BỘ KHOA HỌC VÀ CÔNG NGHỆ
CỤC SỞ HỮU TRÍ TUỆ

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

BẰNG ĐỘC QUYỀN SÁNG CHẾ
Số: 53619

Tên sáng chế: HỆ THỐNG CHUYỂN ĐỔI THỦ NGỮ SANG VĂN BẢN VÀ GIỌNG NÓI TRONG THỜI GIAN THỰC

Chủ Bằng độc quyền: Trường Đại học Bách Khoa - Đại học Quốc gia thành phố Hồ Chí Minh (VN)
268 Lý Thường Kiệt, phường 14, quận 10, thành phố Hồ Chí Minh

Tác giả: 1. Nguyễn Quang Đức (VN)
Trường Đại học Bách Khoa, 268 Lý Thường Kiệt, phường 14, quận 10, thành phố Hồ Chí Minh
2. (Danh sách kèm theo)

Số đơn: 1-2022-00961

Ngày nộp đơn: 17/02/2022

Số điểm yêu cầu bảo hộ: 01 **Số trang mô tả:** 33

Cấp theo Quyết định số: 241768/QĐ-SHTT.IP, ngày: 21/10/2025

Có hiệu lực từ ngày cấp đến hết 20 năm tính từ ngày nộp đơn (Hiệu lực bảo hộ cần duy trì hàng năm).

C.H.X.H.C.N VIỆT NAM
CỤC SỞ HỮU TRÍ TUỆ
INTELLECTUAL PROPERTY OFFICE OF VIET NAM
BỘ KHOA HỌC VÀ CÔNG NGHỆ

VN 1-005

KT. CỤC TRƯỞNG
PHÓ CỤC TRƯỞNG

CỘNG HÒA X.H.C.N VIỆT NAM
CỤC SỞ HỮU TRÍ TUỆ
BỘ KHOA HỌC VÀ CÔNG NGHỆ

Lê Huy Anh

BỘ KHOA HỌC VÀ CÔNG NGHỆ
CỤC SỞ HỮU TRÍ TUỆ

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

BẰNG ĐỘC QUYỀN SÁNG CHẾ
Số: 53619

Tên sáng chế: HỆ THỐNG CHUYỂN ĐỔI THỦ NGỮ SANG VĂN BẢN VÀ GIỌNG NÓI TRONG THỜI GIAN THỰC

Chủ Bằng độc quyền: Trường Đại học Bách Khoa - Đại học Quốc gia thành phố Hồ Chí Minh (VN)
268 Lý Thường Kiệt, phường 14, quận 10, thành phố Hồ Chí Minh

Tác giả: 1. Nguyễn Quang Đức (VN)
Trường Đại học Bách Khoa, 268 Lý Thường Kiệt, phường 14, quận 10, thành phố Hồ Chí Minh
2. (Danh sách kèm theo)

Số đơn: 1-2022-00961

Ngày nộp đơn: 17/02/2022

Số điểm yêu cầu bảo hộ: 01 **Số trang mô tả:** 33

Cấp theo Quyết định số: 241768/QĐ-SHTT.IP, ngày: 21/10/2025

Có hiệu lực từ ngày cấp đến hết 20 năm tính từ ngày nộp đơn (Hiệu lực bảo hộ cần duy trì hàng năm).

KT. CỤC TRƯỞNG
PHÓ CỤC TRƯỞNG

CỤC SỞ HỮU TRÍ TUỆ

VN 1-0053619

Lê Huy Anh

BẰNG ĐỘC QUYỀN SÁNG CHẾ SỐ: 53619

Tác giả khác:

2. MAI THANH PHONG (VN)
268 Lý Thường Kiệt, phường 14, quận 10, thành phố Hồ Chí Minh
3. QUẢN THÀNH THƠ (VN)
268 Lý Thường Kiệt, phường 14, quận 10, thành phố Hồ Chí Minh
4. VÕ THANH HẰNG (VN)
268 Lý Thường Kiệt, phường 14, quận 10, thành phố Hồ Chí Minh
5. BÙI NGÔ HOÀNG LONG (VN)
268 Lý Thường Kiệt, phường 14, quận 10, thành phố Hồ Chí Minh
6. PHAN QUỐC LONG (VN)
268 Lý Thường Kiệt, phường 14, quận 10, thành phố Hồ Chí Minh
7. LÊ ĐỖ THANH BÌNH (VN)
268 Lý Thường Kiệt, phường 14, quận 10, thành phố Hồ Chí Minh
8. NGUYỄN THÀNH LƯU (VN)
268 Lý Thường Kiệt, phường 14, quận 10, thành phố Hồ Chí Minh

(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)**
CỤC SỞ HỮU TRÍ TUỆ

(11) **1-0053619**

(51)[2021.01] **G01L 13/00**

(13) **B**

(21) 1-2022-00961

(22) 17/02/2022

(45) 25/11/2025 452

(43) 25/11/2022 416A

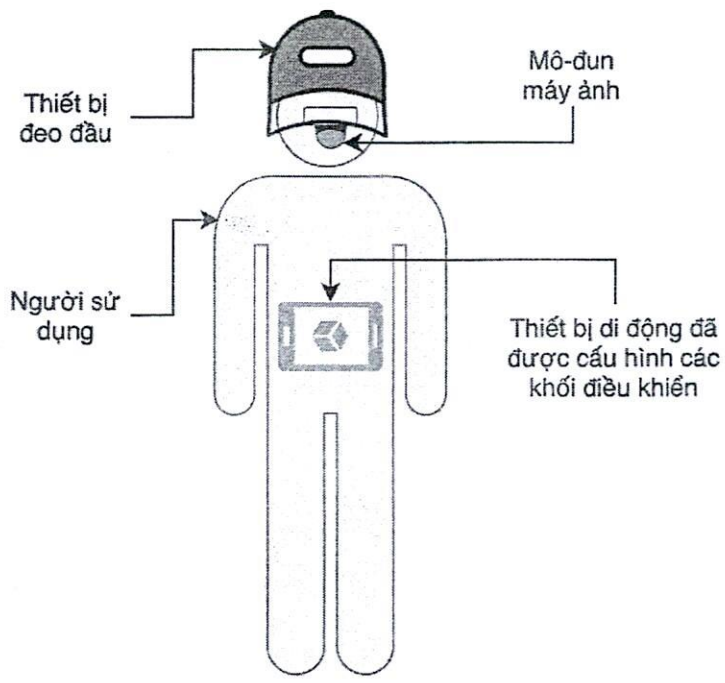
(73) Trường Đại học Bách Khoa - Đại học Quốc gia thành phố Hồ Chí Minh (VN)
268 Lý Thường Kiệt, phường 14, quận 10, thành phố Hồ Chí Minh

(72) Nguyễn Quang Đức (VN); Mai Thanh Phong (VN); Quản Thành Thơ (VN); Võ Thanh Hằng (VN); Bùi Ngô Hoàng Long (VN); Phan Quốc Long (VN); Lê Đỗ Thanh Bình (VN); Nguyễn Thành Lưu (VN).

(54) HỆ THỐNG CHUYỂN ĐỔI THỦ NGỮ SANG VĂN BẢN VÀ GIỌNG NÓI TRONG THỜI GIAN THỰC

(21) 1-2022-00961

(57) Sáng chế là một hệ thống giúp chuyển đổi thủ ngữ hay ngôn ngữ ký hiệu Việt Nam sang văn bản và giọng nói trong thời gian thực, hỗ trợ những người khiếm thanh giao tiếp với người khác dễ dàng hơn. Sáng chế bao gồm bốn thành phần chính là mô-đun máy ảnh được thiết kế nhỏ gọn được gắn vào các thiết bị đeo đầu và ba khối điều khiển việc thu nhận hình ảnh thủ ngữ, chuyển đổi hình ảnh thủ ngữ sang văn bản và chuyển đổi văn bản đã nhận diện sang giọng nói được cấu hình trên thiết bị di động. Mô-đun máy ảnh được cấu hình để ghi nhận hình ảnh thực hiện ngôn ngữ ký hiệu của người khiếm thanh, khiếm thính. Khối điều khiển thu nhận hình ảnh sẽ quản lý việc thu nhận và truyền hình ảnh đã thu nhận về thiết bị di động thông qua mạng Wifi cục bộ. Khối điều khiển việc chuyển đổi thủ ngữ sang văn bản được cấu hình trên thiết bị di động; thông qua các quá trình phân tích, xử lý và tính toán dựa trên các mô hình học sâu được chúng tôi thiết kế để nhanh chóng đưa ra văn bản tương ứng với hình ảnh của ngôn ngữ ký hiệu đã thu thập được. Cuối cùng, khối điều khiển việc chuyển đổi văn bản sang giọng nói tiến hành xử lý văn bản đã nhận diện được thành giọng nói và phát ra các thiết bị nghe tương thích.



Hình 1

Lĩnh vực kỹ thuật được đề cập

Hệ thống chuyển đổi thủ ngữ sang văn bản và giọng nói trong thời gian thực thuộc lĩnh vực công nghệ thông tin đề cập đến việc chuyển đổi ngôn ngữ ký hiệu Việt Nam sang văn bản và lời nói tương ứng trong thời gian thực giúp cho người khiếm thanh, khiếm thính dễ dàng giao tiếp với mọi người. Mục đích của thiết bị là phá bỏ rào cản giao tiếp của những người khiếm thanh, khiếm thính, giúp họ dễ dàng trao đổi, chung sống, hòa nhập với cộng đồng.

Tình trạng kỹ thuật của sáng chế

Một nghiên cứu tiến hành tại Hà Nội, Thái Nguyên, Nghệ An, Thành phố Hồ Chí Minh và thành phố Đà Nẵng từ tháng 8 đến tháng 11 năm 2017 với 574 người khuyết tật chỉ ra rằng những người khuyết tật nói chung, người khiếm thanh, khiếm thính nói riêng rất ít được tiếp cận các dịch vụ y tế, giáo dục, tỷ lệ thất nghiệp cao và bị hoặc cảm nhận bị kỳ thị, ảnh hưởng đến đời sống vật chất, tinh thần [1]. Theo số liệu thống kê năm 2019, nước ta có khoảng 1,5 đến 2 triệu người người khiếm thanh, khiếm thính và người khiếm thính, tuy nhiên số lượng phiên dịch ngôn ngữ ký hiệu chuyên nghiệp lại chỉ có khoảng hơn 10 người, đây là một sự chênh lệch quá lớn [2].

Hiện nay, những người khiếm thanh, khiếm thính hiện vẫn đang sử dụng thủ ngữ (ngôn ngữ ký hiệu Việt Nam) để giao tiếp trong cuộc sống. Tuy nhiên, hệ thống ngôn ngữ ký hiệu này có những điểm hạn chế như: người tương tác buộc phải biết ngôn ngữ ký hiệu, thời gian tương tác lâu, dễ bị hiểu nhầm ý diễn đạt, nhu cầu sử dụng lớn nhưng tại Việt Nam còn khá ít các trung tâm dạy ngôn ngữ ký hiệu, ngành thông dịch viên ngôn ngữ ký hiệu cũng ít được giảng dạy tại các trường đại học.

Chính vì những điều trên, việc phát triển một thiết bị hỗ trợ người khiếm thanh, khiếm thính thuận tiện hơn trong việc giao tiếp với người xung quanh là cần thiết và có ảnh hưởng lớn đến xã hội, đặc biệt là đối tượng người khiếm thanh, khiếm thính. Những

thiết bị, công cụ hỗ trợ được những người khiếm thanh, khiếm thính sử dụng nhiều nhất là các giấy, bảng viết chữ, phần mềm đánh văn bản trên các thiết bị di động. Những giải pháp này tuy có thể giải quyết vấn đề về giao tiếp, tuy nhiên lại không thuận tiện và mất nhiều thời gian. Trong khi đó, thủ ngữ hay ngôn ngữ ký hiệu là một giải pháp giúp giao tiếp hiệu quả giữa những người khiếm thanh, khiếm thính và mọi người. Tuy vậy hạn chế của phương pháp giao tiếp này là số lượng người biết ngôn ngữ ký hiệu không nhiều, do đó việc giao tiếp bằng ngôn ngữ ký hiệu sẽ không hiệu quả.

Những năm qua, nghiên cứu về việc chuyển đổi ngôn ngữ ký hiệu sang hình thái phổ biến (văn bản hoặc giọng nói) được quan tâm nhiều hơn. Trong phạm vi cả nước, mô hình găng tay chuyển đổi thủ ngữ của nhóm nghiên cứu từ Đại học Công nghiệp Hà Nội công bố tháng 09 năm 2019 [4] là một thiết bị có tính ứng dụng cao, tuy nhiên thiết bị vẫn còn gặp khó khăn khi ứng dụng vào thực tế do chưa tối ưu hoá được kích thước của sản phẩm. Trong phạm vi toàn thế giới, các nghiên cứu về sử dụng máy ảnh ghi hình ảnh ngôn ngữ ký hiệu và chuyển đổi trong thời gian thực đạt được nhiều thành tựu, có thể kể đến như các nghiên cứu ở [5] và [6]. Tuy nhiên, các nghiên cứu này chỉ dừng lại ở mức xác định các chữ cái hoặc số đơn lẻ hoặc so trùng chuỗi ký hiệu bàn tay. Vì vậy, những nghiên cứu này không phù hợp với ngôn ngữ ký hiệu Việt Nam do trong ngôn ngữ ký hiệu tiếng Việt có nhiều ký hiệu giống nhau nhưng do đặt ở những vị trí khác nhau tạo ra những từ khác nhau. Ngoài ra, những nghiên cứu kể trên đòi hỏi góc nhìn của máy ảnh quay ngôn ngữ ký hiệu phải đặt tại vị trí chính diện của người thực hiện ngôn ngữ ký hiệu, điều này làm giảm khả năng ứng dụng khi người khiếm thanh, khiếm thính giao tiếp với người khác ở một không gian bất kỳ.

"Hệ thống chuyển đổi thủ ngữ sang văn bản và giọng nói trong thời gian thực" của chúng tôi được thiết kế riêng, phù hợp với đặc điểm của ngôn ngữ ký hiệu Việt Nam. Cụ thể hơn, hệ thống của chúng tôi thiết kế bao gồm mô-đun máy ảnh để ghi hình ngôn ngữ ký hiệu và các khối điều khiển được cấu hình trên thiết bị di động để chuyển đổi hình ảnh ngôn ngữ ký hiệu thành văn bản và giọng nói. Thiết kế của chúng tôi sử dụng vị trí đặt mô-đun máy ảnh chiếu xuống, do đó tạo sự thuận tiện hơn cho người khiếm thanh, khiếm

thính khi sử dụng (Họ có thể gắn kết mô-đun máy ảnh lên nón hoặc các vật dụng khác đặt trên đầu).

Bản chất kỹ thuật của sáng chế

Bản chất của sáng chế là một hệ thống chuyển đổi hình ảnh ngôn ngữ ký hiệu sang văn bản và giọng nói bao gồm các thành phần như sau: Mô-đun máy ảnh được gắn trên thiết bị đeo đầu và được cấu hình để ghi nhận hình ảnh ngôn ngữ ký hiệu; Khối điều khiển được tạo cấu hình để điều khiển quá trình thu nhận hình ảnh ngôn ngữ ký hiệu từ mô-đun máy ảnh và gửi đến thiết bị di động thông qua giao thức WebSocket trên mạng Wifi cục bộ được cấu hình trên thiết bị di động; Khối điều khiển được tạo cấu hình để điều khiển quá trình chuyển đổi chuỗi hình ảnh ngôn ngữ ký hiệu sang văn bản trên thiết bị di động bao gồm 7 bước: Thu nhận chuỗi hình ảnh được truyền về từ mô-đun máy ảnh thông qua giao thức WebSocket và mạng Wifi cục bộ; Trích xuất thông tin vị trí các điểm mốc của các đốt ngón tay trên từng hình ảnh bằng mô hình MediaPipe; Phân tích động tác của bàn tay trên từng hình ảnh dựa vào thông tin vị trí của các điểm mốc bằng mô hình học sâu; Tính toán hướng chỉ của các ngón tay trên từng hình ảnh trên các điểm mốc của các đốt ngón tay; Tính toán vị trí tương đối của bàn tay trên cơ thể trên từng hình ảnh dựa vào vị trí và độ lớn của hình ảnh bàn tay; Tổng hợp thông tin về chuỗi các hành động, hướng chỉ và vị trí của bàn tay và so khớp với dữ liệu các động tác ngôn ngữ ký hiệu đã được định nghĩa sẵn bằng việc vectơ hóa các trạng thái tay và thực thi giải thuật Beam Search kết hợp với Connectionist Temporal Classification trên ma trận xác suất của trạng thái tay; Truy vấn văn bản tương ứng với động tác ngôn ngữ ký hiệu đã được định nghĩa sẵn; và Khối điều khiển được tạo cấu hình để điều khiển quá trình chuyển đổi văn bản đã nhận diện được sang giọng nói và phát ra các thiết bị nghe tương thích.

Mô tả vắn tắt các hình vẽ

Hình 1 mô tả cấu trúc tổng quan của toàn bộ hệ thống bao gồm hai thành phần chính là mô-đun máy ảnh được đặt trên thiết bị đeo đầu và thiết bị di động được cài đặt các khối điều khiển để chuyển đổi ngôn ngữ ký hiệu sang văn bản và giọng nói.

Hình 2 mô tả các ảnh chụp màn hình của giao diện điều khiển quá trình chuyển đổi ngôn ngữ ký hiệu sang văn bản và giọng nói trên thiết bị di động.

Hình 3 mô tả sơ đồ kết nối các linh kiện điện tử có trong mô-đun máy ảnh bao gồm mạch thu phát Wifi ESP32-CAM, máy ảnh OV2640, mạch sạc pin và pin LiPo dung lượng 2200mAh điện áp 3.7V.

Hình 4 mô tả thiết kế của hộp nhựa cấu thành mô-đun máy ảnh.

Hình 5 mô tả cách thức kết nối giữa mô-đun máy ảnh và khối điều khiển thu nhận hình ảnh trên thiết bị di động thông qua mạng Wifi cục bộ được cấu hình trên thiết bị di động.

Hình 6 mô tả một ví dụ về chuỗi hình ảnh mà mô-đun máy ảnh sẽ thu nhận.

Hình 7 mô tả khái quát toàn bộ quá trình chuyển đổi hình ảnh ngôn ngữ ký hiệu thành văn bản.

Hình 8 mô tả kiến trúc của mô hình MediaPipe dùng để trích xuất các điểm mốc trên bàn tay.

Hình 9 mô tả kiến trúc mô hình học sâu dùng để nhận diện động tác tay từ vị trí của các điểm mốc.

Hình 10 mô tả khái quát quy trình xác định hướng chỉ của bàn tay.

Hình 11 mô tả khái quát quy trình xác định vị trí tương đối của bàn tay trên cơ thể.

Hình 12 mô tả các vị trí trên cơ thể tương ứng với từng khu vực trên hình ảnh thu được từ mô-đun máy ảnh được định nghĩa trước.

Hình 13 mô tả các thành phần cấu thành nên một trạng thái tay bao gồm động tác, hướng chỉ và vị trí tương đối của tay.

Hình 14 mô tả khái quát bước so khớp các trạng thái tay với các động tác ngôn ngữ ký hiệu đã được định nghĩa sẵn.

Hình 15 mô tả một ví dụ về phương pháp Beam Search kết hợp với Connectionist Temporal Classification để so khớp các trạng thái tay với các động tác ngôn ngữ ký hiệu đã được định nghĩa sẵn.

Hình 16 mô tả khái quát bước chuyển đổi các động tác ngôn ngữ ký hiệu sang từ tương ứng đã được định nghĩa trước.

Hình 17 mô tả quy trình chuyển đổi văn bản đã nhận diện thành giọng nói.

Mô tả chi tiết sáng chế

❖ Tổng quan hệ thống

Hệ thống là bao gồm hai thành phần: mô-đun máy ảnh được gắn trên thiết bị đeo đầu và các khối điều khiển được cấu hình để chuyển đổi ngôn ngữ ký hiệu trên thiết bị di động. Cấu trúc tổng quan của hệ thống được mô tả như Hình 1. Hình 1a mô tả mô-đun máy ảnh được gắn trên thiết bị đeo đầu (nón hoặc đai đeo) và thiết bị di động đã được cấu hình được đặt ở phần trước ngực hoặc trước bụng của người khiếm thanh, khiếm thính. Giao diện tương tác trên thiết bị di động được trình bày ở Hình 2. Trong Hình 2, toàn bộ giao diện bao gồm 6 màn hình chính: 2a) Màn hình chào; 2b) Màn hình đăng nhập; 2c) Màn hình hiển thị văn bản không có giọng nói; 2d) Màn hình hiển thị văn bản và có giọng nói; 2e) Màn hình tra cứu những dữ liệu những động tác ngôn ngữ ký hiệu tương ứng với văn bản đã được định nghĩa trước và 2f) Màn hình hiển thị thông tin về mô tả động tác ngôn ngữ ký hiệu tương ứng với một từ. Một cách cụ thể, các khối điều khiển được cấu hình trên thiết bị di động bao gồm ba khối điều khiển chính được tóm tắt như sau. Khối điều khiển đầu tiên có nhiệm vụ thu nhận hình ảnh ngôn ngữ ký hiệu. Công việc của khối điều khiển này này bao gồm việc điều khiển ghi nhận hình ảnh ngôn ngữ ký hiệu trên mô-đun máy ảnh và việc điều khiển truyền những hình ảnh đã ghi nhận được về khối điều khiển phân tích và chuyển đổi ngôn ngữ ký hiệu trên thiết bị di động thông qua mạng Wifi cục bộ. Khối điều khiển tiếp theo là phân tích chuỗi các hình ảnh ngôn ngữ ký hiệu thành văn bản. Khối điều khiển này bao gồm các giải thuật và các mô hình học sâu phục vụ cho việc chuyển đổi ngôn ngữ ký hiệu thành văn bản được chúng tôi tự thiết kế. Khối điều khiển cuối cùng đảm nhiệm việc điều khiển chuyển đổi văn bản đã nhận diện thành giọng nói và phát giọng nói qua các thiết bị nghe tương thích.

❖ Thiết kế mô-đun máy ảnh

Mô-đun máy ảnh được cấu thành từ nhiều linh kiện điện tử với số lượng cụ thể được liệt kê ở Bảng 1. Trong Bảng 1, các linh kiện điện tử từ 1 đến 5 có sẵn trên thị trường và chúng tôi tiến hành lắp ráp chúng lại theo sơ đồ thể hiện ở Hình 3. Linh kiện thứ 6 là hộp nhựa dùng để chứa các linh kiện điện tử từ 1 đến 5 được chúng tôi tự thiết kế để phù hợp với việc đeo và có khả năng gắn vào các thiết bị đeo đầu. Hộp nhựa có kích thước dài x rộng x cao là *65 x 75 x 33* theo đơn vị milimet. Hình 4 mô tả thiết kế hộp nhựa của chúng tôi.

Bảng 1: Danh mục linh kiện trong mô-đun máy ảnh

STT	Thiết bị	Số lượng
1	Mạch thu phát Wifi ESP32-CAM	1
2	Máy ảnh OV2640	1
3	Pin Lipo dung lượng 2200mAh, điệp áp 3.7V	1
4	Mạch sạc pin	1
5	Dây nối	nhiều
6	Hộp nhựa chứa linh kiện điện tử	1

❖ Khối điều khiển được cấu hình để thu nhận hình ảnh ngôn ngữ ký hiệu

Quá trình thu nhận hình ảnh ngôn ngữ ký hiệu là quá trình chụp và truyền hình ảnh từ mô-đun máy ảnh đến thiết bị di động. Quá trình này được khái quát như Hình 5 bao gồm bước như sau:

- Bước 1: Thiết bị di động sẽ được cấu hình để phát ra một mạng Wifi cục bộ.
- Bước 2: Mô-đun máy ảnh sẽ tự động kết nối vào mạng Wifi cục bộ đã được cấu hình.
- Bước 3: Thiết bị di động cấp một địa chỉ IP cố định cho mô-đun máy ảnh.

- Bước 4: Thiết bị di động tạo kết nối WebSocket đến mô-đun máy ảnh ở địa chỉ IP đã định.
- Bước 5: Mô-đun máy ảnh bắt đầu quay thu thập hình ảnh các động tác ngôn ngữ ký hiệu và gửi liên tục từng ảnh về thiết bị di động. Hình ảnh sau khi nhận được chuyển trực tiếp sang khối điều khiển phân tích và chuyển đổi hình ảnh ngôn ngữ ký hiệu sang văn bản. Hình 6 mô tả ví dụ về chuỗi bốn khung ảnh được chụp trên mô-đun máy ảnh và truyền về thiết bị di động.

❖ Khối điều khiển phân tích chuỗi các hình ảnh ngôn ngữ ký hiệu thành văn bản

Quá trình phân tích chuỗi các hình ảnh ngôn ngữ ký hiệu thành văn bản sẽ nhận đầu vào là các hình ảnh truyền về từ mô-đun máy ảnh, tiến hành xử lý, phân tích và tính toán để đưa ra văn bản tương ứng với các hình ảnh đầu vào. Hình 7 thể hiện quá trình này một cách tổng quát. Trong đó: Hàng chờ hình ảnh là hàng chờ chứa chuỗi hình ảnh nhận về từ mô-đun máy ảnh; Với mỗi hình ảnh thuộc hàng chờ, nó được thực hiện trích xuất thông tin vị trí các điểm mốc trên bàn tay; sau đó dùng thông tin các điểm mốc để phân tích động tác, tính toán hướng chỉ và tính toán vị trí tương đối của bàn tay với cơ thể; các thông tin về động tác, hướng chỉ và vị trí được tổng hợp lại, so khớp với cơ sở dữ liệu ngôn ngữ ký hiệu và truy xuất từ tương ứng; cuối cùng các từ này được gửi đến khối điều khiển chuyển đổi văn bản sang giọng nói. Cụ thể, ở khối điều khiển này, hình ảnh thủ ngữ sẽ được thực hiện xử lý thông qua 7 bước như sau:

Bước 1: Nhận chuỗi hình ảnh được truyền về từ mô-đun máy ảnh và đưa vào hàng đợi hình ảnh.

Sau khi nhận được hình ảnh gửi về từ mô-đun máy ảnh, chúng tôi tiến hành lưu trữ vào chúng vào hàng đợi hình ảnh. Hàng đợi hình ảnh bao gồm nhiều hình ảnh được lưu trữ theo thứ tự chúng được nhận. Các hình ảnh trong hàng đợi được lần lượt lấy ra để xử lý cho bước tiếp theo theo thứ tự cái nào đưa vào hàng đợi trước thì sẽ được lấy ra trước. Trong trường hợp hàng đợi hình ảnh đã đầy, chúng tôi tiến hành xóa những hình ảnh cũ nhất.

Bước 2: Trích xuất thông tin vị trí các điểm mốc của các đốt ngón tay trên từng hình ảnh.

Để trích xuất thông tin vị trí các điểm mốc của các đốt ngón tay, chúng tôi sử dụng mô hình học máy MediaPipe [7]. Kiến trúc của mô hình này được mô tả như Hình 8. Đầu vào của mô hình này là một chuỗi hình ảnh (flipped_input_video) có tối đa hai bàn tay, đầu ra của mô hình là 21 điểm mốc trên mỗi bàn tay (hand_rect_from_palm_dettections). Giới hạn của hình ảnh đầu vào là hình ảnh có 3 kênh màu Đỏ, Lục, Lam và có độ phân giải tối đa là *1280x720* pixel, tương ứng với ma trận đầu vào có kích thước *1280 x 720 x 3*. Đầu ra với mỗi bàn tay là 21 điểm mốc theo thứ tự như trong Hình 7, mỗi điểm mốc sẽ có ba tọa độ (*x*, *y*, *z*) tương ứng trong không gian ba chiều, tương ứng với ma trận đầu ra có kích thước *21 x 3*. Trước khi đưa hình ảnh vào mô hình MediaPipe này, chúng tôi tiến hành scale ảnh cho vừa kích thước. Đầu tiên, chúng tôi tính toán tỉ lệ scale:

<tỉ lệ scale> = min(1280 / <chiều cao ảnh>, 720 / <chiều rộng ảnh>)

Sau đó, chúng tôi sử dụng phương pháp biến đổi Linear để đưa ảnh về kích thước mới (<chiều rộng cũ> * <tỉ lệ scale>, <chiều cao cũ> * <tỉ lệ scale>). Tiếp đến chúng tôi đưa ảnh đã scale vào mô hình MediaPipe. Cụ thể tính toán trong mô hình này được mô tả ở bài báo [7].

Bước 3: Phân tích động tác của bàn tay trên từng hình ảnh dựa vào thông tin vị trí của các điểm mốc.

Ở bước này, chúng tôi xây dựng một mô hình học sâu bao gồm nhiều lớp mạng để phân tích được động tác của bàn tay đang thực hiện. Đầu vào của mạng học sâu này là ma trận khoảng cách giữa 21 điểm mốc của mỗi bàn tay. ma trận này có kích thước *21 x 21,* trong đó phần ở ở hàng thứ *i* cột thứ *j* được tính bằng công thức sau:

$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$, trong đó x_i, y_i, z_i là toạ độ của điểm mốc thứ *i* và tương tự cho *j*. Sau đó, ma trận khoảng cách được tạo ra sẽ đi qua mạng học sâu gồm 14 lớp tính toán được liệt kê như sau:

- Lớp tính toán 1: Tích chập *3x3* (Conv 3x3) với 32 bộ lọc

Đầu vào là ma trận kích thước *21 x 21 x 1.* Từng ma trận con có kích thước *3 x 3 x 1* được lấy ra và nhân tích chập với ma trận trọng số cùng kích thước và tạo thành ma trận có kích thước *19 x 19 x 1*. Công thức tính tích chập như sau: $a_{mn} =$

$\sum_{j=0}^{2}\sum_{i=0}^{2} w_{ij} d_{(i+m)(j+n)}$, với a_{mn} là vị trí hàng m cột n trên ma trận sau khi tính tích chập, w_{ij} là phần tử hàng i cột j của ma trận trọng số và $d_{(i+m)(j+n)}$ là phần tử ở hàng $i+m$, cột $j+n$ của ma trận khoảng cách. Với 64 bộ lọc, tương ứng với 64 ma trận trọng số *3 x 3 x 1*, chúng tôi có đầu ra cho lớp tính toán 1 là ma trận *19 x 19 x 64*.

- Lớp tính toán 2: Kích hoạt kết quả với hàm *tanh* (Activation *tanh*)

Từng phần tử trên ma trận sau tính toán tích chập được đưa qua hàm *tanh* với công thức:

$b_{mn} = tanh(a_{mn})$.

- Lớp tính toán 3: Chuẩn hoá kết quả về khoảng giá trị 0-1 (BatchNorm)

Với mỗi vị trí hàng i cột j trên ma trận, chúng tôi chuẩn hóa bằng công thức như sau:

$\hat{x}_{ij} = (x_{ij} - \underline{x}_{ij})/\sqrt{\sigma_{ij}^2 + \epsilon}$, trong đó $\hat{x}_{ij}$ là giá trị sau chuẩn hóa, $\underline{x}_{ij}$ là trung bình giá trị, σ_{ij}^2 là phương sai tại hàng i cột j và $\epsilon = 10^{-5}$.

- Lớp tính toán 4: Tích chập *3x3* (Conv 3x3) với 32 bộ lọc

Tính toán tương tự như lớp tính toán 1 với 64 ma trận trọng số có kích thước *3 x 3 x 32*. Đầu ra của lớp này là ma trận có kích thước *17 x 17 x 32*.

- Lớp tính toán 5: Kích hoạt kết quả với hàm *tanh* (Activation *tanh*)

Tính toán tương tự như lớp tính toán 2.

- Lớp tính toán 6: Chuẩn hoá kết quả về khoảng giá trị 0-1 (BatchNorm)

Tính toán tương tự như lớp tính toán 3.

- Lớp tính toán 7: Tích chập *5x5* (Conv 5x5) với 32 bộ lọc

Tính toán tương tự như lớp tính toán 1 với 64 ma trận trọng số có kích thước *5 x 5 x 32* và khoảng cách trượt là 2. Đầu ra của lớp này là ma trận có kích thước *9 x 9 x 32*.

- Lớp tính toán 8: Kích hoạt kết quả với hàm *tanh* (Activation *tanh*)

Tính toán tương tự như lớp tính toán 2.

- Lớp tính toán 9: Chuẩn hoá kết quả về khoảng giá trị 0-1 (BatchNorm)

Tính toán tương tự như lớp tính toán 3.

- Lớp tính toán 10: Làm phẳng ma trận thành vectơ

Ma trận kích thước *9 x 9 x 32* được làm phẳng bằng cách nối các hàng và cột lại với nhau để thành vector 2592 chiều.

- Lớp tính toán 11: Nhân vectơ với ma trận trọng số đầy đủ (Fully-connected)

Đầu vào là vectơ 2592 chiều, sau khi nhân với ma trận trọng số có kích thước *2592 x 128* thì cho ra vectơ đầu ra có số chiều bằng 128. Công thức nhân ma trận như sau: $c = Wx + b$ với W là ma trận trọng số, b là vectơ độ lệch, x là vectơ đầu vào và c là vectơ đầu ra.

- Lớp tính toán 12: Kích hoạt kết quả với hàm *tanh* (Activation *tanh*)

Tính toán tương tự như lớp tính toán 2.

- Lớp tính toán 13: Chuẩn hoá kết quả về khoảng giá trị 0-1 (BatchNorm)

Tính toán tương tự như lớp tính toán 3.

- Lớp tính toán 14: Nhân vectơ với ma trận trọng số đầy đủ (Fully-connected)

Đầu vào là vectơ 128 chiều, sau khi nhân với ma trận trọng số có kích thước *128 x <số lượng động tác tay>* thì cho ra vectơ đầu ra có số chiều bằng *<số lượng động tác tay>*. Công thức nhân ma trận tương tự như lớp tính toán 11.

Kết quả có được sau khi qua 14 lớp tính toán là vectơ có số chiều *<số lượng động tác tay>* thể hiện xác suất của từng động tác mà hình ảnh đầu vào có thể đang thực hiện. Cuối cùng, chúng tôi chọn ra động tác có xác xuất cao nhất làm kết quả trả về. Hình 9 thể hiện kiến trúc mô hình học sâu mà chúng tôi đã thiết kế.

Bước 4: Tính toán hướng chỉ của các ngón tay trên từng hình ảnh nếu động tác của bàn tay có hướng.

Các hướng của bàn tay bao gồm sáu hướng, tức là trái, phải, hướng lên, hướng xuống, hướng trước và hướng sau. Mỗi cử chỉ của bàn tay, kết hợp với các hướng khác nhau, dẫn đến một ý nghĩa khác nhau. Để xác định hướng của bàn tay, chúng tôi sử dụng các điểm mốc đốt ngón tay đã trích xuất ở Bước 2. Chúng tôi thực hiện tính toán khoảng cách giữa cổ tay đến đầu ngón trỏ (điểm mốc 0 và điểm mốc 8 trên Hình 10), có thể được gọi là *vectơ (0, 8)*, sau đó chiếu nó lên các trục Ox, Oy, Oz tương ứng. Chúng tôi sau đó có được hình chiếu của *vectơ (0, 8)* lên ba trục toạ độ lần lượt là p_x, p_y, p_z. Sau đó, chúng tôi tính toán độ dài hình chiếu của ba vectơ p_x, p_y, p_z và so sánh chúng với nhau bằng công thức sau:

$|p_i| = \sqrt{(s_{ix} - e_{ix})^2 + (s_{iy} - e_{jy})^2 + (s_{iz} - e_{jz})^2}$, với s_i là điểm bắt đầu và e_i là điểm kết thúc của p_i và p_i lần lượt là p_x, p_y và p_z.

Cuối cùng, hình chiếu vectơ có độ dài lớn nhất sẽ cho biết bàn tay đang chủ yếu nằm trên trục nào; ngoài ra với chiều từ cổ tay đến đầu ngón trỏ chiếu lên trục tương ứng đó thì ta sẽ biết được bàn tay đó là hướng nào, cụ thể như sau:

- Trục Ox, cùng chiều vectơ đơn vị: Hướng phải (Rightmost)
- Trục Ox, ngược chiều vectơ đơn vị: Hướng trái (Leftmost)
- Trục Oy, cùng chiều vectơ đơn vị: Hướng trước (Farest)
- Trục Oy, ngược chiều vectơ đơn vị: Hướng sau (Nearest)
- Trục Oz, cùng chiều vectơ đơn vị: Hướng lên (Highest)
- Trục Oz, ngược chiều vectơ đơn vị: Hướng xuống (Lowest)

Bước 5: Tính toán vị trí tương đối của bàn tay trên cơ thể trên từng hình ảnh dựa vào vị trí và độ lớn của hình ảnh bàn tay.

Đầu tiên, khung chứa bàn tay trong hình được xác định từ tọa độ của các điểm mốc 0, 5 và 17 trích xuất từ Bước 2 như trong Hình 11. Sau đó, vị trí của khung chứa được so sánh với các vùng trong cơ thể được định nghĩa sẵn (Hình 12) để xác định vị trí tương đối của bàn tay trên cơ thể. Đối với các vùng có đè nhau như trán, ngực, miệng, độ lớn của khung chứa sẽ được dùng để xác định bàn tay thuộc về vùng nào dựa trên được chúng tôi định nghĩa như sau: khung chứa của bàn tay ở vùng trán hay miệng sẽ lớn hơn khi ở vùng ngực vì trán và miệng ở gần máy ảnh hơn. Trong đó nếu đường chéo của khung chứa lớn hơn FOREHEAD_THRESHOLD = 0.5 thì thuộc về phần trán, nếu bé hơn CHEST_THRESHOLD = 0.125 thì thuộc về phần ngực, còn lại thì thuộc về phần miệng. Bảng 2 mô tả vị trí của các cơ quan trên cơ thể được xác định bằng điểm trái dưới và điểm phải trên.

Bảng 2: Vị trí của các cơ quan trên cơ thể được xác định bằng luật. Góc dưới, bên trái khung hình có tọa độ (0,0). Góc trên, bên phải khung hình có tọa độ (1,1)

Vị trí	Trên	Dưới	Trái	Phải
Trán	0.9	0.1	0.3	0.7
Miệng	0.5	0.2	0.35	0.65
Ngực	0.8	0.3	0.4	0.6
Mặt trái	0.2	0.0	0.1	0.35
Mặt phải	0.2	0.0	0.65	0.9
Eo trái	1.0	0.2	0.1	0.3
Eo phải	1.0	0.2	0.7	0.9

Bước 6: Tổng hợp thông tin về chuỗi các hành động, hướng chỉ và vị trí của bàn tay và so khớp với dữ liệu các động tác ngôn ngữ ký hiệu đã được định nghĩa sẵn.

Ở bước này, chúng tôi tổng hợp các trạng thái bàn tay từ ba yếu tố động tác, vị trí và hướng chỉ đã được phân tích và tính toán ở các Bước 3, 4 và 5. Từ đây, mỗi bộ ba phần tử (động tác, vị trí và hướng chỉ) trích xuất được trên một hình ảnh tại một thời điểm sẽ được gọi là trạng thái tay tại thời điểm đó (Hình 13). Khái niệm trạng thái bàn tay này xuất phát từ việc nghiên cứu quá trình xử lý ngôn ngữ tự nhiên, trong đó một từ được cấu tạo bởi nhiều ký tự. Vì vậy, một từ được ghép từ nhiều trạng thái bàn tay. Từ đó, chúng ta sẽ nhận được từ mong muốn lần lượt đi qua các bước xử lý so khớp tiếp theo.

Quá trình so khớp các trạng thái tay và động tác ngôn ngữ ký hiệu bao gồm hai công việc chính được mô tả như ở Hình 14 bao gồm vectơ hóa trạng thái tay và thực thi giải thuật Beam Search kết hợp với Connectionist Temporal Classification (CTC). Để so khớp chuỗi các trạng thái tay với các động tác ngôn ngữ ký hiệu, chúng tôi hiện thực một hàng đợi các trạng thái tay có độ dài bằng 5. Chúng tôi sau đó thực hiện phép biến đổi vectơ hóa (vectorization) để chuyển đổi một hàng đợi gồm 5 trạng thái tay thành một ma trận có kích thước (*5×<số lượng động tác ngôn ngữ ký hiệu>*) làm đầu vào cho bước so khớp (Beam Search + CTC).

Với mỗi trang thái tay trong hàng đợi, chúng tôi tính toán xác suất cho việc nó có thuộc bất kỳ động tác ngôn ngữ ký hiệu nào có sẵn không bằng cách đánh giá độ trùng khớp giữa trang thái tay và động tác ngôn ngữ ký hiệu. Trạng thái tay có động tác, vị trí, hướng chỉ càng giống với động tác ngôn ngữ ký hiệu có sẵn thì điểm sẽ càng cao. Sau đó, chúng tôi dùng hàm *Softmax* với công thức như sau để chuẩn hóa điểm số thành xác suất mỗi trạng thái tay với tất cả động tác ngôn ngữ ký hiệu.

$$f_i(x) = \frac{exp(x_i)}{\sum_j exp(x_j)},$$

trong đó $f_i(x)$ là xác suất của trạng thái tay x thuộc động tác ngôn ngữ ký hiệu i.

Để so khớp với dữ liệu các động tác ngôn ngữ ký hiệu đã được định nghĩa sẵn, chúng tôi thực hiện thuật toán Beam Search để chọn trạng thái tay phù hợp với đầu vào từ cơ sở dữ liệu. Bên cạnh đó, chúng tôi sử dụng mô hình giải mã Connectionist Temporal Classification để loại bỏ trạng thái bàn tay sai hoặc trạng thái bàn tay trùng lặp trước đó, tăng hiệu quả của mô hình. Quá trình so khớp này sẽ tìm ra một tổ hợp các khả năng của trạng thái tay ứng với động tác ngôn ngữ ký hiệu có xác suất cao nhất và lấy đó làm kết quả so khớp trạng thái tay và động tác ngôn ngữ ký hiệu. Một ví dụ minh họa cho quá trình so khớp bằng Beam Search + Connectionist Temporal Classification được mô tả ở Hình 15, trong đó, theo hàng ngang là 5 trạng thái tay từ hàng đợi có màu xanh lục, theo hàng dọc là các động tác ngôn ngữ ký hiệu có sẵn trong cơ sở dữ liệu, và đường đi màu vàng là tổ hợp có xác suất cao nhất và sẽ được trả về như kết quả so khớp.

Bước 7: Truy vấn văn bản tương ứng với động tác ngôn ngữ ký hiệu đã được định nghĩa sẵn.

Sau khi nhận được một tập hợp các động tác ngôn ngữ ký hiệu (trạng thái tay có khả năng xảy ra nhất), tất cả những gì còn lại là ánh xạ chúng đến cơ sở dữ liệu về ngôn ngữ ký hiệu được định nghĩa sẵn trong cơ sở dữ liệu. Việc ánh xạ này là quá trình tìm kiếm từ có đầy đủ các động tác ngôn ngữ ký hiệu đã được nhận diện (Hình 16). Trong

trường hợp không có từ nào khớp với chuỗi động tác đang xem xét, chúng tôi tiến hành bỏ qua và chạy lại từ Bước 1 với dữ liệu hình ảnh mới từ hàng đợi.

❖ Khối điều khiển chuyển đổi văn bản đã nhận diện thành giọng nói và phát ra ngoài

Quy trình chuyển đổi văn bản thành giọng nói được thể hiện trong Hình 17 gồm ba bước cơ bản như sau:

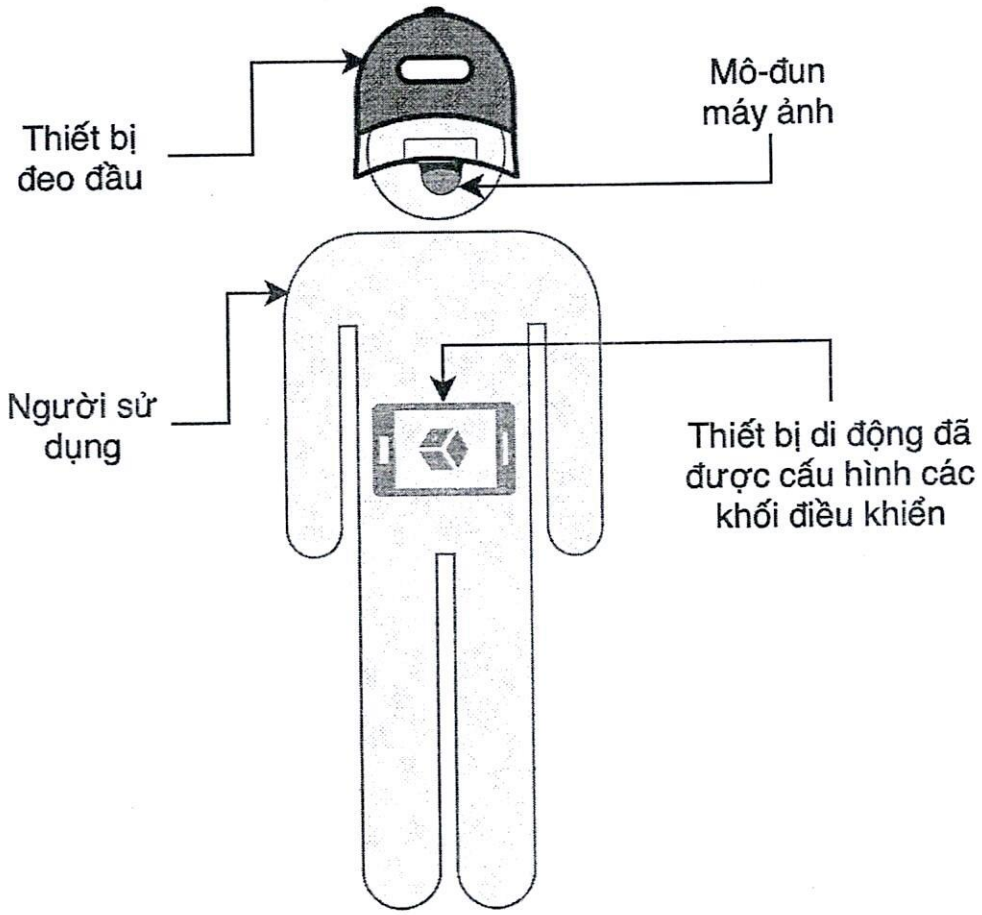
- Bước 1: Thực thi dịch vụ chuyển đổi văn bản thành âm thanh. Đoạn văn bản sau khi đã nhận diện được gửi đến máy chủ chứa dịch vụ chuyển đổi giọng nói của Google. Máy chủ Google thực hiện xử lý và chuyển đổi văn bản thành dữ liệu âm thanh và chuyển về thiết bị di động.
- Bước 2: Khởi động chế độ giọng nói. Hệ thống âm thanh trong thiết bị di động được khởi động và chuẩn bị sẵn sàng cho việc phát âm.
- Bước 3: Phát âm. Các dữ liệu âm thanh được chuyển về từ máy chủ Google được xử lý đổi thành các tín hiệu số. Các tín hiệu số này được truyền đến hệ thống âm thanh trên thiết bị di động, từ đó, âm thanh được phát ra ngoài thông qua loa ngoài của thiết bị di động hoặc các thiết bị nghe tương thích khác.

Hiệu quả đạt được của sáng chế

- Mô-đun máy ảnh có kích thước nhỏ gọn (*65x75x30*) mm, dễ dàng và thuận tiện mang đến bất kỳ đâu.
- Thời lượng pin của mô-đun máy ảnh lên đến 5 giờ cho thời gian sử dụng liên tục.
- Độ chính xác của việc chuyển đổi ngôn ngữ ký hiệu sang văn bản hiệu quả với độ chính xác trên 85% và thời gian phản hồi dưới 3 giây.
- Các khối điều khiển có thể hoạt động tốt khi được cấu hình trên các thiết bị di động giá rẻ dưới ba triệu đồng.
- Kinh phí làm ra toàn bộ hệ thống khá rẻ (chỉ tốn khoảng 600.000 VNĐ cho mô-đun máy ảnh), tạo điều kiện cho những đối tượng kiếm thanh, khiếm thính khó khăn có thể dễ dàng tiếp cận.
- Hệ thống chuyển đổi ngôn ngữ ký hiệu sang văn bản và giọng nói trong thời gian thực có ảnh hưởng to lớn và tích cực đến cộng đồng những người khiếm thanh và khiếm thính, đặc biệt là người khiếm thanh ở Việt Nam.

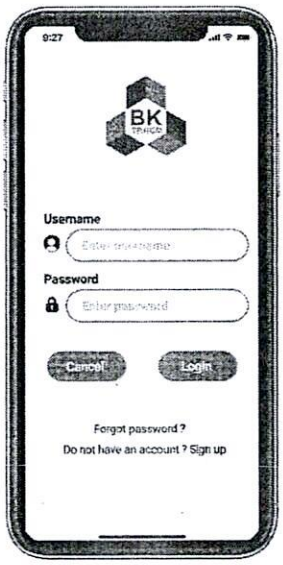
YÊU CẦU BẢO HỘ

I. Hệ thống chuyển đổi thủ ngữ (ngôn ngữ ký hiệu Việt Nam) thành văn bản và giọng nói trong thời gian thực bao gồm những thành phần sau: Mô-đun máy ảnh được gắn trên thiết bị đeo đầu và được cấu hình để ghi nhận hình ảnh ngôn ngữ ký hiệu; Khối điều khiển được tạo cấu hình để điều khiển quá trình thu nhận hình ảnh ngôn ngữ ký hiệu từ mô-đun máy ảnh và gửi đến thiết bị di động thông qua giao thức WebSocket trên mạng Wifi cục bộ được cấu hình trên thiết bị di động; Khối điều khiển được tạo cấu hình để điều khiển quá trình chuyển đổi chuỗi hình ảnh ngôn ngữ ký hiệu sang văn bản trên thiết bị di động bao gồm 7 bước: Thu nhận chuỗi hình ảnh được truyền về từ mô-đun máy ảnh thông qua giao thức WebSocket và mạng Wifi cục bộ; Trích xuất thông tin vị trí các điểm mốc của các đốt ngón tay trên từng hình ảnh bằng mô hình MediaPipe; Phân tích động tác của bàn tay trên từng hình ảnh dựa vào thông tin vị trí của các điểm mốc bằng mô hình học sâu; Tính toán hướng chỉ của các ngón tay trên từng hình ảnh trên các điểm mốc của các đốt ngón tay; Tính toán vị trí tương đối của bàn tay trên cơ thể trên từng hình ảnh dựa vào vị trí và độ lớn của hình ảnh bàn tay; Tổng hợp thông tin về chuỗi các hành động, hướng chỉ và vị trí của bàn tay và so khớp với dữ liệu các động tác ngôn ngữ ký hiệu đã được định nghĩa sẵn bằng việc vectơ hóa các trạng thái tay và thực thi giải thuật Beam Search kết hợp với Connectionist Temporal Classification trên ma trận xác suất của trạng thái tay; Truy vấn văn bản tương ứng với động tác ngôn ngữ ký hiệu đã được định nghĩa sẵn; và Khối điều khiển được tạo cấu hình để điều khiển quá trình chuyển đổi văn bản đã nhận diện được sang giọng nói và phát ra các thiết bị nghe tương thích.

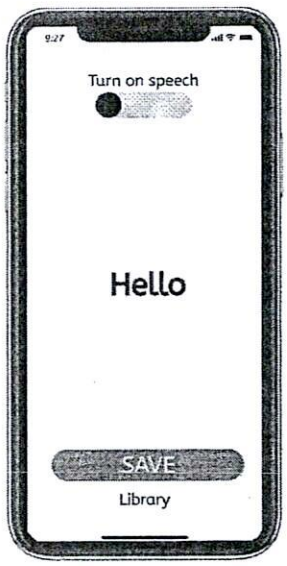

Thiết bị đeo đầu
Mô-đun máy ảnh
Người sử dụng
Thiết bị di động đã được cấu hình các khối điều khiển

Hình 1

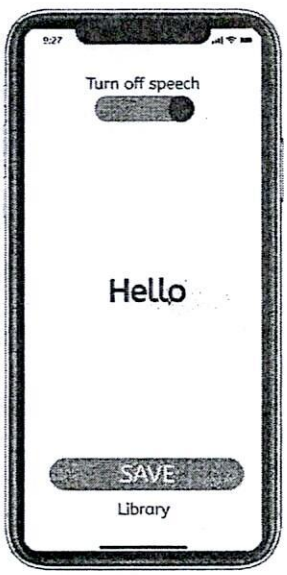
a)



b)

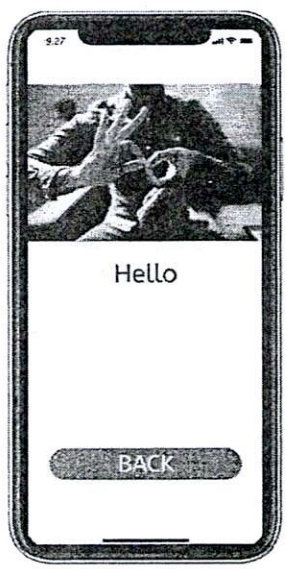


c)



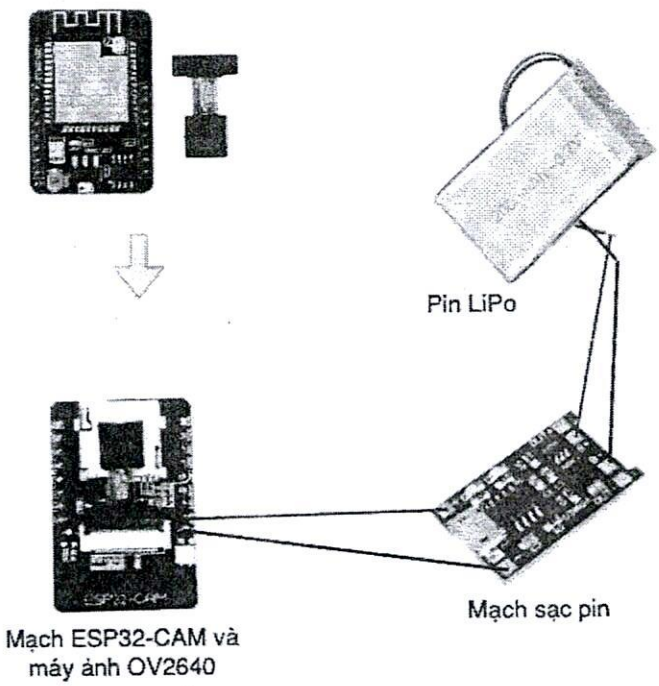
d)

e)

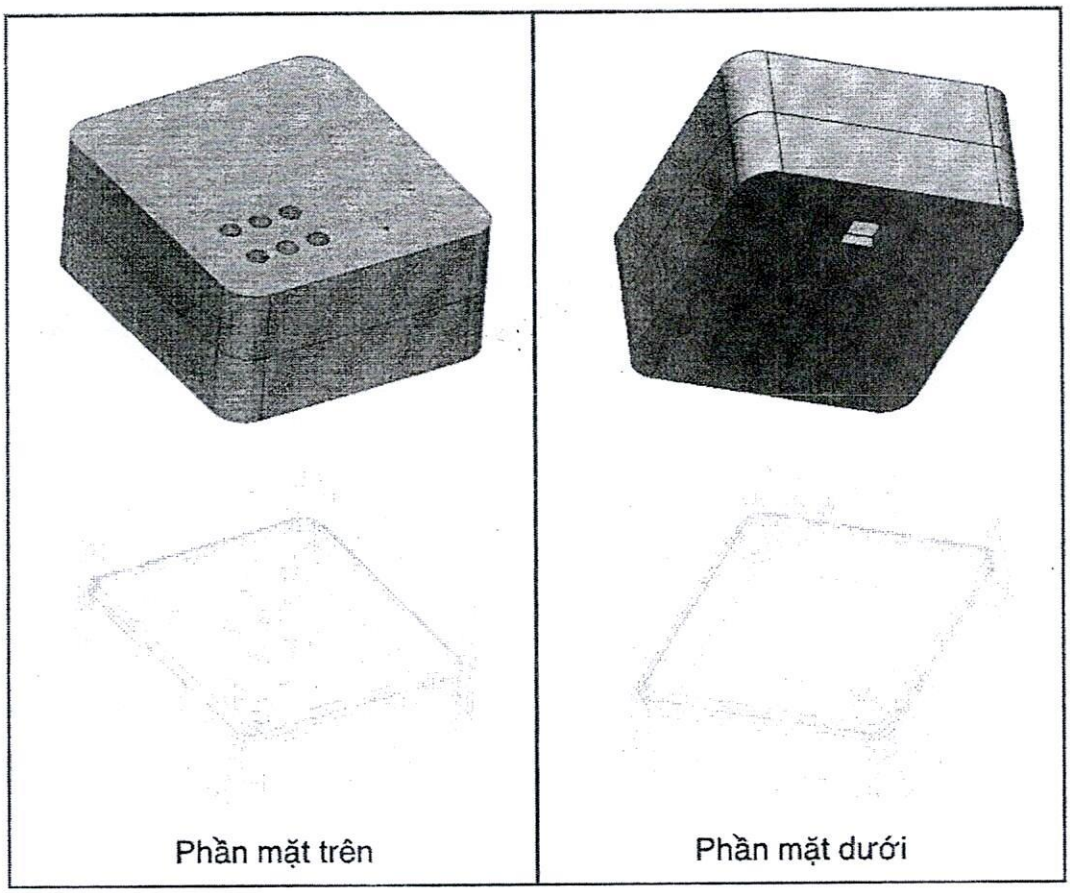


f)

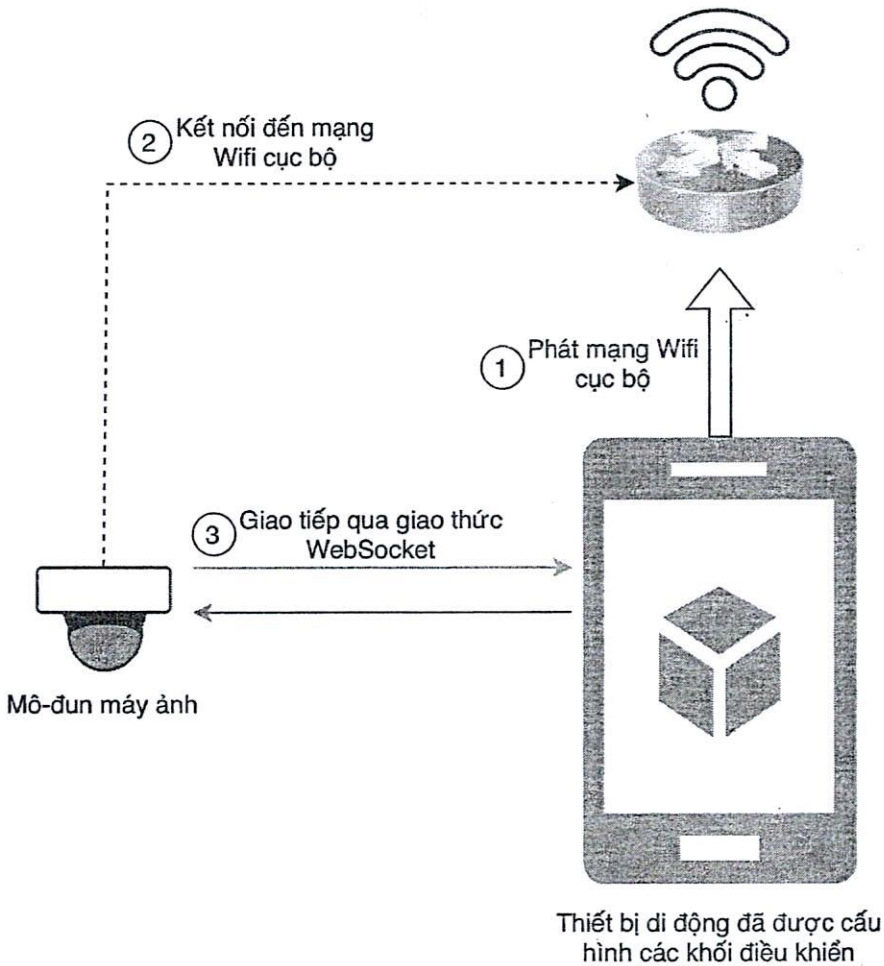
Hình 2



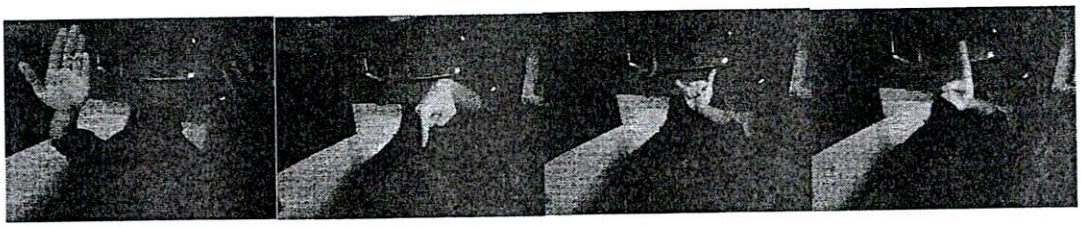
Hình 3



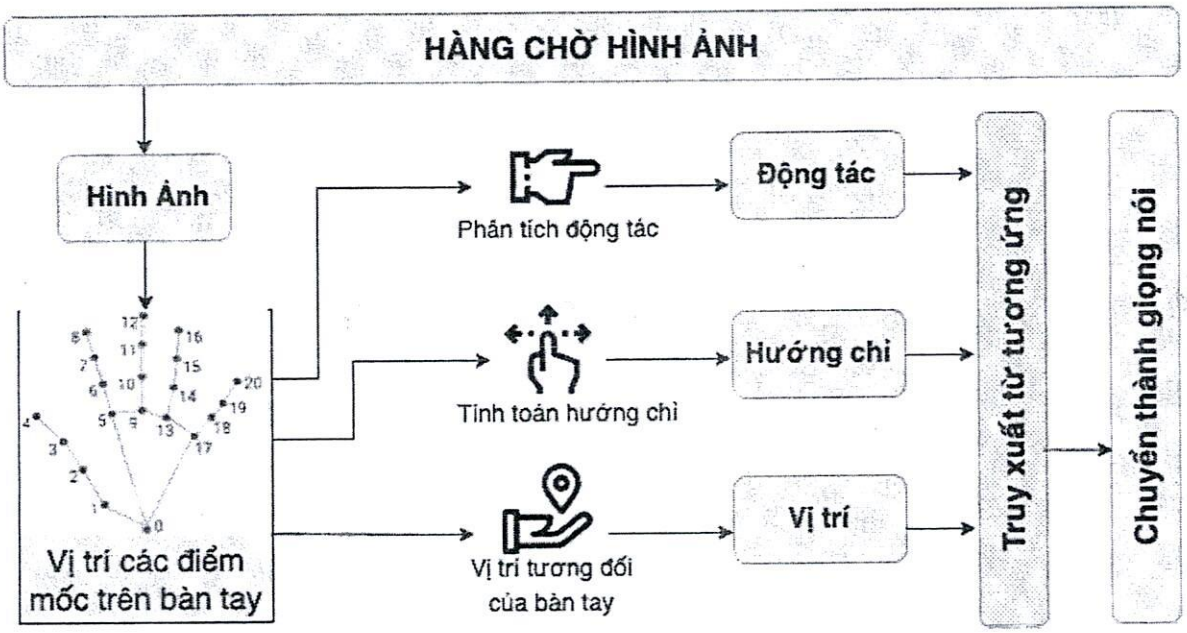
Hình 4


2 Kết nối đến mạng Wifi cục bộ
1 Phát mạng Wifi cục bộ
3 Giao tiếp qua giao thức WebSocket
Mô-đun máy ảnh
Thiết bị di động đã được cấu hình các khối điều khiển

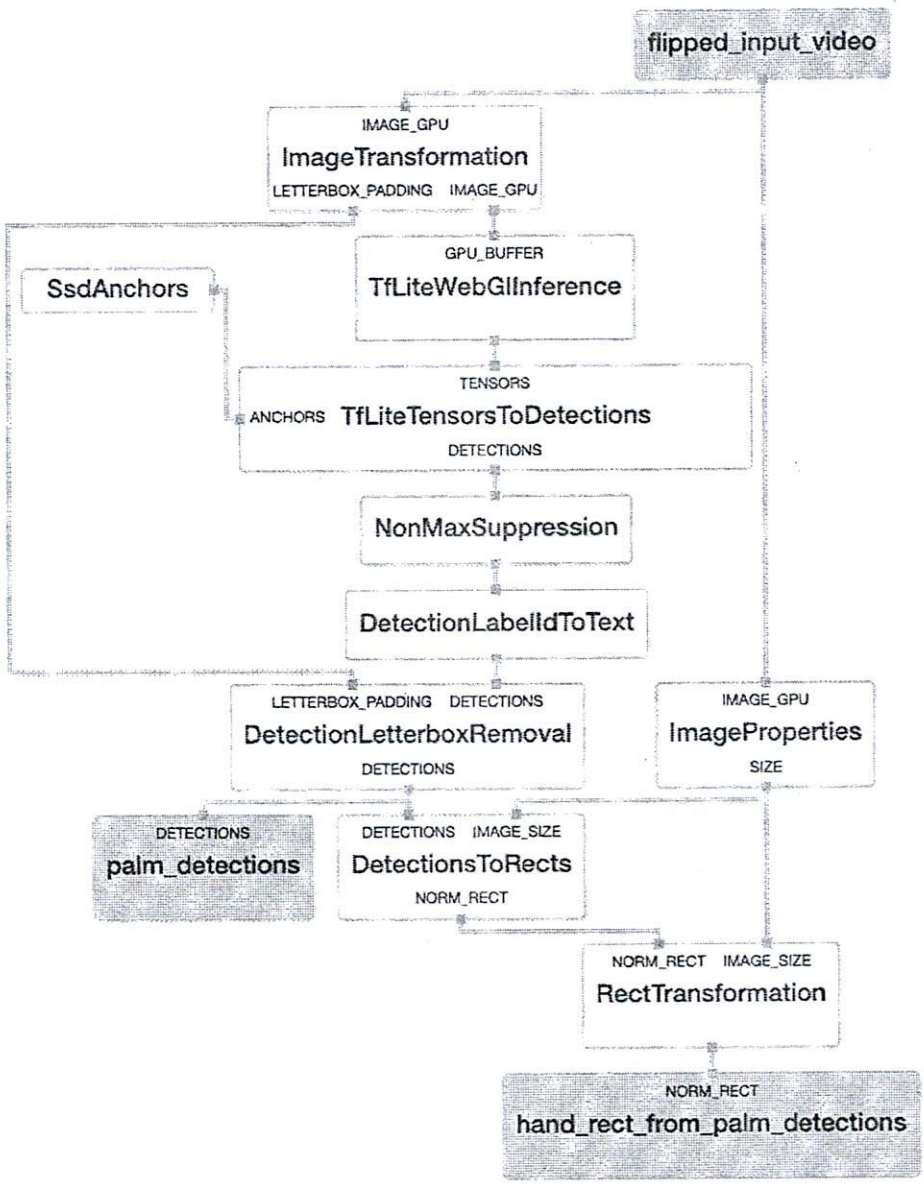
Hình 5



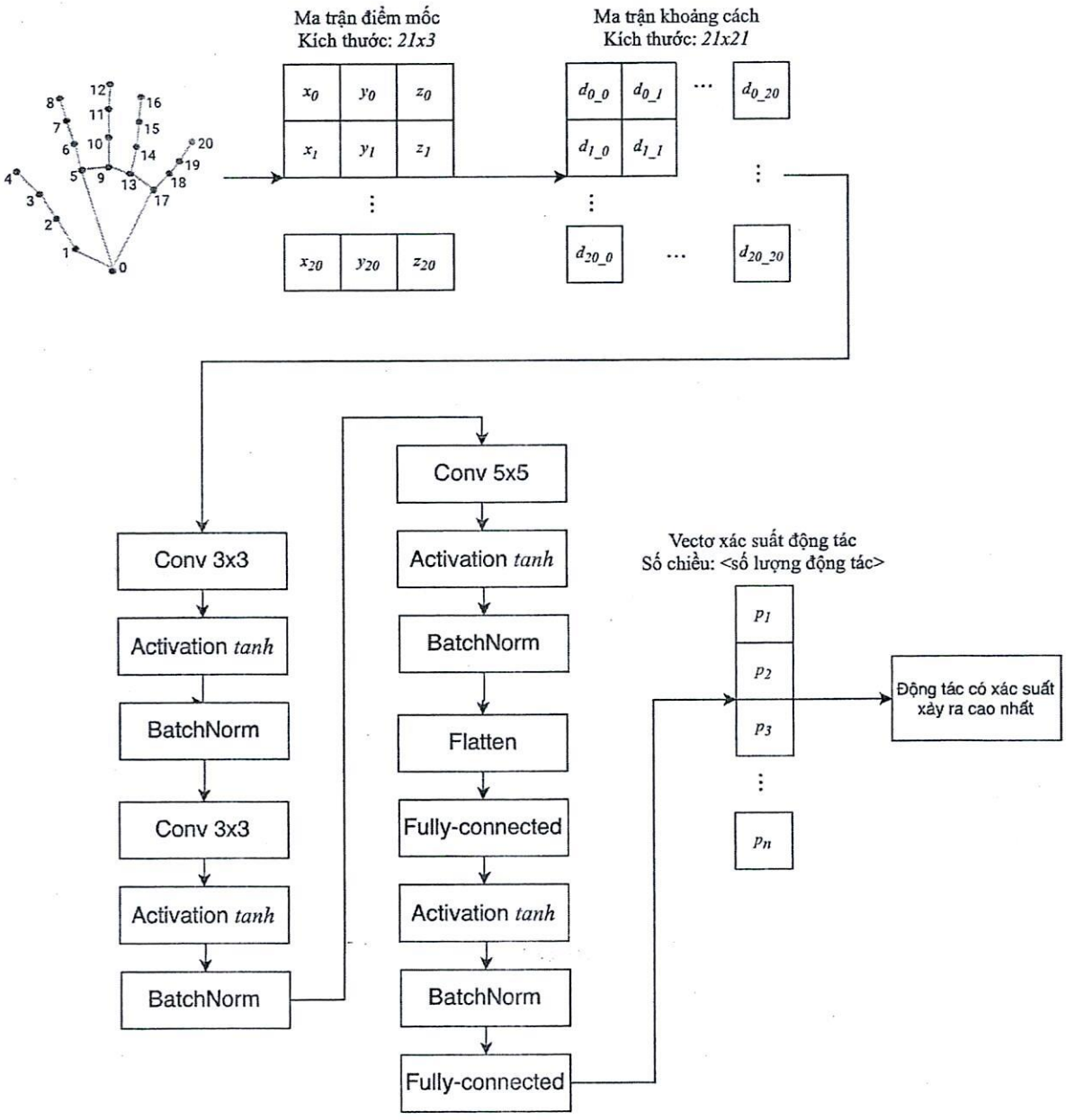
Hình 6


HÀNG CHỜ HÌNH ẢNH
Hình Ảnh
Vị trí các điểm mốc trên bàn tay
Phân tích động tác
Tính toán hướng chỉ
Vị trí tương đối của bàn tay
Động tác
Hướng chỉ
Vị trí
Truy xuất từ tương ứng
Chuyển thành giọng nói

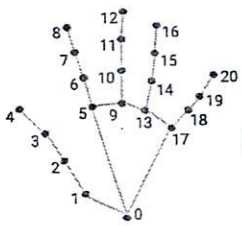
Hình 7


flipped_input_video
IMAGE_GPU
ImageTransformation
LETTERBOX_PADDING IMAGE_GPU
GPU_BUFFER
TfLiteWebGlInference
SsdAnchors
TENSORS
ANCHORS TfLiteTensorsToDetections
DETECTIONS
NonMaxSuppression
DetectionLabelIdToText
LETTERBOX_PADDING DETECTIONS
DetectionLetterboxRemoval
DETECTIONS
IMAGE_GPU
ImageProperties
SIZE
DETECTIONS
palm_detections
DETECTIONS IMAGE_SIZE
DetectionsToRects
NORM_RECT
NORM_RECT IMAGE_SIZE
RectTransformation
NORM_RECT
hand_rect_from_palm_detections

Hình 8


Ma trận điểm mốc
Kích thước: 21x3
Ma trận khoảng cách
Kích thước: 21x21
Conv 3x3
Activation tanh
BatchNorm
Conv 3x3
Activation tanh
BatchNorm
Conv 5x5
Activation tanh
BatchNorm
Flatten
Fully-connected
Activation tanh
BatchNorm
Fully-connected
Vectơ xác suất động tác
Số chiều: <số lượng động tác>
Động tác có xác suất xảy ra cao nhất

Hình 9



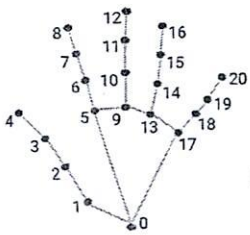
(x, y, z)

- Trái
- Phải
- Hướng lên
- Hướng xuống
- Hướng trước
- Hướng sau

Vẽ vector (0,8), tính toán và so sánh chúng với nhau

Hướng chỉ

Hình 10



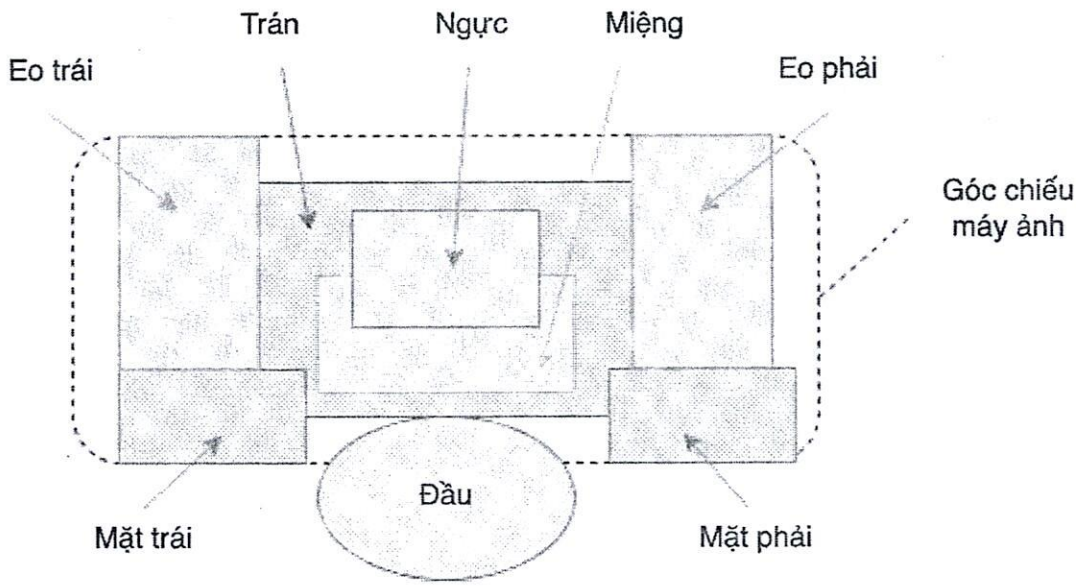
(x, y, z)

Vẽ hình chữ nhật là khung chứa bàn tay xác định từ tọa độ các điểm (0 , 5, 17)

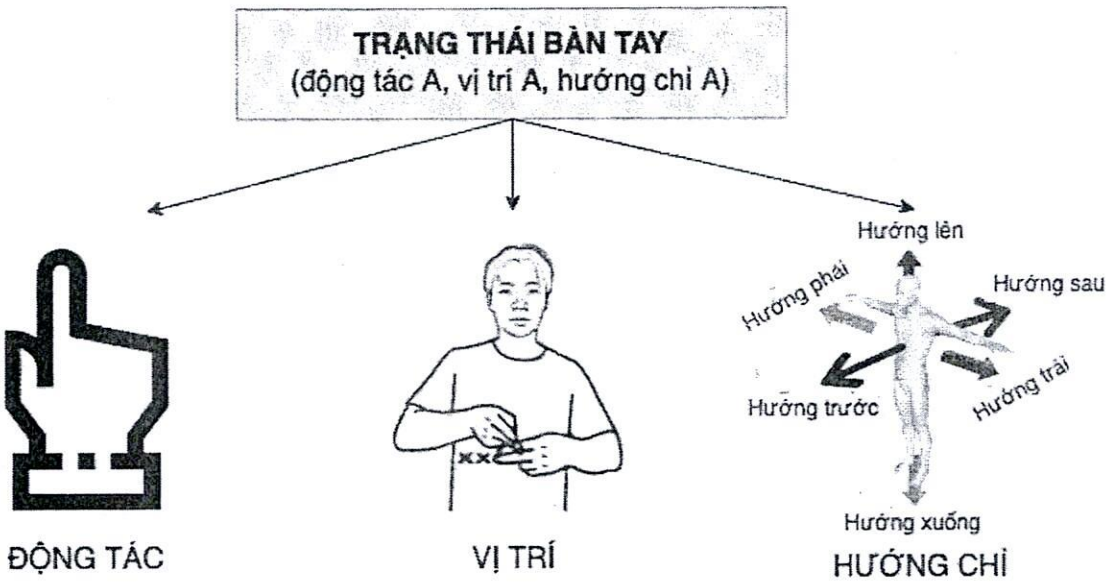
So sánh với các vùng trong cơ thể được định nghĩa sẵn (trán, miệng, ngực, mặt trái, mặt phải, eo trái, eo phải)

Vị trí

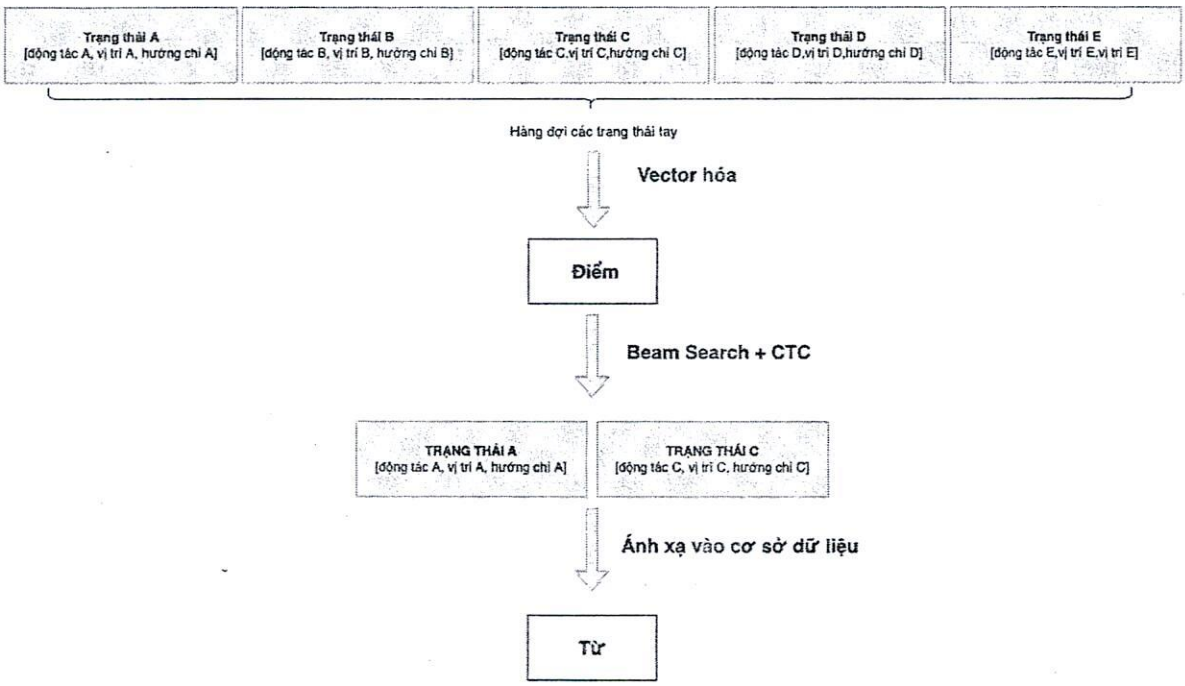
Hình 11


Trán
Ngực
Miệng
Eo trái
Eo phải
Góc chiếu
máy ảnh
Đầu
Mặt trái
Mặt phải

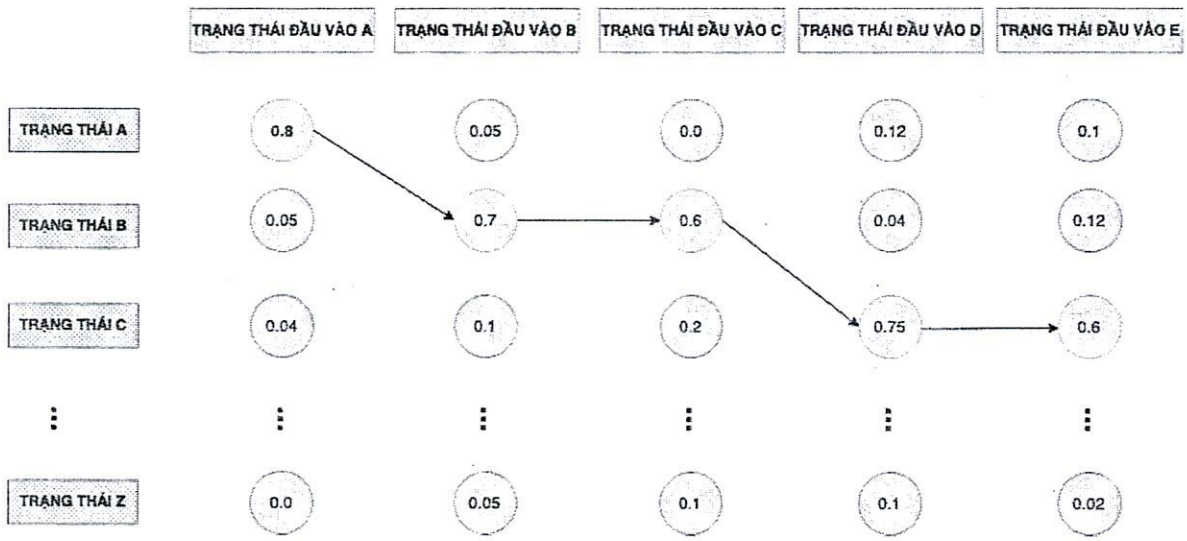
Hình 12


TRẠNG THÁI BÀN TAY
(động tác A, vị trí A, hướng chỉ A)
Hướng lên
Hướng phải
Hướng sau
Hướng trước
Hướng trái
Hướng xuống
ĐỘNG TÁC
VỊ TRÍ
HƯỚNG CHỈ

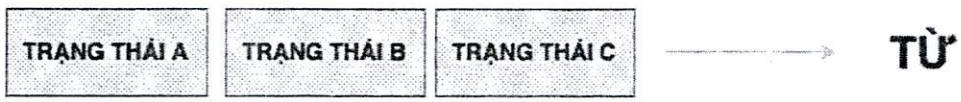
Hình 13


Trạng thái A
[động tác A, vị trí A, hướng chỉ A]
Trạng thái B
[động tác B, vị trí B, hướng chỉ B]
Trạng thái C
[động tác C,vị trí C,hướng chỉ C]
Trạng thái D
[động tác D,vị trí D,hướng chỉ D]
Trạng thái E
[động tác E,vị trí E,vị trí E]
Hàng đợi các trạng thái tay
Vector hóa
Điểm
Beam Search + CTC
TRẠNG THÁI A
[động tác A, vị trí A, hướng chỉ A]
TRẠNG THÁI C
[động tác C, vị trí C, hướng chỉ C]
Ánh xạ vào cơ sở dữ liệu
Từ

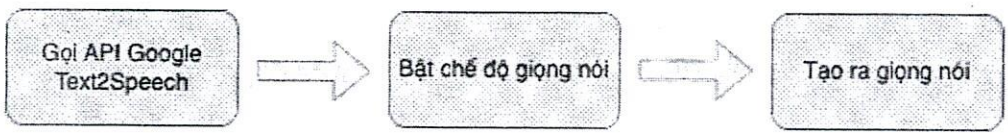
Hình 14


TRẠNG THÁI ĐẦU VÀO A
TRẠNG THÁI ĐẦU VÀO B
TRẠNG THÁI ĐẦU VÀO C
TRẠNG THÁI ĐẦU VÀO D
TRẠNG THÁI ĐẦU VÀO E
TRẠNG THÁI A
TRẠNG THÁI B
TRẠNG THÁI C
TRẠNG THÁI Z
0.8
0.05
0.0
0.12
0.1
0.05
0.7
0.6
0.04
0.12
0.04
0.1
0.2
0.75
0.6
0.0
0.05
0.1
0.1
0.02

Hình 15



Hình 16


Gọi API Google Text2Speech
Bật chế độ giọng nói
Tạo ra giọng nói

Hình 17

53619

SỬA ĐỔI