *Neural Computing and Applications*, 32, 15897-15911. URL: https://doi.org/10.1007/s00521-020-04998-1 (cited pp. 15897-15911).

[Ho08] Hoyle, B., & Waters, D. Mobility AT: The Ultracane does work! *Access Journal*, 9(4), 15-20. URL: https://doi.org/10.1080/10400435.2008.10131938 (cited pp. 15-20).

[Be73] Benjamin, J. M., Ali, N. A., & Schepis, A. F. A laser cane for the blind. *Proceedings of the San Diego Biomedical Symposium*, 12, 53-57. URL: https://doi.org/10.1109/SBIO.1973.453527 (cited pp. 53-57).

[Go04] Golledge, R. G., Marston, J. R., Loomis, J. M., & Klatzky, R. L. Stated preferences for components of a personal guidance system for nonvisual navigation. *Journal of Visual Impairment & Blindness*, 98(3), 135-147. URL: https://doi.org/10.1177/0145482X0409800305 (cited pp. 135-147).

[Wi07] Wilson, J., Walker, B. N., Lindsay, J., Cambias, C., & Dellaert, F. SWAN: System for wearable audio navigation. In *11th IEEE International Symposium on Wearable Computers* (pp. 91-98). URL:

https://doi.org/10.1109/ISWC.2007.4373786 (cited pp. 91-98).

[Ma12] Manduchi, R., & Coughlan, J. (Computer) vision without sight. *Communications of the ACM*, 55(1), 96-104. URL: https://doi.org/10.1145/2063176.2063200 (cited pp. 96-104).

[Da10] Dakopoulos, D., & Bourbakis, N. G. Wearable obstacle avoidance electronic travel aids for blind: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(1), 25-35. URL: https://doi.org/10.1109/TSMCC.2009.2021255 (cited pp. 25-35).

[Ab23] Abidi, M. H., Siddiquee, A. N., Alkhalefah, H., & Srivastava, V. A comprehensive review of navigation systems for visually impaired individuals. URL: https://doi.org/10.1155/2023/9065474 (cited pp. 1-2).

[Kh21] Khan, S., Nazir, S., & Khan, H. U. Analysis of Navigation Assistants for Blind and Visually Impaired People: A Systematic Review. *IEEE Access*, 9, 26712-26730. URL: https://doi.org/10.1109/ACCESS.2021.3057715 (cited pp. 26712-26730).

II

VISION WALK

3 Giới thiệu

Di chuyển trên các con phố đông đúc hoặc trong các không gian công cộng là một thách thức lớn đối với những người khiếm thị. Việc không thể phát hiện ra các vật cản, biển báo giao thông hoặc thay đổi trong môi trường xung quanh thường dẫn đến rủi ro cao về tai nạn và giảm sự độc lập trong việc di chuyển. Khi công nghệ ngày càng phát triển, các giải pháp sử dụng Trí tuệ Nhân tạo (`AI`) đã xuất hiện như một công cụ tiềm năng trong việc hỗ trợ người khiếm thị di chuyển an toàn và hiệu quả.

Dự án `VisionWalk` nhằm giải quyết những khó khăn này bằng cách phát triển một trợ lý dẫn đường sử dụng Trí tuệ Nhân tạo, được thiết kế đặc biệt để hỗ trợ người khiếm thị trong việc di chuyển trong thời gian thực. Bằng cách sử dụng các kỹ thuật thị giác máy tính và các mô hình `AI`, `VisionWalk` sẽ nhận diện các biển báo giao thông quan trọng, các vật cản và các nguy cơ tiềm ẩn trên đường đi, đồng thời cung cấp cảnh báo âm thanh đúng lúc để hướng dẫn người dùng.

VISIONWALK

BEYOND VISION WITHIN REACH

Ứng dụng này sẽ tập trung vào ba bước chính: Nhận diện biển báo giao thông trong thời gian thực, phát hiện các vật cản tĩnh và di động, và cuối cùng là cung cấp cảnh báo và chỉ dẫn thời gian thực giúp người dùng di chuyển an toàn. Sử dụng camera của điện thoại di động được kết nối thông qua ứng dụng, hệ thống sẽ nhận diện và phân loại các biển báo giao thông như biển cấm người đi bộ, vạch qua đường dành cho người đi bộ, cùng với các vật cản như cột đèn, xe cộ, hoặc khu vực công trình thi công. Ngoài ra, ứng dụng còn tính toán khoảng cách và hướng di chuyển của các vật cản và cảnh báo người dùng tương ứng.

Hệ thống sẽ được tối ưu hóa để hoạt động trên các nền tảng di động phổ biến, giúp người dùng dễ dàng tiếp cận mà không cần thiết bị phần cứng đặc biệt. Thông qua dự án này, chúng tôi không chỉ muốn cải thiện sự độc lập và an toàn cho người khiếm thị mà còn góp phần xây dựng một xã hội hòa nhập hơn, nơi công nghệ `AI` có thể được ứng dụng để nâng cao chất lượng sống cho nhóm người yếu thế.

4 Ý tưởng tiếp cận

§4.1 Những công trình nghiên cứu liên quan

Trong những năm gần đây, các nghiên cứu và ứng dụng công nghệ AI cho người khiếm thị đã đạt được nhiều tiến bộ đáng kể, đặc biệt là trong lĩnh vực hỗ trợ di chuyển và phát hiện vật cản. Một số hệ thống và nghiên cứu nổi bật trong lĩnh vực này bao gồm:

1. Wayfinder - Hệ thống hỗ trợ người khiếm thị di chuyển trong môi trường không gian công cộng: Hệ thống Wayfinder sử dụng công nghệ GPS và cảm biến để giúp người khiếm thị di chuyển trong các khu vực công cộng như trung tâm thương mại hoặc sân bay. Tuy nhiên, hệ thống này chủ yếu sử dụng tín hiệu âm thanh để cung cấp thông tin và không thực sự tập trung vào việc nhận diện các vật cản hoặc biển báo giao thông trong môi trường xung quanh.

2. NAVCog - Ứng dụng dẫn đường cho người khiếm thị sử dụng công nghệ Bluetooth Low Energy (BLE): NAVCog là một ứng dụng hỗ trợ người khiếm thị trong việc tìm đường trong các khu vực nội bộ, chẳng hạn như các tòa nhà hoặc bệnh viện. Hệ thống này sử dụng BLE để xác định vị trí và cung cấp hướng dẫn qua tín hiệu âm thanh. Tuy nhiên, ứng dụng này không cung cấp khả năng nhận diện và cảnh báo về các vật cản di chuyển hoặc biển báo giao thông, mà chỉ tập trung vào dẫn đường.

3. BlindSquare - Hệ thống dẫn đường cho người khiếm thị sử dụng GPS: BlindSquare là một ứng dụng di động được thiết kế để giúp người khiếm thị điều hướng môi trường ngoài trời thông qua hệ thống định vị GPS. Ứng dụng cung cấp thông tin về các địa điểm xung quanh và cảnh báo về các vật thể cố định. Tuy nhiên, BlindSquare không tích hợp khả năng phát hiện vật cản động, như xe cộ hoặc người đi bộ khác, trong khi các mối nguy hiểm này thường xuyên gây ra sự cố trong quá trình di chuyển của người khiếm thị.

4. Sónar-based Systems - Hệ thống sử dụng sóng siêu âm và cảm biến khoảng cách: Một số hệ thống hỗ trợ người khiếm thị sử dụng cảm biến siêu âm để phát hiện vật cản trước mặt và cảnh báo cho người sử dụng qua âm thanh. Ví dụ, hệ thống Ultracane sử dụng cảm biến siêu âm để phát hiện các vật cản tĩnh và di chuyển trong phạm vi gần. Tuy nhiên, các hệ thống này có hạn chế trong việc nhận diện các vật cản nhỏ hoặc không thể phát hiện vật cản di chuyển như xe cộ, và độ chính xác của chúng có thể bị ảnh hưởng bởi môi trường xung quanh.

5. Deep Learning for Object Detection: Việc áp dụng các mô hình học sâu (Deep Learning) trong nhận diện đối tượng đã cho thấy hiệu quả rõ rệt trong các bài toán nhận diện vật cản và biển báo giao thông. Các mô hình như YOLO (You Only Look Once) và Faster R-CNN đã được sử dụng trong các ứng dụng nhận diện biển báo và vật thể trong nhiều lĩnh vực khác nhau. Các nghiên cứu như của **Redmon et al. (2016)** và **Ren et al. (2015)** đã chứng minh khả năng của YOLO và Faster R-CNN trong việc phát hiện các đối tượng trong ảnh với độ chính xác cao và tốc độ xử lý nhanh.

6. `Vision-based Assistive Technologies for Blind`: Các nghiên cứu về công nghệ hỗ trợ người khiếm thị dựa trên thị giác sử dụng camera điện thoại và các mô hình học máy để phát hiện biển báo giao thông và vật cản. Ví dụ, nghiên cứu của **Pustokhina et al. (2016)** giới thiệu một hệ thống phát hiện biển báo giao thông bằng cách sử dụng mô hình học sâu với dữ liệu ảnh từ camera di động. Tuy nhiên, những hệ thống này chưa hoàn thiện trong việc nhận diện đa dạng các vật cản di chuyển và tính toán khoảng cách chính xác.

❒ Mặc dù các hệ thống hiện tại đã mang lại nhiều lợi ích cho người khiếm thị trong việc di chuyển và nhận diện vật cản, nhưng hầu hết vẫn thiếu tính năng phát hiện các vật cản động (như xe cộ và người đi bộ) hoặc khả năng nhận diện biển báo giao thông chính xác và thời gian thực.

§4.2 Cách tiếp cận của VisionWalk

Để thực hiện ứng dụng `VisionWalk`, một trợ lý điều hướng thông minh cho người khiếm thị, chúng tôi sẽ sử dụng sự kết hợp của các mô hình học máy và xử lý hình ảnh thời gian thực để nhận diện biển báo giao thông và vật cản, từ đó đưa ra cảnh báo âm thanh cho người sử dụng. Dưới đây là những bước chính trong quá trình tiếp cận và xây dựng hệ thống:

1. Thu thập dữ liệu ứng dụng sử dụng camera điện thoại làm thiết bị thu thập hình ảnh/`video` từ môi trường xung quanh người dùng. Chức năng thu thập này sẽ được diễn ra khi người dùng bấm vào nút **Capture** ở một vị trí dễ để người khiếm thị bấm trên màn hình.

2. Xử lý Dữ liệu và Nhận diện Biển báo:
   - Ảnh trước hết sẽ được tiền xử lý để lọc ra những ảnh hợp lệ mà mô hình ở `Server` có thể cho ra kết quả với độ tin cậy cao. Nói cách khác, những hình ảnh bị mờ hoặc vô tình bị che phủ bởi tay người dùng khi chụp sẽ bị loại và ứng dụng trả ra cảnh báo.
   - Phát hiện vật cản: `Gemini-1.5` - Mô hình `AI` đa phương thức của `Google`, sẽ giúp phát hiện các vật cản trên đường như cột đèn giao thông, biển báo, gốc cây, hay thậm chí các phương tiện giao thông như xe máy, ô tô và người đi bộ.
   - Biển báo giao thông: `Gemini-1.5`, kết hợp với việc huấn luyện các bộ dữ liệu về biển báo, sẽ được sử dụng để nhận diện các loại biển báo giao thông trong ảnh. Các biển báo như *"Cấm người đi bộ"*, *"Trẻ em qua đường"*, và *"Đường trơn trượt"* sẽ được ưu tiên nhận diện. Kết quả trả về bao gồm tên biển báo và vị trí của nó trong hình ảnh, từ đó có thể đưa ra cảnh báo âm thanh cho người dùng.

3. Cảnh báo và Điều hướng: Sau khi các biển báo và vật cản được phát hiện, ứng dụng sẽ sử dụng công nghệ `Text-to-Speech (TTS)` của `Google Cloud` để phát âm thanh cảnh báo. Đồng thời quá trình này đi kèm với xử lý và loại bỏ nhiễu trong cảnh báo âm thanh, sử dụng các công cụ tích hợp sẵn trong `Google Cloud Speech`. Các cảnh báo âm thanh này sẽ giúp người dùng nhận thức rõ ràng các tình huống nguy hiểm hoặc hướng dẫn họ di chuyển an toàn, ví dụ: *"Có biển báo Cấm người đi bộ cách bạn* $3m$ hay *"Có xe máy cách bạn* $1.5m$ *bên phải"*.

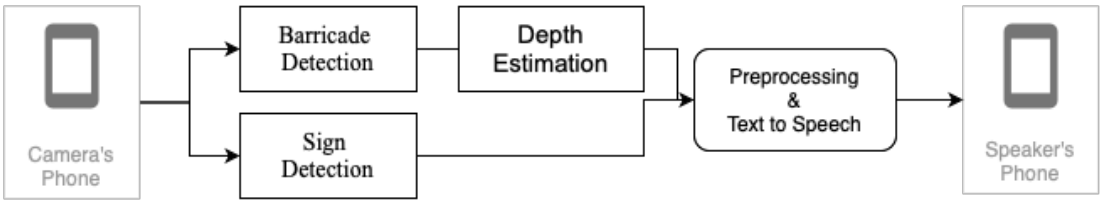
4. Tối ưu hóa và triển khai trên thiết bị di động:
    - Tối ưu hóa mô hình: Để ứng dụng hoạt động mượt mà trên các thiết bị di động phổ thông, các mô hình học máy sẽ được tối ưu hóa bằng các công nghệ như `TensorFlow Lite` hoặc `PyTorch Mobile`. Các kỹ thuật như `Quantization` và `Pruning` cũng sẽ được sử dụng để giảm kích thước mô hình và tăng tốc độ xử lý.
    - Hỗ trợ đa nền tảng: Ứng dụng được phát triển bằng `React Native`, kết hợp với `Expo`, cho phép chạy trên cả `Android` và `iOS`. Quá trình phát triển ứng dụng sử dụng `TypeScript`, đảm bảo mã nguồn dễ bảo trì và mở rộng.
5. Đảm bảo thời gian thực: Để cung cấp cảnh báo kịp thời và hiệu quả, ứng dụng cần xử lý và phân tích dữ liệu trong thời gian thực. Đảm bảo rằng độ trễ khi phân tích hình ảnh ở khoảng chấp nhận được để giúp người dùng nhận cảnh báo nhanh chóng và chính xác.
6. Phát triển và triển khai: Sau khi mô hình được huấn luyện và tối ưu hóa, ứng dụng sẽ được phát triển, thử nghiệm, và triển khai trên các thiết bị di động. Quá trình thử nghiệm sẽ tập trung vào việc cải thiện độ chính xác của các mô hình phát hiện và cảnh báo; giảm thiểu độ trễ, tối ưu hóa tốc độ xử lý của mô hình; cũng như nâng tầm trải nghiệm người dùng.

❒ Tóm lại, ý tưởng tiếp cận của `VisionWalk` là dựa trên sự kết hợp của các mô hình học sâu mới nhất như `Google Cloud`, hay mô hình ngôn ngữ lớn đa chức năng như `Gemini-1.5` để cung cấp khả năng nhận diện chính xác biển báo giao thông và vật cản, đồng thời sử dụng các mô hình đo chiều sâu để tính toán khoảng cách và hướng di chuyển của các vật cản. Điều này giúp người dùng nhận được cảnh báo chính xác và kịp thời về các nguy cơ, từ đó cải thiện khả năng tự di chuyển và an toàn cho người khiếm thị.

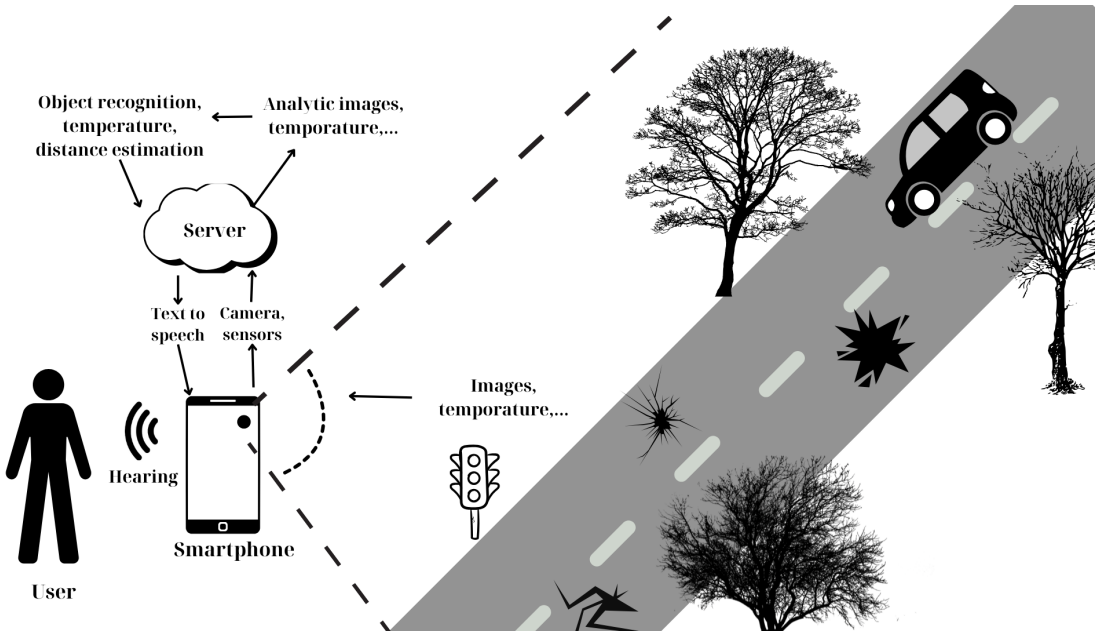
5 Mô hình hoạt động và Công nghệ

§5.1 Mô hình cơ bản

Đầu tiên quan sát một mô hình hoạt động đơn giản như sau mà hầu như tất cả các dự án về định hướng cho người khiếm thị đều sử dụng:



Hình 5.1: Mô hình cơ bản cho một hệ thống định hướng



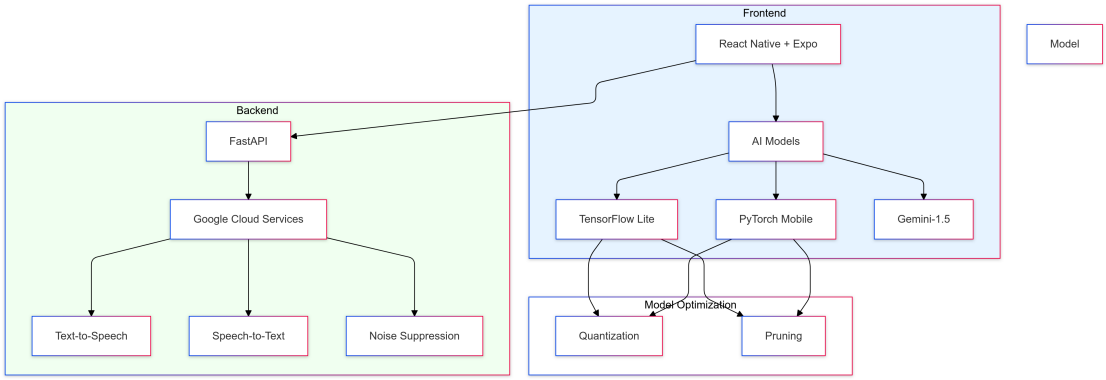
Hình 5.2: Minh họa mô hình

§5.2 Công nghệ sử dụng

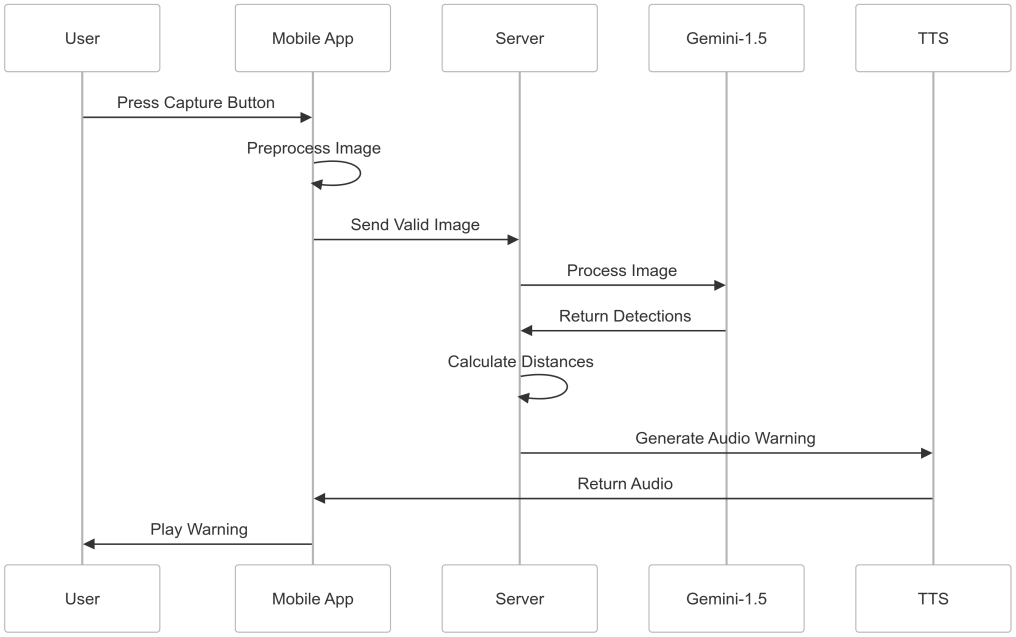
Để làm hoàn thiện hơn mô hình cơ bản, ta cần phải tăng cường bằng những công nghệ hỗ trợ như sau:

1. Xử lý hình ảnh và mô hình AI:
   - `Gemini-1.5` của Google: Phát hiện vật cản và nhận diện biển báo giao thông.
   - `TensorFlow Lite` và `PyTorch Mobile`: Tối ưu hóa mô hình học máy.
2. Công nghệ giao tiếp giữa người với `AI`:
   - `Google Cloud Text-to-Speech (TTS) và Speech-to-Text (STT)`: Để ứng dụng phát âm thanh cảnh báo cho người dùng và ngược lại người dùng có thể tương tác với máy qua giọng nói.
   - `Noise Suppression` (Khử nhiễu âm thanh): Xử lý và loại bỏ nhiễu trong cảnh báo âm thanh, sử dụng các công cụ tích hợp sẵn trong `Google Cloud Speech`.
   - Công nghệ tạo sóng âm khi `AI` nói, như trong các trợ lý ảo Siri hay Alexa: tham khảo tại đây.
3. Phát triển ứng dụng di động: Ứng dụng được phát triển bằng `React Native`, kết hợp với `Expo`, cho phép chạy trên cả `Android` và `iOS`. Việc sử dụng `iOS` đảm bảo mã nguồn dễ bảo trì và mở rộng. Trong đó sử dụng `Framework FastAPI` để xây dựng `API`.
4. Kỹ thuật tối ưu hóa mô hình: `Quantization` và `Pruning` để giảm kích thước và tăng tốc độ xử lý các mô hình.

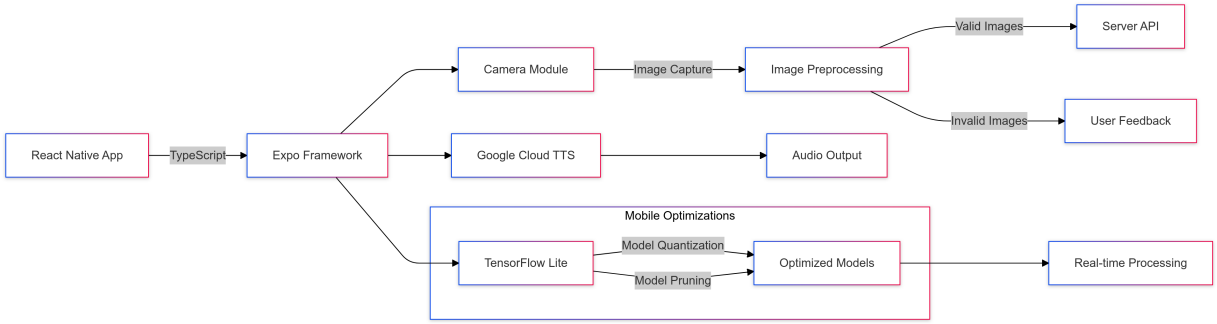
Để hình dung được việc áp dụng các công nghệ này như thế nào, hãy quan sát biểu đồ hoạt động sau:



Hình 5.3: Các mô hình và công nghệ sử dụng trong `VisionWalk`



Hình 5.4: Quy trình xử lý của `VisionWalk`

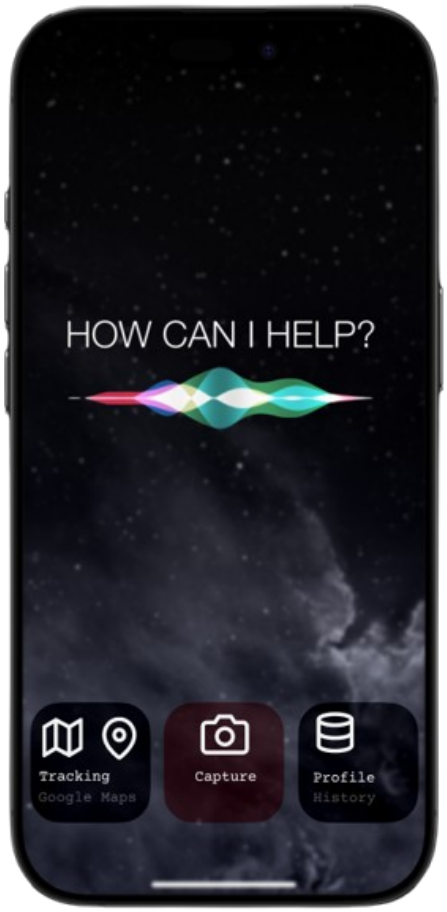


Hình 5.5: Kiến trúc ứng dụng `VisionWalk`

6 Giao diện và Chức năng

§6.1 Giao diện người dùng

Dựa trên ý tưởng đã phác họa ở phần trước, nhóm đã thiết kế được một giao diện người dùng cơ bản cho thiết bị di động như bên dưới.



Hình 6.1: Thiết kế giao diện ứng dụng `VisionWalk`

Vì là ứng dụng dành cho người khiếm thị nên việc thiết kế giao diện cũng phải tuân theo những yêu cầu cơ bản như:

- Thiết kế đơn giản: Giao diện với các biểu tượng lớn, đặt ở những vị trí dễ nhớ để người dùng khiếm thị có thể dễ dàng tiếp cận.
- Chú trọng đến tính khả dụng: Sự tối giản trong thiết kế có thể giúp hạn chế sự nhầm lẫn và cải thiện trải nghiệm người dùng, giúp họ dễ dàng điều hướng qua các tính năng mà không cần quá nhiều thao tác phức tạp.

§6.2 Chức năng

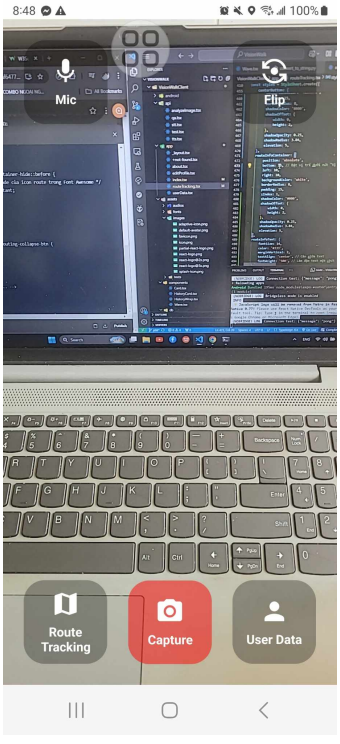
Ngoài chức năng chính là thông qua hình ảnh được gửi lên từ người dùng chụp để phản hồi lại môi trường xung quanh họ, ứng dụng của nhóm còn hướng đến việc tiện nghi hóa những nhu cầu khác của người dùng khiếm thị. Do đó chúng ta có thêm một vài chức năng hữu dụng khác, chẳng hạn như:

- Hỗ trợ `Q&A` giữa người với hệ thống bằng giọng nói
- Hỗ trợ tìm đường và định vị người dùng thông qua chức năng `Tracking` và hỗ trợ người mù trong việc tìm đường đến địa điểm họ muốn đến.
- Theo dõi lịch sử những lần phản hồi của hệ thống đến người dùng và thông tin cơ bản của họ thông qua `Profile` và lưu những thông tin cơ bản của người dùng.

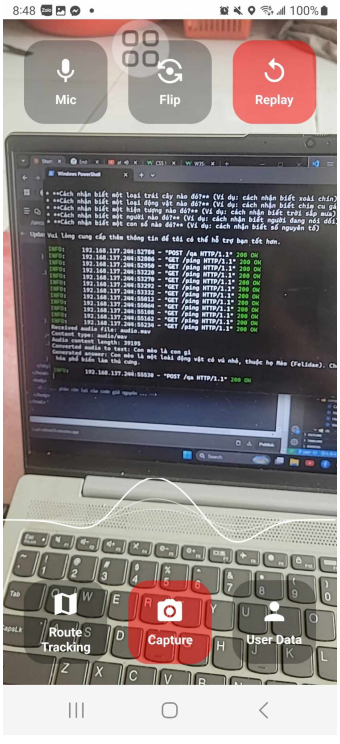
Ta sẽ đi vào chi tiết hơn những tính năng vừa rồi như sau:

① Hỗ trợ `Q&A` giữa người với hệ thống bằng giọng nói:

Ở đây ta sử dụng những mô hình học sâu của **Google Cloud**: **Speech To Text** để chuyển giọng nói của người dùng thành văn bản, **Text to Speech** để chuyển văn bản trả lời của mô hình thành file audio và mô hình ngôn ngữ lớn **Gemini-1.5** được sử dụng để tìm ra câu trả lời cho người dùng.



Hình 6.2: Chưa sử dụng `Q&A`



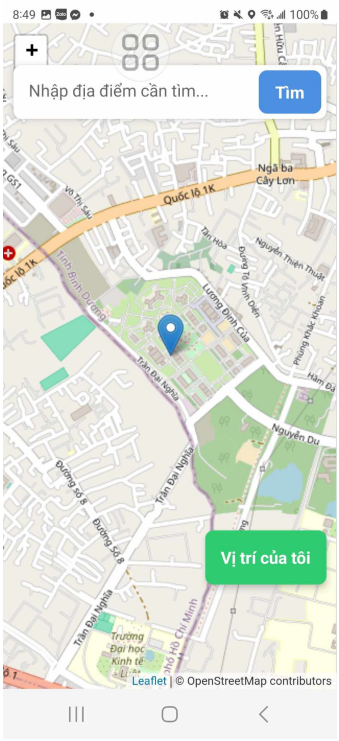
Hình 6.3: Khi sử dụng `Q&A`

Cách dùng rất đơn giản, ta chỉ cần ấn nút **Mic** thì nó sẽ bắt đầu ghi âm (Có thể sẽ xin quyền truy cập mic từ người dùng). Nếu người dùng ấn nút stop hoặc người dùng không nói gì trong 3s thì mic sẽ tự động tắt và sẽ gửi API lên server để server thực hiện chuyển đổi âm thanh sang văn bản, sau đó dùng văn bản đó để gửi lên **gememi-1.5** để tìm câu

trả lời và trả lại cho server văn bản. Cuối cùng ta sẽ chuyển đoạn văn bản đó thành âm thanh để gửi lên cho người dùng để phát audio cho người dùng nghe.
Khi phát audio thì ta có kết hợp sóng siri vào audio visualization (Hiện đang phát triển và chưa hoàn thiện)

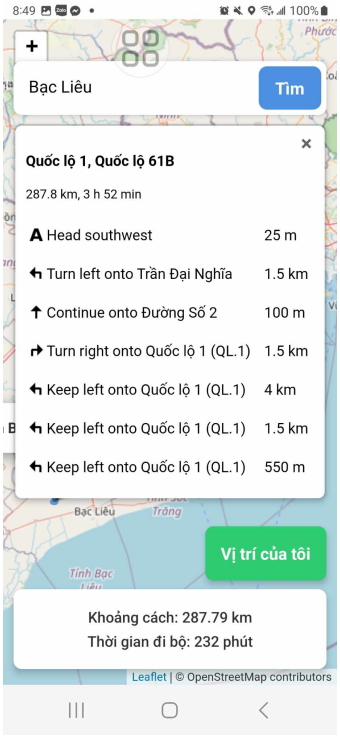
② Chức năng `Tracking`:

Ở đây ta sử dụng open API của **OpenStreetMap**, nó là một API mã nguồn mở miễn phí liên quan đến các dịch vụ của Map như xác định vị trí hiện tại, tìm đường đi ngắn nhất từ vị trí hiện tại của ta đến vị trí mà ta tìm kiếm.
Dựa vào trên thì ta có thể xác định được vị trí của ta trên bản đồ (Hiện chưa phát triển chức năng phát audio nói cho người dùng biết vị trí hiện tại nhưng sẽ phát triển trong tương lai)

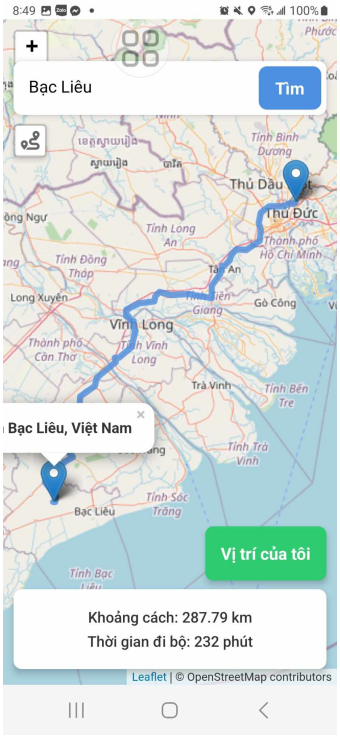


Hình 6.4: Vị trí của người dùng khi vào **Route Tracking**

Khi ta tìm kiếm 1 vị trí nào đó thì ứng dụng sẽ gửi yêu cầu đến **OpenStreetMap** để tìm kiếm vị trí giống nhất so với vị trí ta tìm kiếm. Khi tìm kiếm thành công thì ứng dụng sẽ hiện bảng chi tiết về đường đi để ta có thể dựa vào đó để di chuyển (Trong tương lai sẽ có kết hợp với audio để người khiếm thị có thể sử dụng) và vẽ đường đi lên trên map cho người dùng.



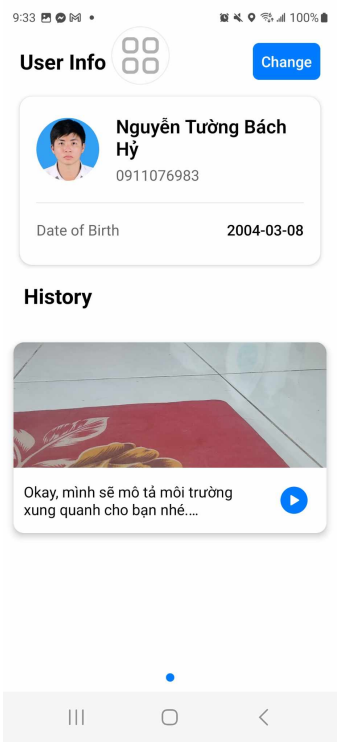
Hình 6.5: Tìm vị trí cần đến (Có bảng chỉ dẫn chi tiết)



Hình 6.6: Tìm vị trí cần đến (Không có bảng chỉ dẫn chi tiết)

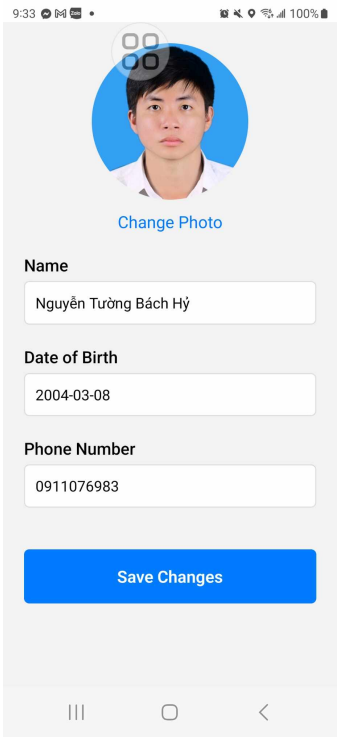
(3) Chức năng `Profile`:

Ở đây ta sử dụng **SQLite** có sẵn trong điện thoại để làm chức năng này (Để giảm chi phí duy trì server). Khi người dùng chụp màn hình gửi lên server để server xử lý ảnh thì server sẽ luôn trả về 2 giá trị là **audio** và **text**, **audio** thì ta lưu vô cơ sở dữ liệu dưới dạng **base64** để khi người dùng muốn nghe lại thì có thể vô phần **User Data** nghe lại. Còn phần text thì ta lưu nó trong cơ sở dữ liệu để người dùng có thể kiểm tra lại những câu trả lời từ mô hình ngôn ngữ lớn (Chức năng này làm để cho người quản lý hội sử dụng để kiểm tra lịch trình của người mù). Và có chức năng cập nhật Profile cho người dùng để họ có thể ghi thông tin của họ lên phòng những trường hợp họ rơi điện thoại hoặc không nhớ thông tin của họ có thể nhìn vào **Profile** để liên hệ.



Hình 6.7: Thông tin người dùng và lịch sử chụp ảnh khi vào **UserData**

Ngoài ra ta còn có thể điều chỉnh profile (Chức năng để người quản lý họ có thể giúp người mù chỉnh lại profile cho họ)



Hình 6.8: Chỉnh sửa thông tin người dùng

III

KẾT QUẢ THU HOẠCH

7 Kết quả thu hoạch

§7.1 Kết quả demo VisionWalk

(1) Khả năng nhận diện: Nhóm đã huấn luyện qua một bộ dữ liệu các hình ảnh tình huống giao thông trên đường bộ, nhưng chủ yếu là vỉa hè (vì đang phục vụ cho đối tượng người khiếm thị). Qua đó nhận được kết quả:

- Bộ lọc hình ảnh thu thập từ người dùng xử lý chính xác >90% các trường hợp ảnh.
- Độ chính xác cao trong việc phát hiện biển báo giao thông (>90%).
- Nhận diện tốt đối với các vật cản ở khoảng cách 15m.

Hình 7.1: Một ví dụ tình huống có biển báo và vật cản

Két quả được trả ra bằng giọng nói AI có nội dung như sau: 'Okay, mình sẽ mô tả môi trường xung quanh cho bạn nhé. Ngay trước mặt chúng ta, cách khoảng 2 mét, có một cột biển báo giao thông hình tròn, màu đỏ trắng, cấm xe hơi. Cạnh đó, bên trái một chút, cũng cách khoảng 2 mét, có một cột biển báo taxi. Hai cột này đứng khá sát vỉa hè nên bạn cẩn thận đừng va phải nhé. Mặt đường ở đây khá bằng phẳng, có vẻ như là đường nhựa, nhưng vẫn nên bước đi cẩn thận. Ngay phía trước, cách khoảng 5-7 mét, có một vài xe máy đang di chuyển, bạn nghe thấy tiếng động cơ không? Hãy đợi chúng đi qua rồi mình mới sang đường nhé. Bên phải, cách chúng ta khoảng 3 mét, là vỉa hè, khá rộng và bằng phẳng. Còn bên trái, cách tầm 10 mét, có một số xe máy đang đỗ, bạn cẩn thận nhé. Dọc

`theo vỉa hè bên trái là các cửa hàng, tường nhà có thể hơi lồi lõm một chút, mình đi sát vào trong một chút để tránh va vào nhé. Hiện tại không thấy có vật cản nào như đống cát, gạch đá hay rãnh nước. Mặt đường cũng có vẻ khô ráo, không trơn trượt. Tuy nhiên, vẫn cần cẩn thận với các phương tiện giao thông di chuyển trên đường. Mình sẽ thông báo cho bạn khi có bất kỳ thay đổi nào nhé.'`

(2) Hiệu suất:

- Thời gian phản hồi chưa được nhanh (khoảng $5s$)
- Tối ưu được bộ nhớ và pin khi dùng ứng dụng
- Hoạt động ổn định trên cả `Android` và `iOS`

(3) Trải nghiệm người dùng:

- Giao diện dễ sử dụng cho người khiếm thị
- Cảnh báo âm thanh rõ ràng, không gây nhiễu

§7.2 Những hạn chế và ý tưởng khắc phục

(1) Tốc độ phản hồi còn chậm: cụ thể là mất tối thiểu $5s$ cho một phản hồi. Điều này nhóm nhận thấy chưa thể là một khoảng chấp nhận được cho sản phẩm này.

❒ Giải pháp:

- **Nén dữ liệu**: Nén dữ liệu phản hồi (như `JSON`, ảnh, video) bằng `gzip` hoặc `Brotli` trước khi gửi về `client` để giảm kích thước và thời gian truyền tải.
- **Tối ưu hóa server**: Cải thiện tốc độ xử lý trên `server` khi xử lý `Speech-to-Text` hay `Text-to-Speech`. Tối ưu hóa `pipeline` xử lý `AI`, giảm số bước xử lý không cần thiết.
- **Batch Processing**: Gom các yêu cầu xử lý nhỏ thành một nhóm hoặc xử lý bất đồng bộ với các tác vụ không cần phản hồi ngay.

(2) Ứng dụng thực tiễn: Việc sử dụng ứng dụng trên thiết bị di động mà không có phần cứng hỗ trợ nào cũng mang lại một số khó khăn. Cụ thể là mô hình được huấn luyện tốt thế nào thì những yếu tố ngoại cảnh như nhiệt độ, độ ẩm, ... cũng không thể mô tả chính xác.

Đặc biệt, trong việc xác định khoảng cách tới các vật cản thì ứng dụng chỉ có thể trả lại phản hồi khoảng cách ước lượng thông qua `Gemini-1.5` mà không được huấn luyện qua các mô hình xác định khoảng cách. Điều đó có thể ảnh hưởng lớn đến kết quả phản hồi.

❒ Giải pháp:

- **Áp dụng các mô hình xác định khoảng cách:** ví dụ như `MonoDepth2` để dự đoán chiều sâu từ camera đơn (`RGB`), hay `Stereo Depth Estimation` để sử dụng camera kép nếu có.

- **Dùng thêm phần cứng là các cảm biến**: Các cảm biến (`sensor`) có thể là Cảm biến siêu âm `HC-SR04` (nhỏ gọn $45 \times 20 \times 15mm$), hay `IMU (MPU6050)` để theo dõi hướng di chuyển của người dùng.

§7.3 Mục tiêu và kết quả mong đợi

(1) Mục tiêu: Ứng dụng `VisionWalk` nhằm mục tiêu hỗ trợ người khiếm thị trong việc di chuyển an toàn và tự tin hơn trong môi trường giao thông. Cụ thể, ứng dụng cung cấp các cảnh báo thời gian thực về biển báo giao thông, vật cản và các nguy hiểm tiềm ẩn trên đường, giúp người khiếm thị nhận thức và tránh được các chướng ngại vật, như xe cộ, vật cản hoặc khu vực không an toàn. Ngoài ra, ứng dụng cũng giúp người dùng giảm sự phụ thuộc vào người đi kèm hoặc các thiết bị trợ giúp truyền thống như gậy dò đường, hỗ trợ họ tự tìm đường, nhận diện không gian, và đưa ra quyết định khi tham gia giao thông. Mục tiêu lâu dài là tăng cường an toàn cho người khiếm thị, đặc biệt ở các khu vực đông đúc hoặc có nguy cơ cao, và giúp họ hòa nhập tốt hơn vào xã hội thông qua việc nâng cao khả năng di chuyển độc lập.

(2) Kết quả mong đợi: Kết quả của ứng dụng `VisionWalk` là khả năng phát hiện chính xác hơn 90% các biển báo giao thông và vật cản thường gặp, bao gồm các biển báo cấm, biển chỉ dẫn, cảnh báo nguy hiểm và các vật cản tĩnh hoặc di động như xe cộ, người đi bộ khác và động vật. Ứng dụng sẽ cung cấp cảnh báo thời gian thực với độ trễ dưới 2 giây, giúp người sử dụng có thể phản ứng kịp thời với các nguy hiểm. Ứng dụng được thiết kế để hoạt động tốt trong mọi điều kiện môi trường, từ ban ngày đến ban đêm và trong các điều kiện thời tiết xấu. Ngoài ra, ứng dụng sẽ được tối ưu hóa để hoạt động mượt mà trên các thiết bị di động phổ thông, không yêu cầu phần cứng đặc biệt, giúp tiết kiệm chi phí và mở rộng khả năng tiếp cận cho người khiếm thị.