Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00082

NHẬN DẠNG NGÔN NGỮ DẤU HIỆU SỬ DỤNG CẢM BIẾN KINECT

Võ Hồng Khanh[1], Phạm Nguyên Khang[1]

[1]Trường Đại học Cần Thơ

hongkhanh@ctu.edu.vn, pnkhang@cit.ctu.edu.vn

TÓM TẮT: *Trong bài báo này, chúng tôi đề xuất một phương pháp nhận dạng ngôn ngữ dấu hiệu tiếng Việt (dành cho người khiếm thính) sử dụng cảm biến Kinect và máy học véctơ hỗ trợ. Dữ liệu thu được từ cảm biến Kinect được trích đặc trưng và sử dụng để huấn luyện bộ nhận dạng dùng mô hình máy học véctơ hỗ trợ. Chúng tôi đề xuất 6 phương pháp biểu diễn đặc trưng của các từ trong ngôn ngữ dấu hiệu dựa trên sự kết hợp giữa 3 phương pháp biểu diễn quỹ đạo bàn tay và 2 phương pháp biểu diễn hình dáng của bàn tay với đặc trưng GIST và HOG. Các thực nghiệm được thực hiện trên tập dữ liệu gồm 50 từ thuộc 6 chủ đề khác nhau, với tổng cộng là 2,400 mẫu dữ liệu. Kết quả cho thấy hệ thống nhận dạng đạt kết quả 99.46%.*

Từ khóa: *Người khiếm thính, Ngôn ngữ dấu hiệu, Nhận dạng ngôn ngữ dấu hiệu, camera Kinect, GIST, HOG.*

I. GIỚI THIỆU

Hỗ trợ người khuyết tật là vấn đề được xã hội rất quan tâm và Công nghệ thông tin là một trong những lĩnh vực tiên phong, góp phần giúp việc tiếp cận thông tin, cũng như giao tiếp giữa người khuyết tật với cộng đồng được thuận lợi hơn. Thông qua các chính sách cụ thể, Nhà nước ta đã và đang rất nỗ lực để giúp người khuyết tật dễ dàng hòa nhập với cộng đồng. Thông tư 28 của Bộ Thông tin và Truyền thông có hiệu lực từ ngày 02/11/2009 "Quy định việc áp dụng các tiêu chuẩn, công nghệ hỗ trợ người khuyết tật tiếp cận, sử dụng công nghệ thông tin và truyền thông" [2].

Với 78,5 triệu dân, Việt Nam có 6,7 triệu người là người khuyết tật (chiếm khoảng 7,8% dân số), trong đó số lượng người khiếm thính khoảng 3 triệu người [5]. Để giao tiếp với cộng đồng, người khiếm thính sử dụng ngôn ngữ cử chỉ. Tuy nhiên việc giao tiếp rất khó khăn vì phần lớn những người bình thường không hiểu được ngôn ngữ dấu hiệu. Với mục đích hỗ trợ cho người khiếm thính, rất nhiều nghiên cứu liên quan đến nhận dạng ngôn ngữ dấu hiệu đã được đề xuất. Nhận dạng tự động ngôn ngữ dấu hiệu là bước cốt lõi khi phát triển hệ thống tương tác người-máy dành cho người khiếm thính. Hệ thống nhận dạng ngôn ngữ dấu hiệu cho phép người khiếm thính điều khiển máy tính bằng ngôn ngữ dấu hiệu và thuận lợi hơn khi giao tiếp với cộng đồng.

Trong bài báo này, chúng tôi sử dụng dữ liệu thu được từ cảm biến Kinect và mô hình máy học véctơ hỗ trợ để phục vụ cho việc nhận dạng ngôn ngữ dấu hiệu. Chúng tôi đề xuất 6 phương pháp biểu diễn đặc trưng của các từ trong ngôn ngữ dấu hiệu. Dựa vào kết quả thực nghiệm, chúng tôi lựa chọn phương pháp có độ chính xác nhận dạng cao nhất.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Để giải quyết bài toán nhận dạng ngôn ngữ cử chỉ, phương pháp phổ biến và ra đời sớm nhất là sử dụng thiết bị thu ảnh hai chiều kết hợp với giải thuật xử lý ảnh. Năm 1994, Davis J. và Shah M. xây dựng hệ thống nhận dạng các tư thế bàn tay cơ bản và đạt độ chính xác 94% [8] dựa vào mô hình phân lớp FSM (Finite State Machine). Năm 2010, Steinberg I., London T. M., Castro D. D. xây dựng hệ thống "Hand gesture recognition in images and video" có độ chính xác nhận dạng là 97,8% [12] dựa vào mô hình phân lớp SVM (Support Vector Machine). Tại Việt Nam, năm 2015, nhóm tác giả Dương Khắc Hưởng, Nguyễn Đăng Bình Trường sử dụng mô hình Markov ẩn để xây dựng hệ thống "nhận dạng ngôn ngữ cử chỉ thông qua quỹ đạo chuyển động liên tục của đối tượng" với độ chính xác 98.4% [3]. Đặc điểm chung của các hệ thống nhận dạng dựa vào phương pháp này là chỉ đạt độ chính xác cao khi hình ảnh được thu nhận trong điều kiện ánh sáng tốt. Độ chính xác nhận dạng của phương pháp này rất thấp khi ảnh được thu trong điều kiện thiếu sáng hoặc màu da tiệp màu với màu áo và màu nền.

Phương pháp cho độ chính xác cao nhất khi giải quyết bài toán này là sử dụng găng tay cảm biến. Năm 2014, Gowri. D, Vidhubala. D xây dựng hệ thống "Sign language recognition for deaf and dumb people" và tự thiết kế bộ găng tay cảm biến để nhận dạng 26 chữ cái tiếng Anh trong ngôn ngữ ASL (American Sign Language) với độ chính xác nhận dạng là 99% [9]. Tuy nhiên vì giá thành đắt đỏ của loại thiết bị này nên khi nghiên cứu chủ đề nhận dạng cử chỉ cần sử dụng găng tay cảm biến, đa số các nhóm nghiên cứu đều sử dụng găng tay cảm biến tự tạo.

Năm 2009, sự ra đời của thiết bị Microsoft Kinect mở ra một hướng nghiên cứu mới để giải quyết bài toán nhận dạng ngôn ngữ cử chỉ. Với giá thành vừa phải, khả năng ứng dụng loại thiết bị này vào các nghiên cứu là khả thi và do đó các nghiên cứu về việc ứng dụng cảm biến thu ảnh ba chiều để nhận dạng ngôn ngữ cử chỉ tăng nhanh chóng về số lượng. Năm 2014, Murata T., Shin J. xây dựng hệ thống "Hand gesture and character recognition based on Kinect sensor" thực hiện nhận dạng chữ số và các chữ cái tiếng Anh dựa vào cảm biến Kinect, kết quả nhận dạng được hiển thị trên thiết bị Palm's Grafiti với độ chính xác 96,9% [11]. Năm 2015, Cao D., Ming C. L., Zhaozheng Y. xây dựng hệ thống "American Sign Language alphabet recognition using Microsoft Kinect" nhận dạng 26 chữ cái trong ngôn ngữ ASL và đạt độ chính xác 92% [6]. Tại Việt Nam, ngày 19/03/2013, Ban chủ nhiệm Chương trình KC.01/11-15 phối

hợp với Văn phòng các Chương trình trọng điểm cấp Nhà nước tổ chức nghiệm thu cấp nhà nước đề tài KH&CN tiềm năng "Nghiên cứu phát triển kỹ thuật nhận dạng cử động của bàn tay người theo thời gian thực" (Mã số: KC.01.TN08) do TS. Trần Nguyên Ngọc - Học viện Kỹ thuật Quân sự làm chủ nhiệm đề tài với độ chính xác của hệ thống nhận dạng là 90.04% [1]. Năm 2015, Phạm Nguyên Khang, Huỳnh Nhật Minh, Võ Trí Thức, Phạm Thế Phi xây dựng hệ thống "Nhận dạng ngôn ngữ dấu hiệu với cảm biến Kinect và đặc trưng GIST" với độ chính xác nhận dạng là 90% [4]. Nhìn chung, phương pháp sử dụng cảm biến Kinect để nhận dạng ngôn ngữ cử chỉ dễ dàng tiếp cận hơn so với phương pháp sử dụng găng tay cảm biến. Bên cạnh đó, dựa vào khả năng thu nhận ảnh hồng ngoại và khả năng tách nền được hỗ trợ sẵn của cảm biến Kinect, hệ thống nhận dạng đạt độ chính xác cao khi thu ảnh trong bóng tối, khi màu da trùng với màu quần áo và màu nền.

Nhìn chung, chủ đề nhận dạng ngôn ngữ cử chỉ đã được nghiên cứu nhiều. Tuy nhiên, việc ứng dụng cảm biến Kinect để nhận dạng ngôn ngữ cử chỉ và đặc biệt là "nhận dạng ngôn ngữ cử chỉ tiếng Việt" thì chưa nhiều đối với cả trong và ngoài nước.

III. ĐỀ XUẤT PHƯƠNG PHÁP BIỂU DIỄN NGÔN NGỮ CỬ CHỈ

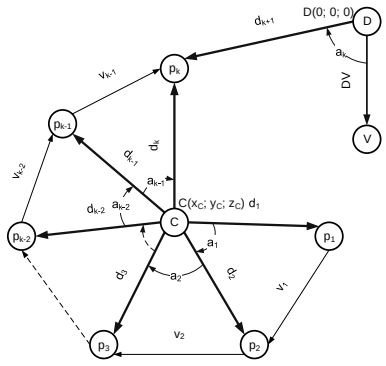
Khi sử dụng ngôn ngữ cử chỉ, để diễn tả một từ thì cần sử dụng một tay hoặc cả hai tay và đôi khi cần kết hợp với biểu cảm của khuôn mặt. Khuôn mặt của mỗi người đa số là khác nhau, nên việc tổng quát hóa khuôn mặt người và trích đặc trưng sẽ làm tăng độ phức tạp của bài toán. Vì vậy, chúng tôi đề xuất không thực hiện trích đặc trưng khuôn mặt, mà chỉ tập trung vào hai việc là biểu diễn quỹ đạo chuyển động hai tay và biểu diễn đặc trưng tư thế hai bàn tay.

A. Biểu diễn quỹ đạo chuyển động của hai tay

Quỹ đạo chuyển động của hai bàn tay được thể hiện thông qua một số đặc trưng như: hình dáng của quỹ đạo, hướng di chuyển, vận tốc di chuyển của bàn tay và vị trí của bàn tay so với đầu của người dùng. Dựa vào dữ liệu ảnh ba chiều thu được từ cảm biến Kinect và khả năng dựng khung xương được hỗ trợ sẵn, tư thế của người dùng được xác định bao gồm vị trí của tay trái, tay phải, đầu, vai cho mỗi khung hình. Vị trí các khớp xương sẽ được biểu diễn lại theo vị trí đầu của người dùng. Việc biểu diễn khung xương theo vị trí tương đối, giúp vị trí các khớp xương không phụ thuộc vào gốc tọa độ. Tuy nhiên, số lượng khung hình dùng để biểu diễn cho mỗi từ là khác nhau. Vì vậy, cần đề xuất phương pháp biểu diễn sao cho đặc trưng không phụ thuộc vào số lượng khung hình của một từ [4]. Để số hóa quỹ đạo chuyển động của bàn tay, quỹ đạo chuyển động của bàn tay được mô tả bằng n điểm P=(p_1, p_2, …, p_n). Ứng với mỗi từ khác nhau, n có thể rất lớn và khác nhau. Do đó, để đặc trưng của quỹ đạo không lệ thuộc vào n, quỹ đạo được chia thành thành k đoạn [4].

Giả sử chọn k = 15, mỗi đoạn lấy một điểm đại diện, sau đó tính các đặc trưng, 3 phương pháp biểu diễn quỹ đạo chuyển động của bàn tay được đề xuất như sau:

B. Phương pháp 1 (PP1)



Hình 1. Phương pháp 1 - trích đặc trưng quỹ đạo bàn tay

Trong phương pháp 1, các đặc trưng của quỹ đạo chuyển động của bàn tay cần lưu trữ bao gồm:

(1) Độ dài các véctơ d_i

(2) Góc giữa các véctơ d_i

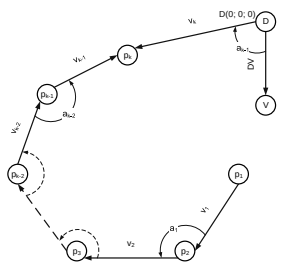
(3) Tỷ lệ giữa độ dài véctơ d_{k+1} và khoảng cách đầu-vai.

(4) Góc giữa véctơ d_{k+1} và véctơ đầu-vai.

Với phương pháp này, khi k = 15 thì véctơ đặc trưng thu được có:

k + k – 1 + (k – 1)*3 + 1 + 1 + 3 = 76 chiều.

C. Phương pháp 2 (PP2)



Hình 2. Phương pháp 2 - trích đặc trưng quỹ đạo bàn tay

Trong phương pháp 2, các đặc trưng của quỹ đạo chuyển động của bàn tay cần lưu trữ bao gồm:

(1) Độ dài các véctơ v_i

(2) Góc giữa các véctơ v_i

(3) Tỷ lệ giữa độ dài véctơ v_k và khoảng cách đầu-vai.

(4) Góc giữa véctơ v_k và véctơ đầu-vai.

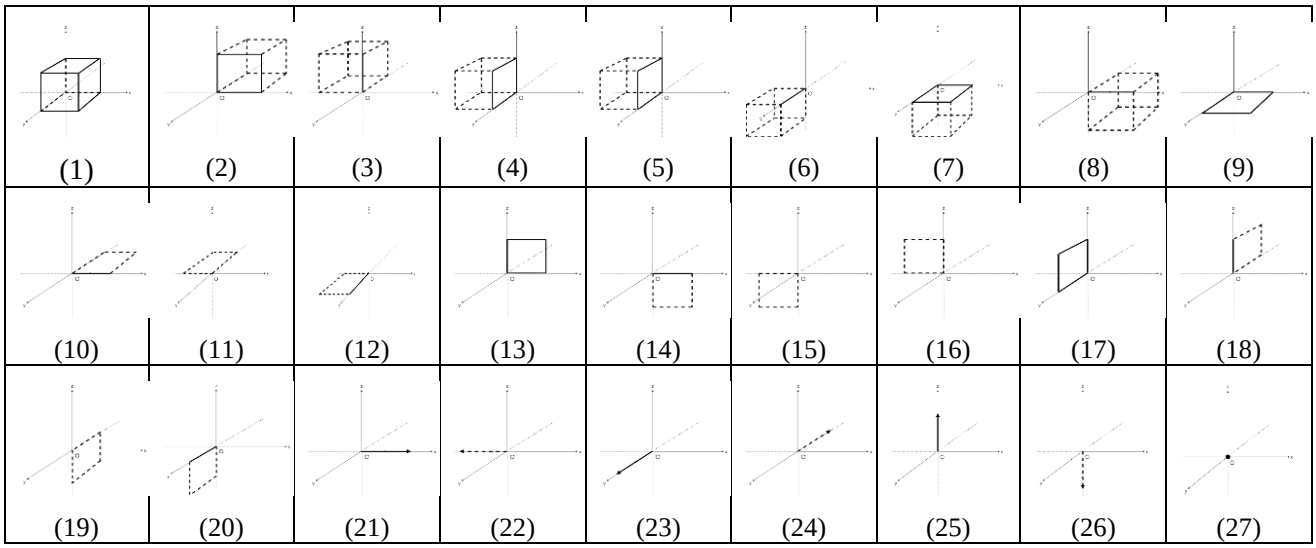
Với phương pháp này, khi k = 15 thì véctơ đặc trưng thu được có tổng cộng:

k – 1 + k – 2 + (k – 2)*3 + 1 + 1 + 3 = 71 chiều.

D. Phương pháp 3 (PP3)

Phương pháp này giống như phương pháp 2, hướng của các vector v_{i+1} so với véctơ v_i được lưu lại thay vì lưu góc giữa các véctơ. Hướng của véctơ sẽ tùy thuộc vào phần không gian mà véctơ thuộc về. Không gian Oxyz thành được chia thành 27 phần.

Bảng 1. Phương pháp 3 - không gian Oxyz được chia ra thành 27 phần



Với phương pháp này, khi k = 15 thì véctơ đặc trưng thu được có: k – 1 + k – 2 + k – 2 + 1 + 1 + 1 = 43 chiều.

E. Biểu diễn đặc trưng tư thế của hai bàn tay

Trong ngôn ngữ cử chỉ, khi biểu diễn một từ thì tư thế bàn tay ở thời điểm bắt đầu, ở giữa và kết thúc là có khả năng phân biệt cao nhất. Do đó, đặc trưng tư thế bàn tay sẽ được trích ở ba thời điểm là thời điểm đầu, thời điểm giữa và thời điểm cuối chuỗi chuyển động của bàn tay để mô tả một từ [4]. Chúng tôi sẽ so sánh độ chính xác của các mô hình phân lớp khi lần lượt sử dụng một trong hai phương pháp rút trích đặc trưng ảnh GIST và HOG.

F. Tách chọn hình ảnh bàn tay

Thiết bị Microsoft Kinect cung cấp nhiều kênh dữ liệu như: màu (Color), hồng ngoại (Infrared), độ sâu (Depth), khung xương (Body), nhận dạng người dùng (BodyIndex). Kết quả khảo sát cho thấy thiết bị hỗ trợ rất tốt cho nghiên cứu:

- Hình ảnh thu được dựa vào kênh hồng ngoại không bị lệ thuộc vào nguồn sáng, ảnh thu được có chất lượng tốt khi thu hình trong điều kiện đủ sáng (ban ngày, có đèn,…) và trong điều kiện ánh sáng yếu (trong bóng tối,…).
- Dữ liệu độ sâu và dữ liệu khung xương giúp phân tách các bộ phận trên cơ thể ra khỏi nền ảnh dễ dàng. Việc tách ảnh bàn tay ra khỏi nền ảnh không gặp khó khăn khi màu nền giống với màu da, màu quần áo.

Từ các nhận định trên, ảnh hai bàn tay được tách dựa vào ba kênh dữ liệu của thiết bị Kinect là: dữ liệu khung xương, độ sâu và hồng ngoại.

G. Biểu diễn đặc trưng tư thế bàn tay dựa vào đặc trưng ảnh GIST

Ảnh mức xám của bàn tay nhận được dựa vào quá trình phân tách và sau đó lần lượt thực hiện các bước sau để trích đặc trưng GIST của ảnh:

Bước 1. Áp dụng phép biến đổi Fourier trên ảnh này.

Bước 2. Áp dụng 20 bộ lọc Gabor lên ảnh kết quả ở bước 1, trong đó, bộ lọc Gabor được tạo ra ở 3 thang (scales) và 8 hướng, thang 1 và thang 2 sử dụng 8 bộ lọc và thang 3 sử dụng 4 bộ lọc.

Bước 3. Thực hiện biến đổi Fourier ngược kết quả nhận ở bước 2.

Bước 4. Chia ảnh kết quả ở bước 3 thành 4 * 4 = 16 ô bằng nhau và tiến hành trích đặc trưng.

Sau khi thực hiện các bước như trên, véctơ đặc trưng ảnh nhận được có tổng cộng (8 + 8 + 4) * 16 = 320 chiều.

Với mỗi bàn tay, đặc trưng tư thế bàn tay được trích ở ba thời điểm: đầu, giữa và cuối chuỗi chuyển động của bàn tay. Do đó, véctơ biểu diễn đặc trưng một bàn tay có tổng cộng 3 * 320 = 960 chiều.

H. Biểu diễn đặc trưng tư thế bàn tay dựa vào đặc trưng ảnh HOG

Để trích đặc trưng HOG ảnh mức xám của bàn tay, cần thực hiện lần lượt các bước sau:

Bước 1. Tính cường độ sáng và hướng biến thiên tại mỗi điểm ảnh.

Bước 2. Vì ảnh đầu vào chỉ chứa duy nhất ảnh bàn tay nên chỉ có ~~1~~một khối và khối này cũng là toàn bộ ảnh. Khối được chia ra thành 4 * 4 = 16 ô bằng nhau.

Bước 3. Tính véctơ đặc trưng cho khối:

- Signed-HOG được sử dụng, chia không gian hướng thành 18 hướng (bins), ~~s~~Sau đó, thực hiện tính toán và gom nhóm đặc trưng tại mỗi ô.
- Thực hiện ghép các véctơ đặc trưng ở từng ô để nhận được véctơ đặc trưng của khối.

Bước 4. Vì ảnh chỉ có một khối nên đặc trưng khối cũng là đặc trưng của toàn bộ ảnh.

Sau khi thực hiện các bước như trên, véctơ đặc trưng ảnh nhận được có tổng cộng 4 * 4 * 18 = 288 chiều.

Với mỗi bàn tay, đặc trưng tư thế bàn tay được trích ở ba thời điểm: đầu, giữa và cuối chuỗi chuyển động của bàn tay. Do đó, véctơ biểu diễn đặc trưng một bàn tay có tổng cộng 3 * 288 = 864 chiều.

I. Biểu diễn đặc trưng của một từ trong ngôn ngữ cử chỉ

Để tạo ra véctơ đặc trưng của một từ trong ngôn ngữ cử chỉ, thực hiện ghép các véctơ đặc trưng thể hiện quỹ đạo bàn tay, tư thế bàn tay. Thứ tự ghép các véctơ lần lượt như sau:

1. Véctơ biểu diễn đặc trưng quỹ đạo chuyển động của bàn tay trái.
2. Véctơ biểu diễn đặc trưng tư thế bàn tay trái.
3. Véctơ biểu diễn đặc trưng quỹ đạo chuyển động của bàn tay phải.
4. Véctơ biểu diễn đặc trưng tư thế bàn tay phải.

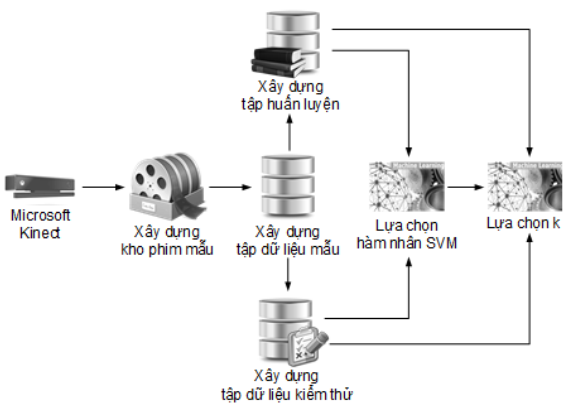
Kết hợp giữa ba phương pháp trích đặc trưng quỹ đạo bàn tay và hai phương pháp trích đặc trưng tư thế bàn tay, có tổng cộng 6 phương pháp biểu diễn đặc trưng của một từ trong ngôn ngữ chỉ. Với mỗi phương pháp, véctơ đặc trưng có số chiều là khác nhau. Cụ thể với k = 15, độ dài véctơ đặc trưng được xây dựng ở mỗi phương pháp là:

1. GIST1: kết hợp giữa PP1 và sử dụng đặc trưng ảnh GIST. Véctơ đặc trưng có: (76 + 960) * 2 = 2,072 chiều.
2. GIST2: kết hợp giữa PP2 và sử dụng đặc trưng ảnh GIST. Véctơ đặc trưng có: (71 + 960) * 2 = 2,062 chiều.
3. GIST3: kết hợp giữa PP3 và sử dụng đặc trưng ảnh GIST. Véctơ đặc trưng có: (43 + 960) * 2 = 2,006 chiều.
4. HOG1: kết hợp giữa PP1 và sử dụng đặc trưng ảnh HOG. Véctơ đặc trưng có: (76 + 864) * 2 = 1,880 chiều.
5. HOG2: kết hợp giữa PP2 và sử dụng đặc trưng ảnh HOG. Véctơ đặc trưng có: (71 + 864) * 2 = 1,870 chiều.
6. HOG3: kết hợp giữa PP3 và sử dụng đặc trưng ảnh HOG. Véctơ đặc trưng có: (43 + 864) * 2 = 1,814 chiều.

IV. QUY TRÌNH THỰC NGHIỆM

A. Xây dựng kho phim mẫu

Chúng tôi tự tìm hiểu và lựa chọn 50 từ trong ngôn ngữ cử chỉ tiếng Việt thuộc 6 chủ đề phổ biến, sau đó tiến hành quay phim lấy mẫu. Chúng tôi xây dựng ứng dụng quay phim để lưu trữ dữ liệu thu nhận từ thiết bị Microsoft Kinect và chỉ lưu lại các thông hữu dụng: khung xương, hồng ngoại, độ sâu, tọa độ các khớp xương thuộc bàn tay, vai, đầu và xương chậu. Sau đó, thực hiện loại bỏ nền ảnh và chỉ giữ lại ảnh hai bàn tay.



Hình 3. Quy trình thực nghiệm

B. Xây dựng tập dữ liệu mẫu

Chúng tôi khảo sát từng phim mẫu (chứa 10 mẫu mô tả một từ) và giữ lại 6 mẫu tốt nhất. Với mỗi mẫu, đường chuyển động của hai tay được phân tích, đặc trưng ảnh hai bàn tay và lưu lại các véctơ đặc trưng cho các trường hợp:

- Các tham số k khác nhau khi trích đặc trưng: thiết bị Mirosoft Kinect cung cấp dữ liệu với tốc độ 30 khung hình mỗi giây. Tuy nhiên, tốc độ ra dấu của mỗi người là khác nhau. Do đó, các tập dữ liệu mẫu được tạo ra ứng với các tham số k khác nhau, để khảo sát sự ảnh hưởng của tham số k đến độ chính xác của các mô hình. Các tham số k được chọn để khảo sát bao gồm: 5, 10, 15, 20, 25 và 30.

- Các phương pháp trích đặc trưng khác nhau: ở mỗi trường hợp của giá trị k, cần xây dựng 6 tập dữ liệu mẫu ứng với 6 phương pháp trích đặc trưng: GIST1, GIST2, GIST3, HOG1, HOG2 và HOG3.

C. Xây dựng tập dữ liệu huấn luyện và tập kiểm thử

Độ chính xác của mô hình được đánh giá theo phương pháp hold-out, nên tập dữ liệu mẫu được chia thành hai tập độc lập, tập huấn luyện chiếm $^2/_3$ dữ liệu và $^1/_3$ dữ liệu còn lại dành cho tập kiểm thử.

D. Khảo sát và lựa chọn hàm nhân SVM

Chúng tôi sử dụng thư viện LibSVMsharp [13] để huấn luyện mô hình và khảo sát 4 hàm nhân:

- LINEAR: tương ứng với hàm nhân Linear : $K(u, v) = u^T * v + c$
- POLY: tương ứng với hàm nhân Polynomial : $K(u, v) = (gamma * u^T * v + coef0)^{degree}$
- RBF: tương ứng với hàm nhân Gaussian : $K(u, v) = e^{-gamma*\|u-v\|^2}$
- SIGMOID: tương ứng với hàm nhân Hyperbolic tangent : $K(u, v) = \tanh(gamma * u^T * v + coef0)$

Để đánh giá mức độ hiệu quả của mỗi loại hàm nhân khi xây dựng mô hình, độ chính xác của mô hình được khảo sát khi thay đổi giá trị các tham số của từng loại hàm nhân. Kết thúc giai đoạn này, dựa vào kết quả khảo sát chúng tôi xác định được hàm nhân và bộ tham số "tối ưu".

E. Khảo sát và lựa chọn tham số k

Với mỗi tham số k, mô hình ứng với các bộ dữ liệu của mỗi 6 phương pháp trích đặc trưng được xây dựng dựa vào hàm nhân "tối ưu". Đối với từng phương pháp trích đặc trưng, độ chính xác của mô hình ứng với từng tham số k được so sánh. Từ đó, xác định được sự ảnh hưởng của tham số k đối độ chính xác của các mô hình. Kết thúc giai đoạn này, chúng tôi xác định được tham số k "tốt nhất". Như vậy, có 6 mô hình "tốt nhất" ứng với 6 phương pháp trích đặc trưng.

V. KẾT QUẢ THỰC NGHIỆM

A. Xây dựng kho phim mẫu

B. Danh sách từ lấy mẫu

Các từ được lấy mẫu thuộc 6 chủ đề phổ biến:
(1) Ẩm thực: bánh mì, bánh tét, bún, kem, kẹo, lẩu, xôi.
(2) Đồ vật: cái cân, cây bút, cây kem, cây kéo, cây nhíp, chiếc chiếu, con diều, cái ghế.
(3) Thời gian: hôm nay, hôm qua, ngày, ngày kia.
(4) Thức ăn: trái bắp, trái bầu, trái bí đao, trái cam, trái chanh, trái dừa, cây mía, trái mướp, trái nho, trái sầu riêng.
(5) Tin học: âm thanh, bàn phím, chuột, lập trình, lệnh, màn hình, mật khẩu, virus.
(6) Từ vựng chung: bệnh viện, cây, con nuôi, đồng ý, em bé, bông hoa, học sinh, mang thai, sinh nhật, thầy giáo.

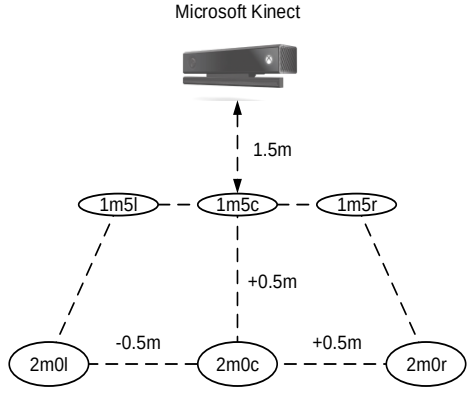
C. Quy ước khi lấy mẫu

Dữ liệu mẫu được thu ở 6 vị trí, mỗi lần lấy mẫu một từ và thực hiện lặp lại 10 lần từ này ở mỗi vị trí. Các mẫu chuyển động của mỗi từ ở mỗi vị trí được lưu lại thành một tập tin KCD (Kinect Custom Data), do đó có 6 phim cho mỗi từ.

D. Xây dựng tập dữ liệu thực nghiệm

Khi thu dữ liệu từ cảm biến Kinect và trích đặc trưng, một số mẫu bị nhiễu nên ảnh bàn tay bị nhòe. Vì vậy, chúng tôi lựa chọn giữ lại các mẫu có ảnh bàn tay rõ nét. Nhằm thuận lợi cho việc xây dựng tập huấn luyện và tập kiểm thử, chúng tôi giữa lại 6 mẫu đối với mỗi phim mẫu. Để đảm bảo sao cho các lớp đều được phân bố vào cả trong tập huấn luyện lẫn tập kiểm thử, ở mỗi phim 4 mẫu ngẫu nhiên được trích ra và đưa vào tập huấn luyện, 2 mẫu còn lại đưa vào tập kiểm thử.

Để độ chính xác của mô hình được đánh giá tốt hơn, với mỗi tập dữ liệu mẫu cần tiến hành sinh 10 bộ dữ liệu huấn luyện và kiểm thử. 10 mô hình phân lớp được huấn luyện dựa vào 10 bộ dữ liệu này, sau đó lấy độ chính xác của mô hình dựa vào giá trị trung bình độ chính xác của 10 mô hình đã huấn luyện.



Hình 4. Các vị trí khi quay phim lấy mẫu

Với mỗi tham số k, có 6 * 10 = 60 bộ dữ liệu. Vì vậy, có 6 * 60 = 360 bộ dữ liệu (6 tham số k bao gồm 5, 10, 15, 20, 25 và 30). Trong nghiên cứu này, dữ liệu mẫu được lấy từ 2 người, đối với mỗi từ mẫu được thu ở 6 vị trí và cần lấy mẫu tổng cộng 50 từ nên:

- Số mẫu của mỗi lớp thuộc tập huấn luyện là 2 * 6 * 4 = 48 mẫu.
- Số mẫu của mỗi lớp thuộc tập kiểm thử là 2 * 6 * 2 = 24 mẫu.
- Số mẫu của mỗi tập huấn luyện là 48 * 50 = 2,400 mẫu.
- Số mẫu của mỗi tập kiểm thử là 24 * 50 = 1,200 mẫu.

Các mẫu được xây dựng theo cấu trúc quy ước của SVM~~,~~. ~~đ~~Do đó, mỗi mẫu của tập huấn luyện và tập kiểm thử bao gồm một nhãn và theo sau là véctơ đặc trưng của một từ trong ngôn ngữ cử chỉ. Số chiều của véctơ đặc trưng tùy thuộc vào từng phương pháp trích đặc trưng. Kết thúc giai đoạn này, chúng tôi có được các bộ dữ liệu huấn luyện và kiểm thử ứng với từng tham số k, và ứng với 6 phương pháp trích đặc trưng.

E. Đánh giá mô hình nhận dạng

F. Khảo sát và xác định hàm nhân SVM "tối ưu"

Để khảo sát độ chính xác của mô hình khi thay đổi giá trị các tham số của từng loại hàm nhân, cần thực hiện 2 bước:

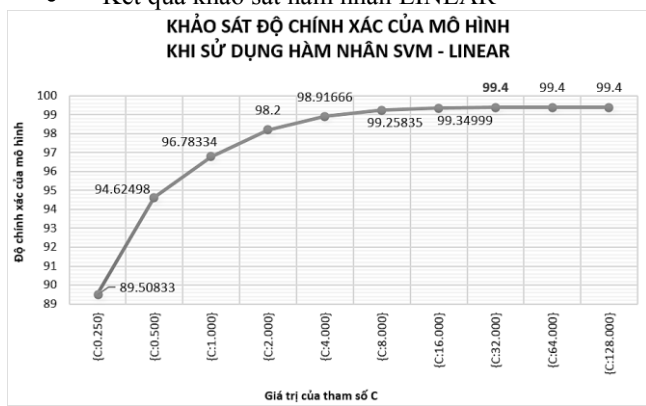
Bảng 2. Các giá trị khảo sát của các tham số của các loại hàm nhân SVM

Hàm nhân	Tham số và các giá trị khảo sát
LINEAR	C = 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128
RBF	C = 1; Gamma = 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128
POLY	C = 1; Degree = 1, 2, 3; Gamma = 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128; Coef0 = 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128
SIGMOID	C = 1; Gamma = 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128; Coef0 = 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128

Bước 1. Khảo sát các hàm nhân trên bộ dữ liệu huấn luyện và kiểm thử ứng với tham số k = 15~~,~~ của phương pháp trích đặc trưng GIST1. Với mỗi hàm nhân thực hiện so sánh và chọn lại một tổ hợp tham số của hàm nhân giúp mô hình có độ chính xác cao nhất. Kết thúc bước này, nhận được có 4 hàm nhân đi kèm với các bộ tham số "tốt nhất".

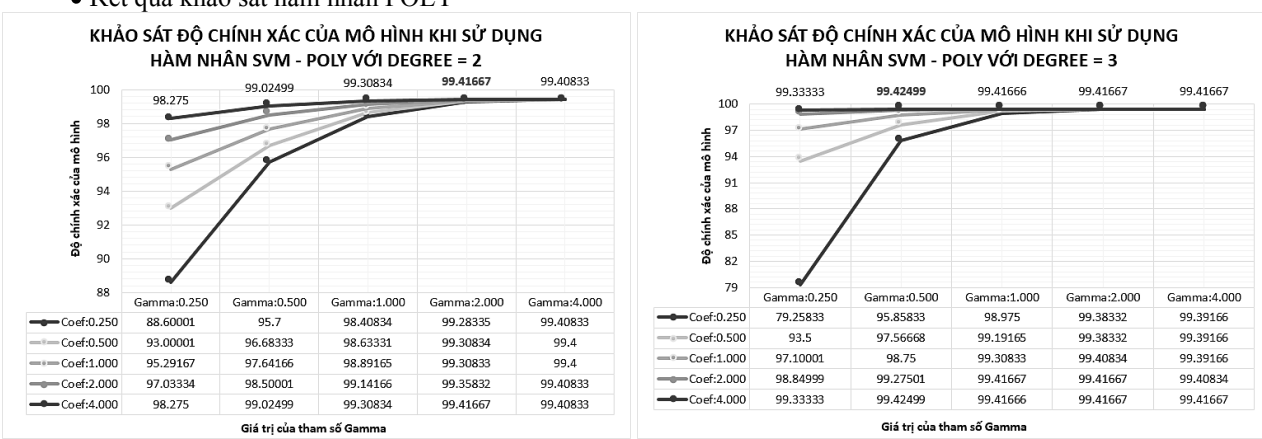
- Tập dữ liệu huấn luyện và kiểm thử: việc huấn luyện mô hình được dựa vào 10 bộ dữ liệu huấn luyện và kiểm thử ứng với tham số k = 15 của phương pháp trích đặc trưng GIST1. Mỗi tập huấn luyện chứa 2,400 mẫu và mỗi tập kiểm thử chứa 1,200 mẫu. Số chiều của véctơ đặc trưng là 2,072 chiều.

- Kết quả khảo sát hàm nhân LINEAR



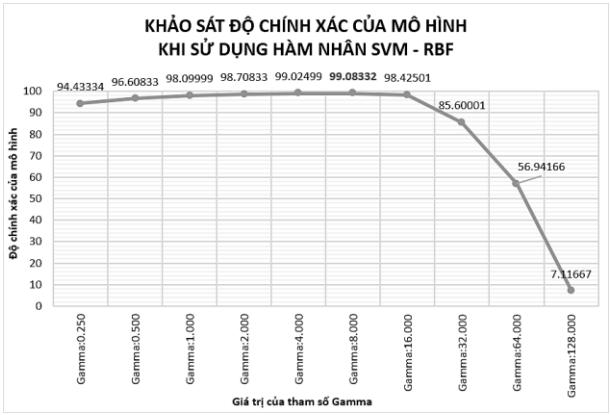
Hình 5. Kết quả khảo sát hàm nhân SVM – LINEAR

- Kết quả khảo sát hàm nhân POLY



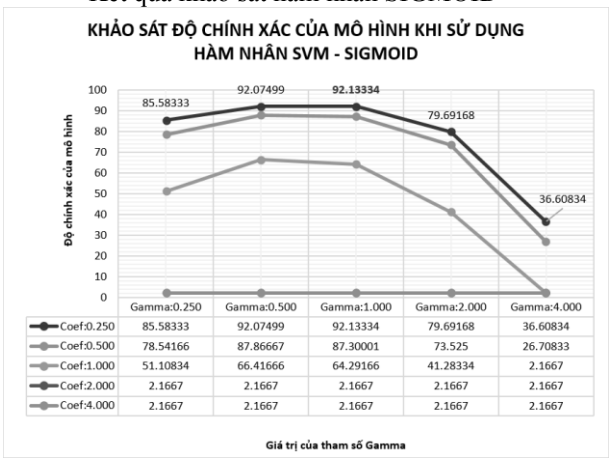
Hình 6. Kết quả khảo sát hàm nhân SVM – POLY với tham số Degree = 2 và Degree = 3

- Kết quả khảo sát hàm nhân RBF



Hình 7. Kết quả khảo sát hàm nhân SVM – RBF

- Kết quả khảo sát hàm nhân SIGMOID

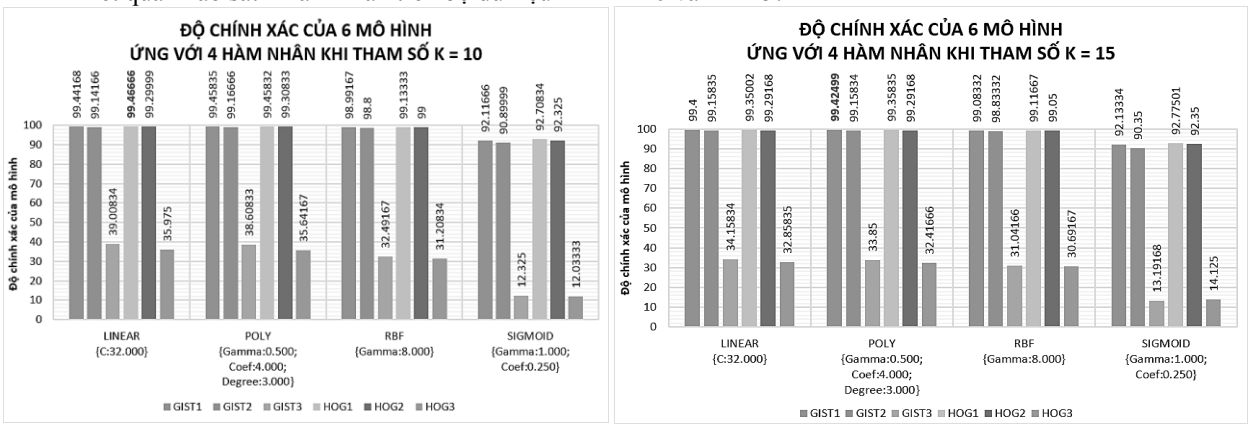


Hình 8. Kết quả khảo sát hàm nhân SVM – SIGMOID

- Tổng hợp: kết thúc quá trình khảo sát, xác định được 4 hàm nhân "tốt nhất" kèm theo bộ tham số "tốt nhất" là:
  - Hàm nhân LINEAR {C: 32}, độ chính xác của mô hình xây dựng dựa vào hàm nhân là 99.4%.
  - Hàm nhân POLY {Gamma:0.5; Coef0:4; Degree:3}, độ chính xác của mô hình xây dựng dựa vào hàm nhân là 99.43%.
  - Hàm nhân RBF {Gamma: 8}, độ chính xác của mô hình xây dựng dựa vào hàm nhân là 99.08%.
  - Hàm nhân SIGMOID {Gamma:1; Coef0:0.25}, độ chính xác của mô hình xây dựng dựa vào hàm nhân là 92.13%.

Bước 2. Khảo sát 4 hàm nhân này lên bộ dữ liệu huấn luyện và kiểm thử ứng với tham số k = 10 và k = 15 của 6 phương pháp trích đặc trưng. Sau đó, so sánh và chọn ra một hàm nhân giúp các mô hình đạt độ chính xác cao nhất.

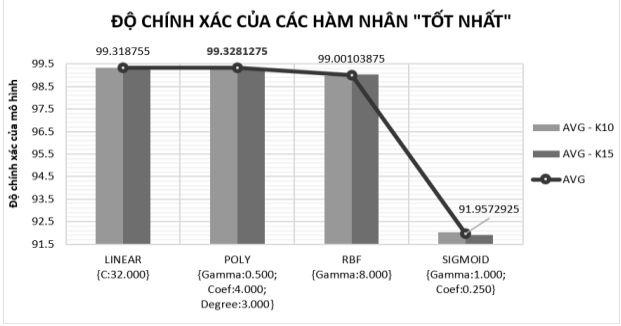
- Kết quả khảo sát 4 hàm nhân trên bộ dữ liệu khi k = 10 và k = 15.



Hình 9. Kết quả khảo sát 4 hàm nhân "tốt nhất"khi k = 10 và k = 15

- Tổng hợp: để tìm ra hàm nhân "tối ưu", cần so sánh độ chính xác trung bình của các mô hình ứng với từng hàm nhân "tốt nhất". Vì độ chính xác của các mô hình ứng với hai phương pháp trích đặc trưng GIST3 và HOG3 quá thấp nên chúng tôi chỉ trình bày kết quả thống kê độ chính xác của các mô hình ứng với các phương pháp trích đặc trưng GIST1, GIST2, HOG1 và HOG2.

Kết thúc hai bước khảo sát, dựa vào dữ liệu thực nghiệm, hàm nhân POLY {Gamma:0.5; Coef0:4; Degree:3} là hàm nhân "tối ưu".

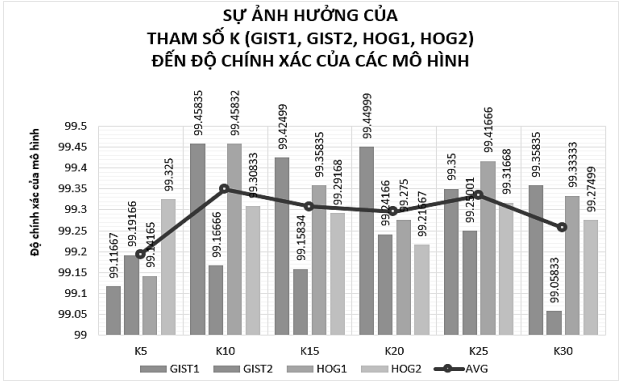


Hình 10. So sánh độ chính xác của 4 hàm nhân "tốt nhất"

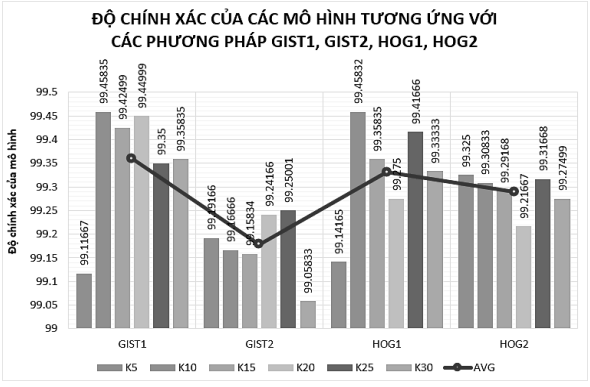
G. Khảo sát và xác định tham số k "tối ưu"

Đối với mỗi tham số k, chúng tôi dựa vào hàm nhân "tối ưu" POLY {Gamma:0.5; Coef0:4; Degree:3} để xây dựng 6 mô hình phân lớp. Sau khi có độ chính xác của các mô hình ứng với các tổ hợp của 6 tham số k và 6 phương pháp trích đặc trưng, khảo sát và xác định tham số k "tối ưu" cũng như phương pháp trích đặc trưng "tốt nhất".

- Kết quả khảo sát: trong 6 phương pháp trích đặc trưng, có 4 phương pháp cho độ chính xác cao bao gồm GIST1, GIST2, HOG1 và HOG2; riêng 2 phương pháp GIST3 và HOG3 có độ chính xác thấp. Do đó, chúng tôi chỉ vẽ biểu đồ cho các phương pháp trích đặc trưng bao gồm GIST1, GIST2, HOG1, HOG2.



Hình 11. Kết quả khảo sát tham số k đối với GIST1, GIST2, HOG1 và HOG2



Hình 12. Kết quả khảo sát độ chính xác của GIST1, GIST2, HOG1 và HOG2

- Tổng hợp: quan sát ba biểu đồ 6.9, biểu đồ 6.10 kết quả cho thấy, với tham số k = 10 thì độ chính xác của các mô hình đạt giá trị cao nhất. Bên cạnh đó, dựa vào biểu đồ 6.11 cho thấy trong 6 phương pháp trích đặc trưng thì GIST1 là phương pháp có độ chính xác cao nhất.

Sau quá trình khảo sát, các mô hình đạt độ chính xác cao nhất khi sử dụng phương pháp trích đặc trưng GIST1 với tham số k = 10. Độ chính xác của các mô hình phân lớp được xây dựng dựa vào phương pháp GIST1 khi tham số k = 10 là 99.46%.

VI. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi đã trình bày 6 phương pháp biểu diễn đặc trưng của các từ trong ngôn ngữ cử chỉ tiếng Việt, từ đó xây dựng các mô hình nhận dạng dựa vào giải thuật SVM. Kết quả thực nghiệm cho thấy phương pháp GIST1 với tham số k bằng 10 kết hợp với hàm nhân SVM LINEAR {C:32} là mô hình phân lớp hiệu quả nhất. Độ chính xác của mô hình nhận dạng dựa vào phương pháp này đạt 99.46% trên bộ dữ liệu kiểm thử với tốc độ nhận dạng mỗi từ là 434ms. Kết quả nhận dạng được trình bày dưới dạng văn bản và đặc biệt, độ chính xác của ứng dụng vẫn được đảm bảo khi nhận dạng trong điều kiện thiếu ánh sáng.

Với kết quả khả quan đạt được, chúng tôi sẽ tiến hành khảo sát với tập từ vựng lớn hơn và bao quát hơn ở các lĩnh vực khác nhau. Để nâng cao độ chính xác khi nhận dạng, chúng tôi sẽ tiến hành khảo sát thêm các tham số của các hàm nhân SVM sau đó lựa chọn ra hàm nhân cũng như mô hình phân lớp có độ chính xác cao hơn. Ngoài ra, chúng tôi sẽ cải tiến hệ thống nhận dạng để phát âm các từ nhận dạng được.

TÀI LIỆU THAM KHẢO

[1] Bộ Khoa học và Công nghệ (2015). Chương trình khoa học công nghệ trọng điểm cấp nhà nước KC07. Bộ Khoa học và Công nghệ, <http://kc01.vpct.gov.vn/News.aspx?ctl=newsdetail&aID=58&p=8>, accessed: 12/08/2016.

[2] Bộ Thông tin và Truyền thông (2009). 28/2009/TT-BTTTT - Thông tư quy định việc áp dụng tiêu chuẩn, công nghệ hỗ trợ người khuyết tật tiếp cận, sử dụng công nghệ thông tin và truyền thông. <http://www.moit.gov.vn/vn/Pages/ChiTietVanBan.aspx?vID=10697>, accessed: 12/08/2016.

[3] Dương Khắc Hưởng, Nguyễn Đăng Bình Trường (2016), *Nghiên cứu nhận dạng ngôn ngữ cử chỉ thông qua quỹ đạo chuyển động liên tục của đối tượng dựa trên mô hình Markov ẩn*, Thesis, ĐH Khoa Học Huế.

[4] Phạm Nguyên Khang, Huỳnh Nhật Minh, Phạm Thế Phi (2015). Nhận dạng ngôn ngữ dấu hiệu với cảm biến Kinect và đặc trưng GIST. *Tạp chí Khoa học Trường Đại học Cần Thơ*, **113**.

[5] UNFPA (United Nations Fund for Population Activities) (2012), *Người khuyết tật ở Việt Nam: Một số kết quả chủ yếu từ Tổng điều tra Dân số và Nhà ở Việt Nam 2009*, Hà Nội.

[6] Cao Dong, Leu M.C., Yin Z. (2015). American Sign Language alphabet recognition using Microsoft Kinect. IEEE, 44–52.

[7] Dalal N., Triggs B. (2005). Histograms of Oriented Gradients for Human Detection. IEEE, 886–893.

[8] Davis J., Shah M. (1994). Recognizing Hand Gestures. *Proceedings of the Third European Conference on Computer Vision (Vol. 1)*, Secaucus, NJ, USA, Springer-Verlag New York, Inc., 331–340.

[9] Gowri.D, Vidhubal.D (2014). Sign language recognition for deaf and dumb people. *Int J Res Eng Technol*, **03**(**19**), 797–799.

[10] Lowe D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *Int J Comput Vis*, **60**(**2**), 91–110.

[11] Murata T., Shin J. (2014). Hand Gesture and Character Recognition Based on Kinect Sensor. *Int J Distrib Sens Netw*, **10**(**7**), 278460.

[12] Steinberg I., London T. M., Castro D. D. (2010). Hand Gesture Recognition in Images and Video. *ResearchGate*.

[13] Can E. LibSVMsharp. GitHub, <https://github.com/ccerhan/LibSVMsharp>, accessed: 13/08/2016.

SIGN LANGUAGE RECOGNITION USING KINECT SENSOR

Vo Hong Khanh, Pham Nguyen Khang

ABSTRACT: *In this paper, we propose a solution for recognizing VSL (Vietnamese Sign Language) based on Microsoft Kinect and SVM (Support Vector Machines). The training data set and test data set will be generated from video that recored by Microsoft Kinect. We present 6 methods that describe sign language features based on 3 ways to describe hands motion path and 2 descriptors for image features GIST descriptor [10] and HOG descriptor [7]. We built data set with 2,400 samples based on 50 words of Vietnamese sign language that divided into 6 topics. 99.46% is the accuracy of the recognition system.*