BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

TÓM TẮT BÁO CÁO TỔNG KẾT ĐỀ TÀI KHOA HỌC VÀ CÔNG NGHỆ CẤP ĐẠI HỌC ĐÀ NẴNG

NGHIÊN CỨU VÀ CẢI TIẾN KỸ THUẬT NHẬN DẠNG NGÔN NGỮ CỬ CHỈ SỬ DỤNG KINECT

Mã số: D2015-02-118

Chủ nhiệm đề tài: ThS. VÕ ĐỨC HOÀNG

Đà Nẵng, 3/2016

BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

TÓM TẮT BÁO CÁO TỔNG KẾT
ĐỀ TÀI KHOA HỌC VÀ CÔNG NGHỆ
CẤP ĐẠI HỌC ĐÀ NẴNG

NGHIÊN CỨU VÀ CẢI TIẾN KỸ THUẬT NHẬN DẠNG NGÔN NGỮ CỬ CHỈ SỬ DỤNG KINECT

Mã số: D2015-02-118

Xác nhận của cơ quan chủ trì đề tài	**Chủ nhiệm đề tài**
(ký, họ và tên, đóng dấu)	*(ký, họ và tên)*

ThS. Võ Đức Hoàng

Đà Nẵng, 3/2016

MỞ ĐẦU

1. Tính cấp thiết của đề tài

Ngôn ngữ ký hiệu là ngôn ngữ cử chỉ tay với dấu hiệu truyền trực quan bằng tay sử dụng hình dạng của bàn tay, hướng và sự di chuyển của bàn tay, cánh tay hoặc cơ thể, nét mặt và miệng để truyền đạt ý nghĩa từ thay vì sử dụng âm thanh. Ngôn ngữ ký hiệu là ngôn ngữ hoàn toàn khác biệt và độc lập với ngôn ngữ nói hay ngôn ngữ viết. Ngôn ngữ này được sử dụng phổ biến trong cộng đồng người khiếm thính bao gồm: thông dịch viên, bàn bè, gia đình của người điếc cũng như trong cộng đồng người có khuyết tật về nghe. Tuy nhiên có rất nhiều trở ngại lớn để tạo ra sự giao tiếp giữa người khiếm thính và người bình thường bởi vì người bình thường không thể hiểu được ngôn ngữ cử chỉ. Nhận dạng ngôn ngữ cử chỉ là thực sự cần thiết để tạo ra một hệ thống tương tác giữa người bình thường và người khiếm thính hay sự giao tiếp giữa người và máy. Hiện nay các hệ thống nhận dạng ngôn ngữ cử chỉ thường sử dụng hai phương pháp sau:

- Dựa trên dữ liệu cảm biến: phương pháp này được thực hiện bằng cách sử dụng hàng loạt các cảm biến được tích hợp trên một găng tay để phát hiện các chuyển động khi thao tác cử chỉ.
- Dựa trên tầm nhìn máy tính: máy tính được gắn máy máy với chức năng là đầu vào của dữ liệu (ảnh, phim). Các tập tin được lưu trữ và xử lý phương phương pháp xử lý hình ảnh và xuất các thông tin, ý nghĩa về ký hiệu của ngôn ngữ ra thiết bị bên ngoài.

Trong hơn thập kỷ qua, nhiều công trình nghiên cứu đã hướng tới phát triển một hệ thống nhận dạng với nhiều ngôn ngữ ký hiệu khác nhau và là thách thức lớn cho nhiều lĩnh vực nghiên cứu như: phương pháp lấy cử chỉ tay, phân loại học máy, giao tiếp của người và máy, xử lý ngôn ngữ tự nhiên... Hầu hết đa số các hệ thống nhận dạng đều giải quyết các cử chỉ một cách riêng biệt và tỉ lệ nhận dạng thành công thấp, chịu sử ảnh hưởng của môi trường thực hiện. Yêu cầu cấp thiết hiện nay là một hệ thống nhận dạng ngôn ngữ ký hiệu liên tục, phải dịch một chuỗi cử chỉ thành một cụm từ hoặc một câu văn bản có ý nghĩa.

Kỹ thuật nhận dạng Ngôn ngữ ký hiệu đang còn ở phạm vi hẹp đối với câu, cụm từ và tỉ lệ nhận dạng còn thấp. Thông thường yếu tố quyết định tỉ lệ nhận dạng tốt phụ thuộc vào quá trình thu nhận ảnh và tiền xử lý để trích xuất đặc trưng. Các nghiên cứu trước thường sử dụng các máy ảnh có độ phân giải cao để thu nhận ảnh, tuy nhiên đến cuối năm 2010 khi Microsoft phát hành thiết bị Kinect đã làm thay đổi phương thức thu nhận dữ liệu đầu vào cho nghiên cứu nhận dạng Ngôn ngữ ký hiệu. Thiết bị Kinect sử dụng webcame 3D, thiết bị thu phát hồng ngoại và thiết bị thu âm thanh. Đối với công cụ tích hợp (SDK) của Kinect có thể xử lý và cho người dùng trích lấy dữ liệu về các vị trí chuyển động của cơ thể bao gồm: 2 bàn tay, 2 khủy tay, đầu, thân và 2 chân hoặc kể cả hình dạng bàn tay có chiều sâu 3D.

Yêu cầu của đề tài là chú trọng phát triển các phương pháp nhận dạng ngôn ngữ cử chỉ đã có và cải tiến một số nghiên cứu giải pháp, thuật toán giúp chuyển đổi ngôn ngữ ký hiệu thành văn bản nhằm tạo ra sự giao tiếp thuận tiện giữa người khuyết tật và người bình thường. Việc nghiên cứu cải tiến các phương pháp nhận dạng cử

chỉ tay có ý nghĩa quan trọng, giúp người khiếm thính hòa nhập tốt với cộng đồng.

2. Mục tiêu và nhiệm vụ đề tài

Mục tiêu

- Tìm hiểu ngôn ngữ ký hiệu tiếng Việt và các nghiên cứu về nhận dạng.
- Nghiên cứu cải tiến các giải pháp, thuật toán cho việc nhận dạng ngôn ngữ cử chỉ sử dụng Kinect.
- Ứng dụng nhận dạng ngôn ngữ ký hiệu trong giao tiếp ở người khiếm thính.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu

- Nghiên cứu về nhận dạng ngôn ngữ cử chỉ.
- Nghiên cứu về thiết bị Kinect và SDK của thiết bị để phát triển.
- Nghiên cứu và xây dựng bộ dữ liệu cho nhận dạng ngôn ngữ cử chỉ tiếng Việt.

Phạm vi nghiên cứu

- Nghiên cứu về ngôn ngữ ký hiệu tiếng Việt.
- Nghiên cứu về các phương pháp thu nhận dữ liệu và xử lý ảnh.
- Nghiên cứu về nhận dạng ngôn ngữ ký hiệu dành cho người khiếm thính Việt Nam, sử dụng thiết bị Kinect để nâng cao kết quả nhận dạng.

4. Cách tiếp cận, phương pháp nghiên cứu

Cách tiếp cận

- Nghiên cứu về giải pháp cải tiến thuật toán cho nhận dạng ngôn ngữ cử chỉ với Kinect.
- Xây dựng công cụ nhận dạng ngôn ngữ cử chỉ.
- Thử nghiệm, đánh giá hiệu quả nhận dạng của công cụ mới so với các nghiên cứu trước.

Phương pháp nghiên cứu

- Tìm hiểu về lý thuyết xử lý và nhận dạng ảnh.
- Phát triển ứng dụng và cải tiến thuật toán nhận dạng bằng Kinect.
- Khảo sát các mô hình, thuật toán nhận dạng cử chỉ.

5. Nội dung dung

- Nghiên cứu tổng quan về nhận dạng ngôn ngữ cử chỉ.
- Khảo sát các phương pháp thu nhận dữ liệu.
- Khảo sát và đánh giá các phương phấp nhận dạng đã được nghiên cứu.
- Đề xuất nghiên cứu đối với ngôn ngữ cử chỉ tiếng Việt.
- Đánh giá hiệu quả.

6. Cấu trúc đề tài

Nội dung luận văn được trình bày bao gồm các phần chính như sau:

Chương 1: Nêu tổng quan về các phương pháp nghiện cứu về nhận dạng ngôn ngữ ký hiệu đã có tại Việt Nam và trên thế giới. Đồng thời nêu lên các đặc điểm của ngôn ngứ ký hiệu tiếng Việt để có thể đề xuất các phương pháp thu nhận dữ liệu và trích xuất đặc trưng cho quá trình phân loại và nhận dạng.

Chương 2: Trình bày tổng quan về cử chỉ tĩnh của ngôn ngữ ký hiệu tiếng Việt, cụ thể ở đây là Bảng chữ cái và chữ số. Thông qua các nghiên cứu về nhận dạng cử chỉ tĩnh của ngôn ngữ ký hiệu, chúng tôi đã trình bày đề xuất về thu nhận dữ liệu, cải tiến phương pháp trích xuất đặc trưng và nâng cao tỉ lệ nhận dạng.

Chương 3: Trình bày các phương pháp về nhận dạng cử chỉ liên tục của ngôn ngữ ký hiệu. Tuy kết quả nghiên cứu chưa đạt tỉ lệ thành công cao nhưng đây là tiền đề để phát triển các nghiên cứu tiếp theo

Phần kết luận tổng hợp tất cả các quá trình nghiên cứu và đưa ra các đề xuất cho nghiên cứu trong thời gian tiếp theo.

CHƯƠNG 1
NGHIÊN CỨU TỔNG QUAN

1.1. Tổng quan

Ngôn ngữ ký hiệu là ngôn ngữ cử chỉ tay với dấu hiệu truyền trực quan bằng tay sử dụng hình dạng của bàn tay, hướng và sự di chuyển của bàn tay, cánh tay hoặc cơ thể, nét mặt và miệng để truyền đạt ý nghĩa từ thay vì sử dụng âm thanh. Ngôn ngữ ký hiệu là ngôn ngữ hoàn toàn khác biệt và độc lập với ngôn ngữ nói hay ngôn ngữ viết. Sự khác biệt cơ bản là hạn chế vốn từ vựng của ngôn ngữ ký hiệu. Ngôn ngữ ký hiệu có sự khác biệt rất lớn giữa các quốc gia như Mỹ (ASL), Đức (GSL), Trung Quốc (CSL), Việt Nam (VSL) ..... và giữa các vùng miền trong một quốc gia Việt Nam như Hà Nội, Hải Phòng, Cần Thơ, Hồ Chí Minh về từ vựng hay cách biểu diễn cử chỉ. Ngôn ngữ này được sử dụng phổ biến trong cộng đồng người khiếm thính bao gồm: thông dịch viên, bàn bè, gia đình của người điếc cũng như trong cộng đồng người có khuyết tật về nghe. Tuy nhiên, hiện nay ngôn ngữ này không được phổ biến trong cộng đồng giao tiếp do đó có một rào cản lớn giữa người khiếm thính và người bình thường.

Sự giao tiếp bằng ngôn ngữ ký hiệu rất đa dạng không chỉ liên quan đến ký hiệu bàn tay mà còn được định nghĩa là mô hình cụ thể hay chuyển động của bàn tay, nét mặt hoặc cơ thể. Ngôn ngữ ký hiệu có thể chia làm hai phần là tư thế tay và cử chỉ tay. Thể hiện tư thế tay được định nghĩa là một hình dạng cụ thể của bàn tay vào một thời điểm tức thì, một cử chỉ tay được định nghĩa là hệ quả của tư thế tay di chuyển trong một miền thời gian.

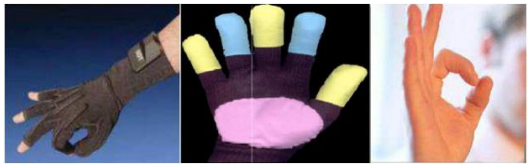
Trong hơn thập kỷ qua, nhiều công trình nghiên cứu đã hướng tới phát triển một hệ thống nhận dạng với nhiều ngôn ngữ ký hiệu

khác nhau và các nhà nghiên cứu đã kết luận rằng một hệ thống như vậy là thách thức lớn cho nhiều lĩnh vực nghiên cứu khác nhau như: phương pháp lấy cử chỉ tay, phân loại học máy, sự giao tiếp của người và máy, xử lý ngôn ngữ tự nhiên... Hầu hết đa số các hệ thống nhận dạng đều giải quyết các cử chỉ một cách riêng biệt. Yêu cầu cấp thiết hiện nay là một hệ thống nhận dạng ngôn ngữ ký hiệu liên tục, phải dịch một chuỗi cử chỉ thành một cụm từ hoặc một câu văn bản có ý nghĩa. Sự phức tạp trong nhận dạng ngôn ngữ ký hiệu phát sinh từ thực tế là vốn từ vựng của ngôn ngữ ký hiệu ít, cách biểu diễn các từ đồng âm nhưng khác nghĩa, sự phân chia cách biểu diễn liên tục nhiều từ... Nhận dạng ngôn ngữ ký hiệu liên tục đã trở thành một lĩnh vực nghiên cứu quan trọng với trọng tâm là nhận dạng cử chỉ tay và nhận dạng cử chỉ tương tác với cảm xúc con người. Khi có thiết bị Kinect, một bộ điều khiển trò chơi dành cho XBOX nhằm tạo tương tác giữa người chơi và máy tính thì nó thu hút rất nhiều nhà nghiên cứu bởi vì thiết bị có thể nhận dạng chuyển động của con người và thu nhận hình ảnh có chiều sâu (3D).

1.2. Các phương pháp thu nhận dữ liệu

Bước đầu tiên quan trọng của việc xử lý nhận dạng ngôn ngữ ký hiệu là thu thập dữ liệu thô. Dữ liệu thô sau đó được phân tích bằng cách sử dụng các thuật toán khác nhau để trích xuất đặc trưng và đưa vào các mô hình thống kê để nhận dạng. Trước đây trong nghiên cứu nhận dạng ngôn ngữ ký hiệu có thể chia thành 2 lĩnh vực dựa vào phương pháp thu nhận dữ liệu: một là dựa vào dữ liệu các cảm biến có thể đặt trên các bộ phận của cơ thể người, hai là dựa trên thị giác máy tính. Trong phương pháp thu nhận dựa cảm biến đặt trên cơ thể có thể là các cảm biến sinh học điện cơ, cảm biến điện tử hay

là các găng tay điện tử, găng tay màu. Còn trên thị giác máy tính, thì máy ảnh được sử dụng là thiết bị đầu vào bao gồm ảnh và đoạn phim. Các đoạn phim được lưu trữ trước khi xử lý, được tách thành các phân đoạn đặc biệt và xử lý tương tự như xử lý hình ảnh. Nhìn chung, chúng ta có thể phân loại thành 3 nhóm cơ bản sau: *găng tay cảm biến, găng tay màu và thị giác máy tính* (Hình 1).



Hình 1: Các kỹ thuật thu nhận dữ liệu đầu vào.

Phương pháp thu nhận dữ liệu dựa trên găng tay cảm biến yêu cầu người dùng phải đeo một thiết bị găng tay cồng kềnh. Găng tay được trang bị các cảm biến để cảm nhận sự chuyển động của bàn tay và các ngón tay và truyền các thông tin vào máy tính. Phương pháp này dễ dàng cung cấp chính xác tọa độ vị trí lòng bàn tay, ngón tay và hướng, hình dạng bàn tay. Ưu điểm của phương pháp này là độ chính xác cao và tốc độ xử lý nhanh. Tuy nhiên khi sử dụng phương pháp này, yêu cầu găng tay của người dùng phải được kết nối trực tiếp với máy tính nên cản trở sự tương tác của người thực hiện và khoảng cách giữa người và máy, đặc biệt chi phí của thiết bị khá cao.

Phương pháp thu nhận dữ liệu dựa trên găng tay màu sắc đã khắc phục được các nhược điểm của găng tay cảm biến và đây là sự kết hợp giữa phương pháp thu nhận dữ liệu dựa trên găng tay và thị giác máy tính. Găng tay thường là màu trắng và được đánh dấu bởi các màu khác nhau giữa các ngón tay và lòng bàn tay. Một máy ảnh màu có thể nhận biết và theo dõi sự chuyển động, hình dạng, vị trí

của lòng bàn tay, ngón tay. Sự tiện lợi của phương pháp này là người dùng không bị phụ thuộc nhiều vào khoảng cách so với máy tính và chi phí cho chế tạo găng tay nhỏ. Về bản chất hai phương pháp sử dụng găng tay là tương tự như nhau, nhưng khi sử dụng găng tay màu phải trải qua giai đoạn tiền xử lý. Tuy nhiên cách tiếp cận này không được tự nhiên (do phải sử dụng găng tay) và không được nhiều người sử dụng (do vấn đề về vệ sinh).

Phương pháp tiếp cận dựa trên thị giác máy tính, người thực hiện không cần đeo bất kỳ một thiết bị gì. Các thao tác cử chỉ được thực hiện một cách tự nhiên như trong giao tiếp của cuộc sống. Thay vào đó, một hay nhiều máy quay được sử dụng để chụp ảnh hay quay các đoạn phim của bàn tay, sự di chuyển của bàn tay, cánh tay. Đây là phương pháp nhận dạng ngôn ngữ ký hiệu (cử chỉ) đơn giản, tự nhiên và tiện lợi nhất cho người sử dụng, được sử dụng rộng rãi nhất. Mặc dù phương pháp này đơn giản nhưng lại đặt ra rất nhiều thách thức cho quá trình tiền xử lý như: phải loại bỏ hình ảnh nhiễu bởi nền, phụ thuộc vào điều kiện ánh sáng, màu da và trang phục mặc trên người. Yêu cầu hệ thống xử lý phải có cấu hình cao, tốc độ xử lý nhanh và hiệu quả.

Tuy nhiên đến cuối năm 2010 khi Microsoft phát hành thiết bị Kinect đã làm thay đổi phương thức thu nhận dữ liệu đầu vào cho nghiên cứu nhận dạng Ngôn ngữ ký hiệu. Trong thời gian gần đây, các thông tin thu được từ cảm biến chiều sâu được sử dụng nhiều trong các nghiên cứu. Việc phân đoạn bàn tay được thực hiện dựa trên ảnh chiều sâu và thuật toán theo dõi hình ảnh không gian 3 chiều. Thiết bị Kinect cũng thu nhận dữ liệu dựa trên phương pháp thị giác máy tính. Tuy nhiên, thiết bị Kinect sử dụng webcam 3D,

thiết bị thu phát hồng ngoại và thiết bị thu âm thanh. Đối với công cụ tích hợp (SDK) của Kinect có thể xử lý và cho người dùng trích lấy dữ liệu về các vị trí chuyển động của cơ thể bao gồm: 2 bàn tay, 2 khủy tay, đầu, thân và 2 chân hoặc kể cả hình dạng bàn tay có chiều sâu 3D. Khi đã có dữ liệu thu nhận vào ta sử dụng các phương pháp học máy để có thể nhận dạng. Một ưu điểm chính của thiết bị Kinect là đã khắc phục được các yếu tố gây nhiễu trong quá trình thu nhận dữ liệu như: ảnh nền, ánh sáng, màu da, vị trí cổ tay, ngón tay.

1.3. Các phương pháp phân loại và nhận dạng ngôn ngữ ký hiệu

Có nhiều phương pháp được sử dụng để phân loại nhận dạng ngôn ngữ ký hiệu, các phương pháp này dựa trên các thông số sau khi trích chọn đặc trưng từ các dữ liệu đã xử lý sau khi thu nhận bằng các phương pháp ở phần 1.2. Các phương pháp như: Mạng nơ ron nhân tạo (ANN), Mô hình Markov ẩn (HMM), Máy vector hỗ trợ (SVM), Đường cong theo thời gian động (DTW), mô hình hỗn hợp Gaussian (GMM)... Hầu hết các phương pháp này đều dựa trên mô hình thống kê và tự học, có khả năng tự tối ưu hóa các thông số qua quá trình đào tạo để nâng cao khả năng phân loại và nhận dạng dựa vào các thông số ẩn.

1.4. Ngôn ngữ ký hiệu tiếng Việt

Lịch sử phát triển của ngôn ngữ ký hiệu nói chung và ngôn ngữ ký hiệu tiếng Việt nói riêng đã trải qua nhiều giai đoạn thăng trầm. Từ thế kỉ 16, Geronimo Cardano - nhà vật lý học người Padua, đã tuyên bố người khiếm thính có thể học tập thông qua giao tiếp bằng ký hiệu. Đến năm 1620, Juan Pablo de Bonet xuất bản cuốn sách đầu tiên về ngôn ngữ ký hiệu, đồng thời công bố bảng chữ cái

năm 1620 dựa trên nền tảng là ngôn ngữ ký hiệu đã được cộng đồng người điếc phát triển theo bản năng từ trước.

Ở Việt Nam, ngôn ngữ ký hiệu đã được đưa vào giáo dục và sử dụng từ rất sớm: từ năm 1866, một linh mục người Pháp là cha Azemar đã quy tụ khoảng 5 trẻ khiếm thính để dạy ngôn ngữ và đạo đức. Sau đó, một trong những trẻ này đã sang Pháp để học tập phương pháp dùng ngôn ngữ ký hiệu điệu bộ. Đến năm 1886, khi anh về nước, linh mục đã tuyên bố mở trường dạy trẻ khiếm thính tại Thuận An. Trung tâm này chính là cái nôi của người khiếm thính tại Việt Nam. Nơi đây hơn một trăm năm qua, biết bao thế hệ những người khiếm thính đã được nuôi dưỡng và giáo dục.

Từ những năm 2000, Việt Nam bắt đầu triển khai những nỗ lực của mình nhằm hoàn thiện và hệ thống hóa ngôn ngữ ký hiệu Việt Nam. Các câu lạc bộ, nhóm học tập bắt đầu hình thành và phát triển. Một số tài liệu khá công phu xuất hiện như: bộ 3 tập Ký hiệu cho người điếc Việt Nam, từ điển ngôn ngữ ký hiệu Việt Nam.

Hình 2: Bảng chữ cái ngôn ngữ ký hiệu tiếng Việt.

Bảng chữ cái ngôn ngữ ký hiệu là một loại cử chỉ tay. Tương tự như ngôn ngữ viết tiếng Việt xuất phát từ ký tự Latin, ngôn ngữ ký hiệu tiếng Việt được xây dựng tương tự như ngôn ngữ ký hiệu Mỹ (ASL) đã được sử dụng rộng rãi ở một số quốc gia. Bảng chữ cái bao gồm 23 chữ cái, các từ ghép, dấu mũ và dấu thanh. Các chữ cái Ă, Â, Ê, Ô, Ơ, Ư, CH, GH, NGH là sự kết hợp từ 2 hoặc 3 cử chỉ tay liên tục.

Ngoài các ký hiệu biểu diễn bảng chữ cái ngoài ra còn có các biểu diễn cử chỉ được sử dụng để mô tả các đối tượng, con người... Các dấu hiệu này có thể chia thành hai nhóm dựa trên bản chất của cử chỉ: dấu hiệu tự nhiên và dấu hiệu thông qua giáo dục.

Dấu hiệu tự nhiên là các dấu hiệu hay cử chỉ mà con người học hỏi từ các dấu hiệu trong tự nhiên được sử dụng để mô tả các hành động chung trong các hoạt động hằng ngày như: ăn uống, ca hát, khóc, ngủ, đói bụng...

Dấu hiệu thông qua giáo dục dùng để diễn tả các khái niệm trừu tượng hoặc các đối tượng trong thực tế cuộc sống như đẹp, xấu xí, thích, hạnh phúc... Những cử chỉ này không thể hiểu được đối với người bình thường và người khiếm thính nếu không qua các lớp đào tạo.

Ta có thể phân tích cử chỉ là một chuỗi các hình ảnh tĩnh. Mỗi hình ảnh chứa thông tin của một dấu hiệu cụ thể bao gồm hình ảnh bàn tay, vị trí tay và biểu hiện khuôn mặt... Các thông tin này được trích xuất đặc trưng và lưu trữ để so sánh với các đặc trưng của các hình ảnh trước và sau trong cử chỉ đó. Dựa trên tổng hợp các đặc điểm này sẽ đề xuất cách phân tích và nhận dạng cho hợp lý.

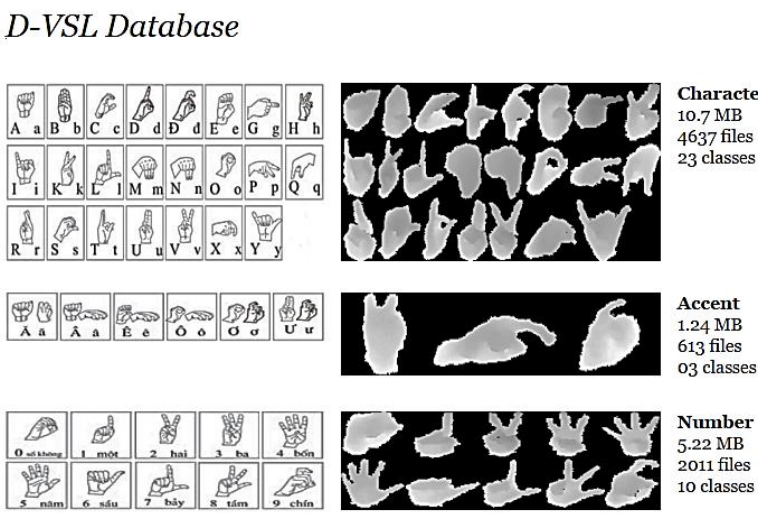
CHƯƠNG 2
NHẬN DẠNG CỬ CHỈ TĨNH

2.1. Tổng quan

Trong chương này tôi đề xuất một phương pháp tiếp cận, có thể thực hiện trong thời gian thực để nhận biết các cử chỉ tĩnh của ngôn ngữ ký hiệu. Thay vì sử dụng dữ liệu RGB như nhiều giải pháp khác, đầu vào của hệ thống là hình ảnh chiều sâu thu nhận bởi thiết bị Microsoft Kinect. Để mô tả cử chỉ tay, tôi sử dụng kỹ thuật xếp hạng ma trận tương đương (rank-order correlation matrix - ROCM). Căn cứ vào tính chất của bảng chữ cái ngôn ngữ ký hiệu tiếng Việt và cách thu nhận dữ liệu, có thể sử dụng các cách phân loại và nhận dạng khác nhau. Trong nghiên cứu này tôi sử dụng cách phân loại nhiều vec-tơ hỗ trợ học máy (Multiple support vector machines - SVMs) kết hợp với kỹ thuật MAX-WINS để nhận dạng. Các thí nghiệm được thực hiện trên trên ba bộ dữ liệu hình ảnh chiều sâu của cơ sở dữ liệu ngôn ngữ ký hiệu tiếng Việt (D_VSL) và nhận được nhiều kết quả khả quan.

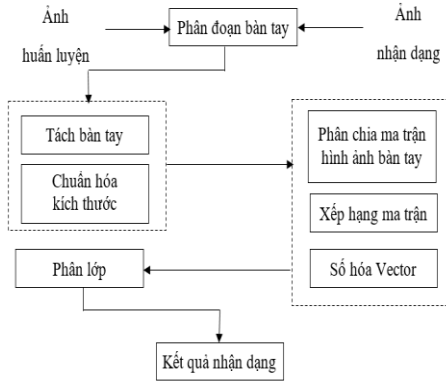
Bảng chữ cái ngôn ngữ ký hiệu tiếng Việt bao gồm các ký tự đơn tương tự như ngôn ngữ ký hiệu Mỹ gồm 23 lớp ký tự (dữ liệu bảng thứ nhất) và các ký tự có sự kết hợp của hai biểu tượng đơn (bảng thứ nhất và thứ 2) bao gồm các ký tự có mũ, các dấu thanh và các ký tự ghép. Ý tưởng tiếp cận của tôi là nhận dạng các ký tự đơn và lần lượt kết hợp thêm các ký tự ghép. Đầu vào của hệ thống là hình ảnh chiều sâu được thu nhận bởi cảm biến chiều sâu thiết bị Microsoft Kinect. Trong thiết bị này một máy phát tia hồng ngoại (IR) và một camera thu để đo được độ sâu tại mỗi điểm ảnh của ảnh. Hình ảnh thu được không bị ảnh hưởng bởi ảnh sáng. Mỗi hình ảnh

được tạo ra ở độ sâu 30fps với độ phân giải 640*480.



Hình 3: Bộ dữ liệu hình ảnh chiều sâu cử chỉ tĩnh.

2.2. Quy trình nhận dạng cử chỉ tĩnh



Hình 4: Sơ đồ khối nhận dạng cử chỉ tĩnh.

2.2.1. Phân đoạn bàn tay

Trong nhiều cách tiếp cận, bàn tay được phát hiện bằng cách sử dụng bộ lọc màu da. Các nghiên cứu thường tiếp cận như vậy tuy nhiên kết quả có thể bị ảnh hưởng bởi điều kiện môi trường. Để tránh sự hạn chế này, trong nghiên cứu của tôi đề xuất sử dụng thông tin ảnh chiều sâu. Thiết bị Kinect sử dụng cảm biến chiều sâu với khoảng cách thu nhận từ 0.8m đến 4.0m và tích hợp các thuật toán để thu nhận. Khi thực hiện các thao tác thể hiện ngôn ngữ ký hiệu, bàn tay là phần cơ thể gần thiết bị Kinect nhất.

2.2.2. Tiền xử lý

Tách bàn tay: Sau khi chọn phạm vi thu nhận ảnh thích hợp, ảnh thu được có thể bị nhiễu nhẹ do phụ thuộc vào môi trường và cảm biến. Sử dụng bộ lọc hình thái không gian để loại bỏ nhiễu và làm mịn ảnh, đồng thời sử dụng thuật toán xác định biên và làm mịn đối tượng. Cuối cùng ta có hình ảnh bàn tay dựa trên khung của nó.

Chuẩn hóa kích thước: Có nhiều phương pháp để thay đổi kích thước của hình ảnh bàn tay trước giai đoạn trích xuất đặc trưng. Một điểm bất lợi về hình ảnh thu được từ bàn tay là kích thước thu được với tỉ lệ chiều đứng và chiều ngang khác nhau (bàn tay thể hiện đứng hay ngang) do vậy sẽ ảnh hưởng rất lớn đến các bước xử lý tiếp theo. Vì vậy cần xử lý để đưa hình ảnh bàn tay về một kích thước chuẩn là cần thiết.

2.2.3. Trích xuất đặc trưng

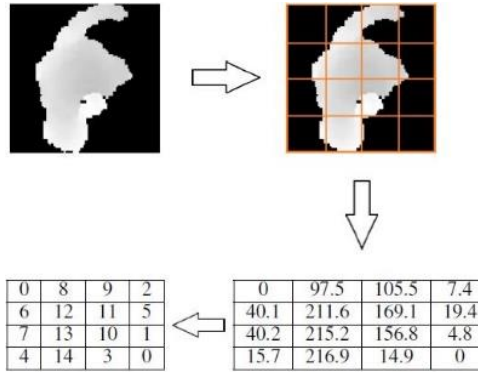
Chia lưới (ma trận) hình ảnh: Trong nghiên cứu này tôi sử dụng một lưới vuông để chia ảnh chiều sâu bàn tay thành d ô. Sau đó tính toán giá trị các ô này dựa trên giá trị trung bình của các điểm ảnh thuộc ô đó. Kết quả thu được là ma trận vuông có các giá trị trung bình tương ứng từng ô.

Thống kê thông tin: Để mô tả giá trị một ô, tương ứng với một khu vực hình ảnh. Xét một tập hợp n điểm ảnh với các giá trị độ sâu xi tương ứng, hai thuộc tính được mô tả như hình 5.

Sau khi tính toán cho tất cả các ô kết quả thu được là hai ma trận vuông cấp d. Ma trận thứ nhất, M_atm bao gồm d2 giá trị trung bình, ma trận thứ hai M_atsd bao gồm d2 giá trị độ lệch tiêu chuẩn.

Xếp hạng ma trận: Mỗi ma trận vuông cấp 2 được chuyển đổi thành ma trận xếp hạng tương ứng có cùng kích thước dựa vào giá trị các phần tử để xếp hạng. Các giá trị của ma trận M_atm được sắp xếp theo thứ tự tăng dần sau đó đánh giá trị thứ hạng được bắt đầu từ 0 và chuyển giá trị xếp hạng tương ứng vào ma trận M_atsd.

Tạo vector: Để tương thích với kỹ thuật phân loại, mỗi ma trận xếp hạng sẽ được biểu diễn như một vector, được đặt tên là vector kết hợp. Mỗi phần tử của vector mô tả mối quan hệ giữa hai ô lân cận, tương ứng với hai yếu tố liên tiếp của ma trận xếp hạng.



Hình 5: Xếp hạng giá trị trung bình ma trận 4*4.

2.2.4. Phân lớp và nhận dạng

Mô hình học máy hỗ trợ vec-tơ (Support Vector Machine – SVM)

là một mô hình mạnh mẽ dùng để sử dụng trong phân tích dữ liệu và nhận dạng mẫu, phân loại dựa vào các giá trị đặc trưng.

Có 5 mô hình được đề cập trong nghiên cứu, trong đó mỗi mô hình là một SVM đa lớp được xây dựng được từ từng lớp riêng biệt. Cụ thể mô hình 1 được tạo ra dựa trên 23 ký tự đơn (một bàn tay) từ A đến Y. Mô hình hai được xây dựng dùng để phân loại mẫu của 6 lớp bao gồm: dấu mũ, dấu mũ ngược, dấu móc, ký tự H, ký tự G và ký tự R. Ba mô hình còn lại tương ứng với 3 tập ký tự đó là {A, E, O}, {O, U} and {C, G, K, N, P, T}. Phương pháp đề xuất là sự kết hợp của năm lớp SVM, trong đó các mô hình nhận dạng tay trái phụ thuộc vào nhận dạng kết quả của tay phải.

2.3. Kết quả thực nghiệm

Trong thử nghiệm, tôi phát triển hệ thống dựa trên ngôn ngữ lập trình C# và Accord.NET Framework. Tập dữ liệu đầu tiên có tên là Accent (trọng âm) bao gồm 03 động tác (613 ảnh) tương ứng với ba điểm nhấn bao gồm dấu mũ, dấu mũ ngược, dấu móc (hình 15.a) và 23 động tác (4637 hình ảnh) tương ứng với 23 ký tự chữ cái tiếng Việt (hình 15.b). Tất cả dữ liệu được thu bởi máy ảnh chiều sâu của Kinect.

Các thử nghiệm được kiểm tra với năm mô hình được mô tả ở trên, trong đó mỗi mô hình được kiểm tra với những kích cỡ khác nhau của việc chia ma trận xếp hạng. Các kết quả được thể hiện như Bảng 1.

Bảng 1: Độ chính xác khi thử nghiệm 5 mô hình với 5 cách chia ma trận

SVM \ ROCM	2×2	3×3	4×4	5×5	6×6
Model1	42.44 %	92.95 %	**94.22** %	94.16 %	48.76 %
Model2	63.90 %	**98.27** %	96.96 %	88.49 %	62.01 %
Model3	89.20 %	**99.67** %	96.18 %	75.91 %	56.64 %
Model4	99.01 %	**100.0 %**	**100.0 %**	90.57 %	74.44 %
Model5	71.85 %	**99.01 %**	94.45 %	86.09 %	63.66 %

Với mô hình 1, SVM phân loại thu được độ chính xác cao nhất là 94.22% tương ứng với ma trận xếp hạng 4 x 4, nhưng đối với mô hình 2-5 thì độ chính xác cao nhất thuộc về giá trị 3 x 3. Từ kết quả này ta nhận thấy rằng việc phân chi ma trận ô trên mỗi hình ảnh cử chỉ sẽ tương thích để thu được kết quả tốt nhất. Không có cách phân chia chung nào cho kết quả tốt nhất.

Tương tự đối với việc nhận dạng các ký tự số từ 0 đến 9 dữ liệu bao gồm 2011 mẫu bao gồm 10 cử chỉ tay.

Bảng 2: Độ chính xác khi thử nghiệm 10 cử chỉ số với 5 cách chia ma trận

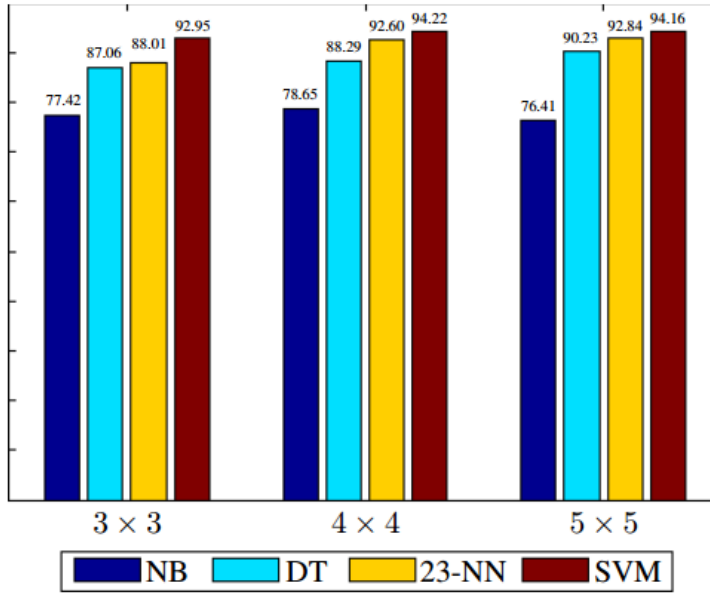
Kích cỡ ma trận	2 × 2	3 × 3	4 × 4	5 × 5	6 × 6
Tỉ lệ	52.61 (%)	**97.61 (%)**	95.72 (%)	80.46 (%)	61.66 (%)

Bên cạnh đó, đối với mô hình 1 gồm 23 ký tự đây là mô hình có số lượng dữ liệu lớn nhất, tôi tập trung phân tích, thực hiện các kỹ thuật phân loại khác nhau để so sánh và đánh giá hiệu quả. Các kỹ thuật được lựa chọn để phân loại học máy gồm: k-Nearest Neighbors (k-NN), decision tree (DT) và Naive Bayes (NB). Việc so sánh khi thử nghiệm 23 ký tự tương ứng với kỹ thuật "Xếp hạng ma trận" với các kích thước 3*3, 4*4, 5*5 và kết quả thể hiện trong hình 13, phương pháp SVM luôn cho kết quả tốt nhất.

2.4. Kết luận

Trong nghiên cứu cải tiến của phần này, tôi đề xuất một phương pháp mới để nhận dạng ngôn ngữ ký hiệu tiếng Việt dựa trên hình ảnh chiều sâu. Một kỹ thuật khai thác tính năng mới dựa trên xếp hạng các ô dựa trên lưới ô vuông được chia được đặt tên là ROCM – Rank Order Correlation Matrix để mô tả sự tương quan giữa các ô trong ảnh chiều sâu. Có hai đóng góp chính của tôi được sử dụng ở đây. Một là xây dựng

quá trình nhận dạng cử chỉ tay bao gồm bốn giai đoạn: phân đoạn, tiền xử lý, trích xuất đặc trưng và phân loại. Hai là xây dựng quy tắc để phân loại và nhận dạng bảng chữ cái ngôn ngữ ký hiệu tiếng Việt. Cụ thể, vị trí tay được phát hiện và thu nhận bằng cách áp dụng một bộ lọc khoảng cách trên hình ảnh chiều sâu thu được từ thiết bị Kinect. Các kích thước của hình ảnh bàn tay sau đó được chuẩn hóa về hình ảnh là hình vuông. Sau khi chia hình ảnh thành ma trận các ô vuông (2*2, 3*3, 4* hay 5*5) một vec-tơ đặc trưng được tạo ra bằng cách ghép vec-tơ giá trị trung bình và vec-tơ độ lệch tương ứng. Cuối cùng, sử dụng mô hình phân loại SVM đa lớp với chiến lược MAX-WIN để phân loại và nhận dạng. Cách tiếp cận của tôi đã cho kết quả với độ chính xác cao và có thể tích hợp để xử lý trong thời gian thực.



Hình 6: Độ chính xác của các kỹ thuật phân loại khác nhau

CHƯƠNG 3

NHẬN DẠNG CỬ CHỈ LIÊN TỤC

3.1. Tổng quan

Ngoài sự biểu diễn ngôn ngữ ký hiệu với các cử chỉ tĩnh để ghép thành các từ, cụm từ có ý nghĩa. Ngôn ngữ ký hiệu còn biểu diễn thông tin qua cử chỉ, điệu bộ, nét mặt thay cho lời nói. Tất cả các ngôn ngữ kí hiệu trên thế giới đều có 5 phương tiện và cách thức biểu hiện sau:

1. Vị trí của bàn tay.
2. Hình dạng bàn tay.
3. Hướng của lòng bàn tay.
4. Hướng của chuyển động lòng bàn tay.
5. Biểu hiện của nét mặt.

Nghiên cứu phần này hướng đến xử lý ngôn ngữ ký hiệu liên tục (động) trong thời gian thực, hay nói cách khác hướng đến nhận dạng từ vựng của ngôn ngữ ký hiệu tiếng Việt. Không giống như ngôn ngữ ký hiệu ở dạng tĩnh đã có những mức thành công nhất định, xử lý nhận dạng từ vựng ngôn ngữ ký hiệu liên tục khá phức tạp. Từ vựng trong ngôn ngữ ký hiệu Tiếng Việt bao gồm nhiều cử chỉ phức tạp như: hành động cánh tay, hình dạng bàn tay, các ngón tay, khẩu hình miệng, cảm xúc khuôn mặt,… Khác với cách biểu diễn ngôn ngữ bằng bảng chữ cái, các từ ngữ trong từ điển ngôn ngữ ký hiệu tiếng Việt rất đa dạng và phong phú.

Công cụ sử dụng trong thu nhận dữ liệu đầu vào là Camera Kinect v2 gồm: camera màu, camera hồng ngoại, và một dãy microphone gồm 4 microphone. Camera màu có thể ghi lại 30 frame ảnh RGB với độ phân giải 1920 x 1080 trong 1 giây. Camera màu cũng có thể lưu ảnh dưới dạng Raw Bayer, YUV và ảnh xám 16 bit.

Cảm biến chiều sâu có thể ghi lại 30 frame ảnh với độ phân giải 512 x 424 mỗi giây, góc nhận diện giới hạn được mở rộng 70 bề ngang và 60 bề dọc. Khoảng cách giới hạn của camera chiều sâu mặc định từ 0.5 mét đến 4.5 mét và có thể được sử dụng trong chế độ gần từ 0.4 mét đến 3 mét. Khoảng cách hoạt động tốt nhất của cảm biến là từ 1.2 mét đến 3.5 mét.

Trong phạm vi nghiên cứu, tính năng theo dõi chuyển động khung xương của Kinect SDK được sử dụng. SDK có thể xử lý dữ liệu thô đến từ camera chiều sâu và camera màu để bắt chuyển động khung xương con người. Với Kinect v2, ta có thể bắt được 6 khung xương người trong cùng một thời điểm và theo dõi 25 điểm tương ứng với các vị trí quan trọng của bộ phận cơ thể. Các vị trí được tính toán tương đối với cảm biến của thiết bị trong hệ tọa độ Đề-Các (x,y,z).

3.2. Quy trình nhận dạng cử chỉ liên tục

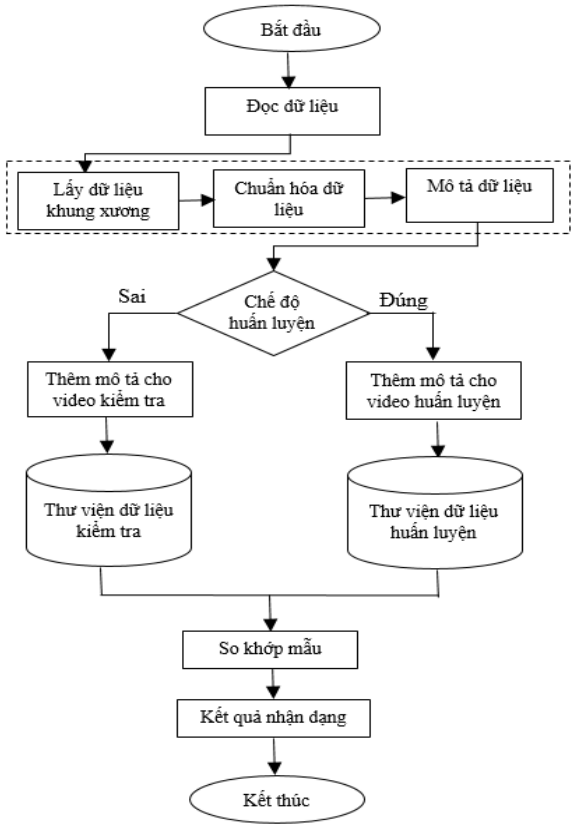
Sơ đồ quy trình nhận dạng cử chỉ liên tục được trình bày trong hình 7, bao gồm các bước cơ bản như sau : Đọc dữ liệu, Trích xuất đặc trưng, so khớp phân loại và nhận dạng.

3.2.1. Đọc dữ liệu

Mặc dù Kinect v2 có thể nhận biết được 25 vị trí khớp trong khung xương nhưng sau khi khảo sát từ điển ngôn ngữ ký hiệu tiếng Việt, chúng tôi kết luận rằng chuyển động của đôi tay là yếu tố quan trọng nhất, các thành phần khác của khuôn mặt như khẩu hình miệng hay chuyển động mắt không được sử dụng. Do đó, chúng tôi chỉ sử dụng 4 điểm liên quan đến tay gồm 2 điểm bàn tay trái và phải, 2 điểm khuỷu tay trái và phải.

Dữ liệu khung xương được thu bởi Kinect với tốc độ 30 khung

hình mỗi giây. Tuy vậy, hệ thống mà chúng tôi xây dựng chỉ chọn và xử lý 5 khung xương trong số đó. Do đó, việc thu nhận dữ liệu được thực hiện cứ sau mỗi 0.2 giây. Cụ thể, cứ thu được 6 khung hình thì hệ thống tiến hành tính khung xương trung bình và đưa vào mô-đun nhận dạng. Lưu ý rằng mỗi khung hình được thu nhận sẽ được kiểm tra có chứa các thành phần bàn tay, khuỷu tay và tâm cơ thể hay không. Nếu có điểm bất kỳ không được thu nhận, hệ thống sẽ tự động điền thông tin đó bằng dữ liệu từ khung hình trước.

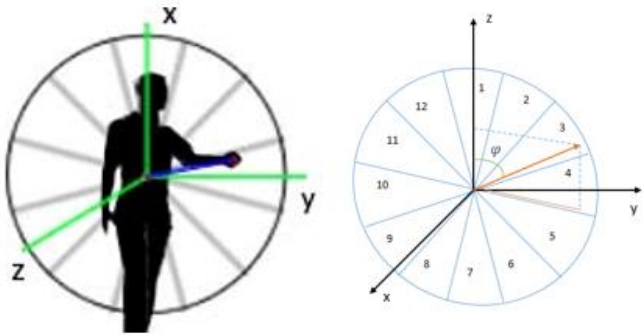


Hình 7: Sơ đồ hoạt động của hệ thống nhận dạng cử chỉ liên tục

3.2.2. Trích xuất đặc trưng

Công việc chính ở giai đoạn này là chuyển thông tin khung xương ở hệ tọa độ Đề-Các sang hệ tọa độ cầu. Camera Kinect v2 với cảm biến chiều sâu cho phép làm việc với dữ liệu chiều sâu của đối tượng. Do đó, ta có thể sử dụng dữ liệu 3D để xử lý ngôn ngữ ký hiệu tiếng Việt. Thông tin về khung xương đã đề cập ở trên có thể biểu diễn trong hệ tọa độ Đề-Các với 3 thông số (x, y, z). Tuy nhiên, phương pháp này bộc lộ nhược điểm là chỉ có thể sử dụng dữ liệu trong trường hợp vị trí và khoảng cách của đối tượng với camera Kinect là không thay đổi. Do đó, ta cần phải đổi hệ quy chiếu từ máy quay sang hệ quy chiếu của đối tượng: lấy tâm người làm gốc tọa độ, các dữ liệu về bàn tay và khuỷu tay được quy về theo hệ tọa độ này.

Trong toán học, một hệ tọa độ cầu Spherical là một hệ tọa độ cho không gian 3 chiều mà vị trí một điểm được xác định bởi 3 số: khoảng cách theo hướng bán kính từ gốc tọa độ r, góc nâng từ điểm đó từ một mặt phẳng cố định θ, và góc kinh độ của hình chiếu vuông góc của điểm đó lên mặt phẳng cố định đó φ.



Hình 8: Chia vùng chuẩn hóa dữ liệu góc kinh độ φ

Dữ liệu ban đầu đưa vào là dữ liệu số thực ở hệ tọa độ Đề-Các, chúng ta chuyển chúng sang hệ tọa độ cầu với tâm là tâm cơ thể của đối tượng. Đối với góc θ và φ ta chia thành 12 góc nhỏ với mỗi góc $30°$ (hình 8). Với bán kính r, ta sẽ nhân với 10 và lấy phần nguyên của nó (vì dữ liệu thô tính bằng đơn vị mét). Giải thích về việc chuẩn hóa dữ liệu này nhằm đồng bộ dữ liệu Training và Test cũng như để loại bỏ nhiễu không cần thiết. Như vậy sau quá trình chuẩn hóa dữ liệu, dữ liệu đưa vào sẽ bao gồm các số nguyên.

Sau khi chuẩn hóa dữ liệu, việc tiếp theo chúng ta phải mô tả dữ liệu đã được chuẩn hóa. Chúng ta sẽ có một vec-tơ gồm 12 phần tử chứa dữ liệu của 4 điểm tại một thời điểm.

$$\vec{J} = \{r_{LE}, \theta_{LE}, \varphi_{LE}, r_{RE}, \theta_{RE}, \varphi_{RE}, r_{LH}, \theta_{LH}, \varphi_{LH}, r_{RH}, \theta_{RH}, \varphi_{RH}\}$$

Dữ liệu sẽ là một mảng các vec-tơ $\vec{J}$ ở mỗi thời điểm khác nhau.

$$Data = \{\vec{J_1}, \vec{J_2}, \vec{J_3}, \dots, \vec{J_n}\}$$

Các dữ liệu huấn luyện sẽ được lưu vào một file và sẽ được gán nhãn với mỗi từ ngữ của ngôn ngữ ký hiệu.

3.2.3. Phân loại

Dynamic Time Warping (DTW) là thuật toán dùng để tính độ tương đồng giữa hai chuỗi đặc trưng, khác nhau về chiều dài. DTW được áp dụng trong nhiều lĩnh vực và là ý tưởng nguyên thuỷ của nhận dạng tiếng nói. Mục đích của DTW dùng để tìm kiếm một ánh xạ giữa hai vec-tơ đặc trưng (có chiều dài khác nhau) với khoảng cách ngắn nhất.

Thuật toán láng giềng gần với hệ số k (k-Nearest Neighbors (kNN)) được sử dụng rất phổ biến trong lĩnh vực khai phá dữ liệu. kNN

là phương pháp để phân lớp các đối tượng dựa vào khoảng cách gần nhất giữa đối tượng cần xếp lớp với tất cả các đối tượng trong dữ liệu huấn luyện.

Một đối tượng được phân lớp dựa vào k láng giềng của nó. K là số nguyên dương được xác định trước khi thực hiện thuật toán. Người ta thường dùng khoảng cách Euclidean để tính khoảng cách giữa các đối tượng.

Cách thực áp dụng thuật toán kNN để tìm nhãn của cử chỉ trong đề tài là tìm k vec-tơ mô tả cử chỉ trong mỗi một lớp cử chỉ và gần nhất với cử chỉ đưa vào nhận dạng dựa trên khoảng cách DTW. Tính khoảng cách trung bình của k vec-tơ đó và coi đó là khoảng cách của mẫu cử chỉ đưa vào với lớp cử chỉ. Với 10 lớp cử chỉ, ta tìm lớp nào có khoảng cách đến mẫu đưa vào là nhỏ nhất và coi đó là lớp của cử chỉ cần nhận dạng.

Đề xuất của nghiên cứu là xây dựng một cải tiến của phương pháp phân loại kNN kết hợp với thuật toán DTW (kNN-DTW) được xem như là một hàm chi phí. Khi thu nhận một dữ liệu kiểm tra, hệ thống sẽ phân loại và xếp hạng dữ liệu đầu vào với tập dữ liệu có k gần nhất. Để kiểm tra sự chắc chắn ta tiếp tục sử dụng DTW để so khớp và đưa ra kết quả nhận dạng. Dữ liệu đưa vào gồm 2 phần chính là dữ liệu khuỷu tay và dữ liệu bàn tay trong cùng 1 mảng vec-tơ.

3.3. Kết quả

Phương pháp được thử nghiệm với 10 từ trong bộ từ điển Ngôn ngữ ký hiệu Tiếng Việt. Mỗi từ được lấy 30 mẫu bao gồm 20 mẫu training và 10 mẫu test. Dữ liệu được phân loại bằng thuật toán DTW và phương pháp phân cụm Nearest Neighbor với trọng số 80% cánh tay, 20% khuỷu tay. Cấu hình hệ thống: Windows 8 Profesional, Intel Core i5 2.5GHz, RAM 4G, Kinect v2 for Windows. Hệ thống hoạt động cho

ra kết quả trong thời gian thực:

Bảng 3: Kết quả nhận dạng ngôn ngữ ký hiệu tiếng Việt

TỪ	KẾT QUẢ
Buổi sáng	90%
Bàn hội nghị	85%
Bánh chưng	95%
Cầu vượt	90%
Giao thông	95%
Ấm áp	90%
Ăn mặc	80%
Thành phố	95%
Biểu quyết	100%
Tình nguyện	100%

Hệ thống làm việc thời gian thực ổn định, các kết quả chính xác đến 92%. Với bộ thư viện dữ liệu nhỏ (khoảng 20 mẫu), thuật toán DTW có thể xử lý nhanh chóng và đưa ra kết quả ngay lập tức. Nhược điểm của thuật toán này là với bộ dữ liệu lớn hơn hệ thống sẽ trở nên quá tải. Hơn nữa, các dấu hiệu về hình dáng bàn tay, biểu cảm khuôn mặt, khẩu hình miệng bị lược bỏ trong thực tế cũng khá quan trọng để nhận dạng ngôn ngữ ký hiệu tiếng Việt. Tuy vậy, nghiên cứu đã nêu ra được phương pháp giải quyết ngôn ngữ ký hiệu tiếng Việt trong thời gian thực.

Để hệ thống có thể hoạt động tốt hơn cần phải bổ sung thêm các tính năng về nhận diện hình dáng bàn tay. Ngoài ra việc xử lý thời gian thực với nguồn thư viện lớn cũng phải được xem xét.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong nghiên cứu này, đối với cử chỉ tĩnh tôi đề xuất một phương pháp mới để nhận dạng ngôn ngữ ký hiệu tiếng Việt dựa trên hình ảnh chiều sâu. Một kỹ thuật khai thác tính năng mới dựa trên xếp hạng các ô dựa trên lưới ô vuông được chia được đặt tên là ROCM – Rank Order Correlation Matrix để mô tả sự tương quan giữa các ô trong ảnh chiều sâu. Có hai đóng góp chính của tôi được sử dụng ở đây. Một là xây dựng quá trình nhận dạng cử chỉ tay bao gồm bốn giai đoạn: phân đoạn, tiền xử lý, trích xuất đặc trưng và phân loại. Hai là xây dựng quy tắc để phân loại và nhận dạng bảng chữ cái ngôn ngữ ký hiệu tiếng Việt. Cụ thể, vị trí tay được phát hiện và thu nhận bằng cách áp dụng một bộ lọc khoảng cách trên hình ảnh chiều sâu thu được từ thiết bị Kinect. Các kích thước của hình ảnh bàn tay sau đó được chuẩn hóa về hình ảnh là hình vuông. Sau khi chia hình ảnh thành ma trận các ô vuông (2*2, 3*3, 4*4 hay 5*5) một vec-tơ đặc trưng được tạo ra bằng cách ghép vec-tơ giá trị trung bình và vec-tơ độ lệch tương ứng. Cuối cùng, sử dụng mô hình phân loại SVM đa lớp với chiến lược MAX-WIN để phân loại và nhận dạng.

Đối với cử chỉ liên tục, tôi đề xuất một phương pháp thu nhận dữ liệu cho các cử động của ngôn ngữ ký hiệu tiếng Việt được trên dữ liệu khung xương thu nhận từ Kinect để nhận dạng. Thay đổi hệ tọa độ phụ thuộc và vị trí người thực hiện so với thiết bị sang vị trí tương đối so với trọng tâm con người để khắc phục ảnh hưởng của vị trí. Cuối cùng, sử dụng mô hình kNN kết hợp với DTW phân loại và nhận dạng. Cách tiếp cận của tôi đã cho kết quả với độ chính xác cao và có thể tích hợp để xử lý trong thời gian thực.

Tuy nhiên nhược điểm của thuật toán này là với bộ dữ liệu lớn

hơn hệ thống sẽ trở nên quá tải. Hơn nữa, các dấu hiệu về hình dáng bàn tay, biểu cảm khuôn mặt, khẩu hình miệng bị lược bỏ trong thực tế cũng khá quan trọng để nhận dạng ngôn ngữ ký hiệu tiếng Việt. Để hệ thống có thể hoạt động tốt hơn cần phải bổ sung thêm các tính năng về nhận diện hình dáng bàn tay, khẩu hình miệng. Ngoài ra việc xử lý thời gian thực với nguồn dữ liệu lớn cũng phải được xem xét.

Hướng nghiên cứu trong thời gian tiếp theo để ghi nhận ngôn ngữ ký hiệu:

- Xây dựng bộ cơ sở dữ liệu hoàn chỉnh cho nhận dạng ngôn ngữ ký hiệu tiếng Việt.
- Nghiên cứu về phân đoạn video để loại bỏ nhiễu và tăng tỉ lệ thành công khi nhận dạng.
- Tập trung vào nghiên cứu, cải tiến thuật toán để nâng cao kết quả nhận dạng với cử chỉ động với dữ liệu lớn. Hệ thống sẽ kết hợp nhận dạng khuôn mặt, bàn tay (phải/trái) và các bộ phận khác của cơ thể cùng một lúc.