Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

NHẬN DẠNG NGÔN NGỮ DẤU HIỆU VỚI CAMERA KINECT VÀ ĐẶC TRƯNG GIST

Phạm Nguyên Khang[1], Huỳnh Nhật Minh[1], Võ Trí Thức[1] và Phạm Thế Phi[1]
[1] Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ

Thông tin chung:
Ngày nhận: 19/09/2015
Ngày chấp nhận: 10/10/2015

Title:
Sign language recognition using camera Kinect and Gist feature

Từ khóa:
Ngôn ngữ ký hiệu, camera Kinect, máy học véc-tơ hỗ trợ, nhận dạng cử chỉ

Keywords:
Sign language, Kinect, support vector machines, gestures recognition

ABSTRACT

We present, in this paper, a novel method for sign language recognition. From data acquired with Kinect camera, features of hand movement are extracted. We also propose a new feature to describe hand movement. The feature is computed by dividing the orbit of hand movement into k segments. For each segment, we compute the orientation histogram. The feature is hence independent to length of orbit. Moreover, to improve the discrimant power we also extract the visual information of hand shape with GIST feature. These features are then used to train a recognition model with support vector machines. The experimentations are realized with 280 samples collected from 5 students in Can Tho Disabled Children School. The numerical results show that the proposed method gives an 90% in term of accuracy.

TÓM TẮT

Trong bài báo này, chúng tôi đề xuất một phương pháp mới cho việc nhận dạng ngôn ngữ dấu hiệu. Với dữ liệu được thu nhận từ camera Kinect, chúng tôi trích các đặc trưng chuyển động của bàn tay. Chúng tôi đề xuất một phương pháp biểu diễn quỹ đạo chuyển động của bàn tay bằng cách chia quỹ đạo thành k (e.g. k = 4) đoạn và sau đó tính tổ chức đồ (orientation histogram) của hướng di chuyển cho từng đoạn. Với phương pháp này, đặc trưng chuyển động không phụ thuộc vào độ dài của quỹ đạo. Ngoài ra, để tăng cường khả năng phân biệt, thông tin trực quan (visual) về hình dạng của bàn tay cũng trích xuất với đặt trưng GIST. Tất cả các đặc trưng trên được sử dụng để huấn luyện bộ nhận dạng được huấn luyện bằng mô hình máy học véc-tơ hỗ trợ. Chúng tôi đã thu thập dữ liệu từ 5 bạn học viên trường dạy trẻ khuyết tật thành phố Cần Thơ. Bộ dữ liệu gồm 14 từ, mỗi người thực hiện 4 lần. Tổng cộng là 280 phần tử. Thực nghiệm cho thấy kết quả nhận dạng đạt 90%.

1 GIỚI THIỆU

Theo thống kê, Việt Nam hiện có hơn 2.5 triệu người khiếm thính. Như mọi người bình thường người khiếm thính cũng muốn được đi học, giao tiếp với người những người xung quanh, sử dụng máy tính,… Ngôn ngữ người khiếm thính sử dụng để giao tiếp hiện nay là ngôn ngữ dấu hiệu. Nhằm giúp đỡ các người khiếm thính, nhiều nghiên cứu liên quan đến nhận dạng ngôn ngữ dấu hiệu đã đề xuất. Nhận dạng tự động ngôn ngữ dấu hiệu là một bước không thể thiếu trong các hệ thống tương tác người-máy cho người khiếm thính (hoặc mở rộng hơn: các hệ thống tương tác người máy sử dụng dấu hiệu). Người khiếm thính có thể dùng ngôn ngữ dấu hiệu (ngôn ngữ thông thường của họ) để điều khiển máy tính, nhập văn bản, tìm kiếm thông tin bằng ngôn ngữ dấu hiệu,... Ngoài ra, hệ thống

nhận dạng có thể kết hợp với hệ thống tổng hợp ngôn ngữ dấu hiệu để tạo thành một hệ thống tương tác người-máy hoàn chỉnh giúp người khiếm thính có thể "nói chuyện" được với máy tính, giúp họ hoà nhập cộng đồng tốt hơn trong kỷ nguyên công nghệ thông tin.

Hệ thống nhận dạng ngôn ngữ dấu hiệu (sign language recognition system) dựa trên chuỗi hình ảnh hướng đến nhận dạng các từ trong ngôn ngữ dấu hiệu từ hình ảnh thu từ camera hoặc từ các đoạn video đã thu được từ trước. Trong vài thập kỷ qua, nhiều công trình nghiên cứu về lĩnh vực này đã được đề xuất và thu được một số kết quả khả quan. Đầu tiên, các nghiên cứu tập trung vào bài toán học có giám sát hoàn toàn với tập học đã được gán nhãn trước. Huỳnh Hữu Hưng *và ctv*. (2012) nhận dạng ngôn ngữ ký hiệu từ ảnh tĩnh bằng mạng nơ-ron. Theo các tác giả, kết quả là khả quan (98% độ chính xác). Tuy nhiên, các tác giả chỉ mới đề cập đến nhận dạng ảnh tĩnh chứ không phải đoạn video. Dương Văn Hiếu (2009) đề xuất một mô hình nhận dạng ngôn ngữ dấu hiệu tiếng việt với mô hình markov ẩn mờ (Fuzzy Hidden Markov Model). Tuy nhiên, kết quả đạt được còn hạn chế. Tomas Pfister *et al*. (2012) đã đề xuất một phương pháp tách người ra dấu và xác định các vị trí quan trong như: đầu, vai, bàn tay, cùi chỏ dựa trên màu sắc và mô hình học máy rừng ngẫu nhiên. Một phát triển của phương pháp này được công bố trong (Charles *et al*., 2013).

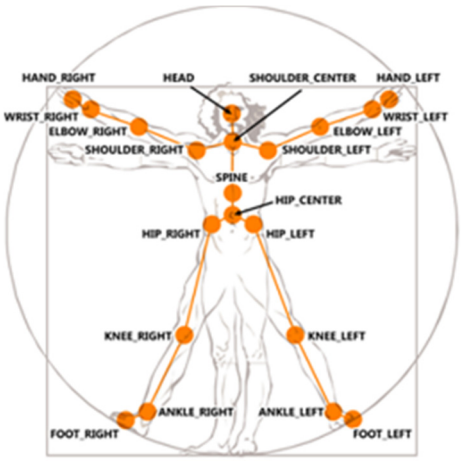
Gần đây, với sự ra đời của camera Kinect, việc trích vị trí các khớp xương trên cơ thể người có thể được thực hiện dễ dàng. Hàng loạt công trình liên quan đến việc ứng dụng dữ liệu thu được từ Kinect đã được công bố như: Nhận dạng tư thế người (Lan et al. 2013), Nhận dạng cử chỉ (Wang *et al*., 2012; Hussein *et al*., 2013) và cả nhận dạng ngôn ngữ dấu hiệu. Trong (Agarwal và Thakur, 2013), các tác giả trình bày một phương pháp để nhận dạng các số (từ 0 đến 9) trong ngôn ngữ dấu hiệu.

Trong bài báo này, chúng tôi sử dụng dữ liệu thu được từ camera Kinect (Shotton *et al*., 2011) để phục vụ cho việc nhận dạng ngôn ngữ dấu hiệu. Hai đóng góp chính của bài báo là: (i) đề xuất 4 phương pháp trích đặc trưng quỹ đạo của bàn tay và (ii) kết hợp các đặc trưng quỹ đạo chuyển động của bàn tay và đặc trưng hình dáng của bàn tay nhằm làm tăng khả năng phân biệt của các đặc trưng. Phần tiếp theo của bài báo được trình bày như sau: Thu nhận dữ liệu bằng camera Kinect được trình bày trong phần 2; tiếp theo đó, chúng tôi trình bày 4 phương pháp trích đặc trưng quỹ đạo chuyển động và trích đặc trưng hình dáng bàn tay với đặc trưng GIST; mô hình máy học véc-tơ hỗ trợ được trình bày trong phần 4; phần 5 dành cho kết quả thực nghiệm và sau cùng là kết luận và hướng phát triển.

2 THU NHẬN DỮ LIỆU VỚI CAMERA KINECT

2.1 Camera Kinect

Thiết bị Kinect cho phép chụp ảnh màu và ảnh độ sâu cùng một lúc. Ngoài ra, với phiên bản hiện tại Kinect còn cho phép thu được vị trí của 20 khớp xương trên cơ thể các khớp xương được thể hiện ở Hình 1. Với mỗi khớp, thông tin chính sẽ là vị trí của nó trong tọa độ Oxyz.



Hình 1: Mô hình 20 khớp xương camera Kinect có thể thu nhận

2.2 Thu nhận ngôn ngữ dấu hiệu với camera Kinect

Khi người ra dấu (signer) đứng đối diện với camera, dữ liệu thu thập được chính xác hơn mặc dù Kinect cho phép người ra dấu quay một góc 30^o so với chính diện. Dữ liệu thu thập được từ Kinect là một chuỗi các khung, mỗi khung bao gồm 3 kênh: (i) dữ liệu khung xương, mỗi khung xương gồm toạ độ của 20 khớp xương, (ii) ảnh màu (tương đương với hình ảnh thu được với máy ảnh thông thường) và ảnh độ sâu tính từ camera. Từ dữ liệu khung xương, ta có thể trích xuất dễ dàng góc quay giữa các khớp xương để phục vụ các tác vụ khác như: nhận dạng tư thế.

Đối với ngôn ngữ ký hiệu, vị trí của hai tay và đầu là đáng quan tâm nhất. Vì thế trong nghiên cứu này, chúng chỉ quan tâm đến vị trí của hai tay và đầu. Ngoài ra, đa phần các từ trong ngôn ngữ dấu hiệu chỉ cần dùng một tay là đủ để biểu diễn.

3 TRÍCH ĐẶC TRƯNG

Trích đặc trưng là một bước không thể thiếu trong bất kỳ bài toán nhận dạng nào. Đối với ngôn ngữ ký hiệu, dữ liệu chúng ta thu nhận được là một đoạn video ngắn khoảng 30 – 35 khung (frame). Ta cần phải trích thông tin quan trọng từ dữ liệu này. Thông tin này phải phản ánh được bản chất của từ tương ứng phải có khả năng phân biệt cao giữa từ này với từ khác. Dựa vào trực quan, chúng ta có thể dễ dàng thấy rằng: mỗi từ trong ngôn ngữ dấu hiệu là một cử chỉ (gesture) được thực hiện bằng tay và đôi khi kết hợp với đầu. Vì thế, trong nghiên cứu này chúng tôi đề xuất một phương pháp mới để trích đặc trưng và biểu diễn các từ dưới dạng véc-tơ đặc trưng có cùng số chiều. Vì mỗi từ tương ứng với một cử chỉ, nên quỹ đạo chuyển động của tay là một trong các thông tin quan trọng cần trích xuất. Bên cạnh đó, hình dáng của bàn tay lúc bằng đầu và kết thúc một từ cũng là một thông tin có tính phân biệt cao.

3.1 Đặc trưng chuyển động của bàn tay

Kênh khung xương của dữ liệu thu được từ camera Kinect cho phép ta trích được vị trí của bàn tay theo thời gian. Như thế ta có được quỹ đạo chuyển động của bàn tay như một danh sách các điểm trong không gian 3 chiều. Vì số lượng khung của mỗi từ không giống nhau nên ta không thể sử dụng trực tiếp danh sách điểm này như đặc trưng của quỹ đạo. Ta cần phải trích các đặc trưng sao cho nó độc lập với số lượng khung ảnh của một từ. Chúng tôi nghiên cứu các đặc trưng có tính chất này theo ba hướng: (i) canh lề và nội suy các quỹ đạo và (ii) tổng hợp thông tin theo thời gian và (iii) kết hợp cả hai hướng trên. Với tiếp cận, đó chúng tôi đề xuất 4 phương pháp mới cho việc trích xuất đặc trưng quỹ đạo chuyển động.

3.1.1 Phương pháp 1

Phương pháp này thuộc họ canh lề và nội suy. Giả sử quỹ đạo chuyển động của bàn tay được một tả bằng n điểm P = (p_1, p_2, ..., p_n). Để trích đặc trưng, chúng tôi chia quỹ đạo thành k (ví dụ k = 15) đoạn và mỗi đoạn lấy 1 điểm đại diện sau đó tính các đặc trưng sau:

Tâm của quỹ đạo:

$$x_c = \frac{x_1 + x_2 + \cdots + x_k}{k} \quad (1)$$

$$y_c = \frac{y_1 + y_2 + \cdots + y_k}{k} \quad (2)$$

$$z_c = \frac{z_1 + z_2 + \cdots + z_k}{k} \quad (3)$$

Khoảng cách trung bình đến tâm:

$$\bar{d} = \frac{\sum_{i=1}^{k} d(p_i, c)}{k} \quad (4)$$

với $d(p_i, c)$ là khoảng cách từ điểm p_i đến tâm c.

Vận tốc chuyển động tại từng điểm:

$$v_i = p_i - p_{i-1} \quad (5)$$

Góc giữa các điểm:

Tích vô hướng:

$$v_1 \cdot v_2 \quad (6)$$

Tích hữu hướng của v_1 và v_2:

$$[v_1, v_2] \quad (7)$$

với v_1 là vectơ tạo từ 2 điểm p_i và p_{i-1} và v_2 là vectơ tạo từ 2 điểm p_i và p_{i+1}.

Như thế, với phương pháp này véc-tơ đặc trưng thu được có 3 + 1 + 14*3 + 13 + 13*3 = 98 chiều. Để đặc trưng tâm của quỹ đạo bất biến với phép tịnh tiến chúng tôi tính vị trí tương đối (hiệu) của tâm quỹ đạo so với vị trí của đầu.

3.1.2 Phương pháp 2

Phương pháp này tổng hợp các thông tin theo thời gian. Với mỗi điểm trong danh sách điểm của quỹ đạo (ngoại trừ điểm đầu tiên), ta tính hướng chuyển động của quỹ đạo tại điểm này và phân bố nó vào một trong 8 hướng ứng với 8 phần trong không gian Oxyz. Đếm số lượng điểm rơi vào từng phần ta có được một véc-tơ 8 phần tử mô tả quỹ đạo của chuyển động.

3.1.3 Phương pháp 3

Phương pháp này kết hợp cả phương pháp 1 và 2. Ý tưởng chính là chia quỹ đạo chuyển động của bàn tay thành k (ví dụ k = 4) phần, với mỗi phần ta tính 8 đặc trưng theo phương pháp 2. Như vậy, với mỗi một mẫu sẽ có tổng cộng 8*k đặc trưng. Phương pháp này chính là tổng quát hoá của phương pháp 2. Nếu chọn k = 1, ta có kết quả như phương pháp 2.

3.1.4 Phương pháp 4

Tương tự phương pháp 3, nhưng thay vì đếm số lượng điểm rơi vào từng phần, ta sẽ cộng dồn độ lớn của gradient tại điểm đang xét. Ý tưởng của phương pháp này dựa trên tinh thần của đặc trưng cục bộ SIFT (Lowe, 2004).

3.2 Đặc trưng hình dáng bàn tay

Ngoài thông tin về quỹ đạo chuyển động của tay, hình dáng của bàn tay cũng là một thông tin quan trọng để phân biệt từ này với từ khác. Chúng tôi đề xuất sử dụng đặc trưng GIST (Oliva và

Torralba, 2001) để trích đặc trưng về hình dáng của bàn tay tại 3 thời điểm: bắt đầu, ở giữa và kết thúc một từ. Trong ngôn ngữ ký hiệu hình dáng bàn tay lúc bắt đầu và kết thúc là quan trọng nhất có khả năng phân biệt cao.

Đặc trưng GIST thuộc nhóm đặc trưng biến đổi toàn cục và khai triển chuỗi. Khác với đặc trưng SIFT (Lowe, 2004), GIST là một đặc trưng toàn cục biểu diễn nội dung ảnh được Oliva & Torralba đề xuất năm 2001. Đặc trưng GIST thể hiện dưới dạng một véc-tơ và mỗi véc-tơ này được tính toán từ kết quả của việc áp dụng các bộ lọc Gabor lên ảnh. Từ dữ liệu ảnh đầu vào, sau khi trích đặc trưng sẽ cho ra một véctơ có 960 chiều. Các bước tiến hành như sau:

– Ảnh đầu vào sau khi được tiền xử lý sẽ được tách ra thành 3 kênh màu Red-Green-Blue riêng biệt.

– Áp dụng phép biến đổi Fourier trên mỗi kênh màu.

– Ứng với mỗi ảnh Fourier áp dụng lần lượt 20 bộ lọc Gabor lên ảnh. Bộ lọc Gabor được tạo ra ở 3 scales và 8 hướng. Trong đó, scale 1 và scale 2 sử dụng 8 bộ lọc, scale 3 sử dụng 4 bộ lọc.

– Cuối cùng, kết quả của mỗi bộ lọc được đưa qua phép biến đổi Fourier ngược, sau đó chia thành 16 vùng bằng nhau và trích đặc trưng. Kết quả của mỗi vùng là một đặc trưng.

Như vậy, số chiều của đặc trưng GIST là: 3*(8+8+4)*16 = 960 chiều.

4 MÁY HỌC VÉC-TƠ HỖ TRỢ

4.1 Mô hình máy học véc-tơ hỗ trợ

Mô hình máy học véc-tơ hỗ trợ (Support vector machines hay viết tắt là SVM) là một mô hình học tự động do (Vapnik *et al*., 1995) đề xuất và phát triển. Xét bài toán phân lớp tuyến tính nhị phân. Cho tập huấn luyện gồm m phần tử: $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$, mỗi phần tử là một véc-tơ trong không gian n chiều. Mỗi phần tử thuộc về một trong hai lớp (+1: lớp dương, -1: lớp âm) như Hình 2. Nhiệm vụ của bài toán phân lớp tuyến tính là tìm một siêu phẳng tách rời hai lớp sao cho các phần thuộc cùng lớp nằm về một phía của siêu phẳng. Cũng cùng một mục tiêu đó, mô hình SVM cho bài toán phân lớp tuyến tính nhị phân cũng tìm một siêu phẳng tách rời hai lớp dữ liệu. Tuy nhiên, để tăng cường khả năng tổng quát hoá, mô hình SVM cố gắng tìm một siêu phẳng tối ưu trong tất cả các siêu phẳng có khả năng tách rời tập dữ liệu. Siêu phẳng tối ưu theo mô hình SVM là siêu phẳng mà khoảng cách từ nó đến phần tử gần nó nhất là lớn nhất. Để tìm được siêu phẳng tối ưu, ta định nghĩa hai siêu phẳng hỗ trợ song song nhau: một cho lớp dương (d^+) và một cho lớp âm(d^-):

$$(d^+): w^T x + b = +1 \quad (1)$$

$$(d^-): w^T x + b = -1 \quad (2)$$

trong đó: w là véc-tơ pháp tuyến của 2 siêu phẳng và b là hệ số tự do (còn được gọi là độ lệch – bias).

Không giảm tính tổng quát, ta luôn có thể giả sử tất cả các phần tử thuộc lớp âm nằm về bên trái của siêu phẳng d^- và tất cả các phần tử thuộc lớp dương nằm về phía bên phải của siêu phẳng d^+. Ta sẽ tìm đồng thời d^+ và d^- sao cho khoảng cách giữa chúng là lớn nhất. Khoảng cách giữa hai siêu phẳng được gọi là *lề* (margin):

$$margin = \frac{1}{\|w\|} \quad (8)$$

Bài toán tối ưu của SVM chính là bài toán quy hoạch toàn phương:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad (9)$$

với ràng buộc:

$$y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1$$

trong đó $y^{(i)}$ là nhãn hay lớp của phần tử i.

Giải bài toán tối ưu này ta thu được w và. Để dự báo nhãn của một phần tử mới x, ta xét dấu của $w^T x^{(i)} + b$ hay:

$$predict(x) = sign\left(w^T x^{(i)} + b\right) \quad (10)$$

Trường hợp, dữ liệu không khả tách tuyến tính (ta không thể nào tách rời dữ liệu bằng một siêu phẳng mà không có phần tử nằm sai phía), mô hình SVM có thể mở rộng bằng cách thêm vào mô hình các biến lỗi $z^{(i)}$ (ta xem khoảng cách từ các phần tử nằm sai phía so với siêu phẳng hỗ trợ của chúng như là lỗi). Bài toán tối ưu đối với SVM bây giờ trở thành bài toán tối ưu 2 mục tiêu: lề lớn nhất và lỗi nhỏ nhất. Ta có thể kết hợp 2 mục tiêu lại thành một tiêu duy nhất nhờ vào tham số điều chỉnh sự kết hợp này. Bài toán tối ưu của SVM trong trường hợp này sẽ là:

$$\min_{w} \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{m} z^{(i)} \quad (11)$$

với ràng buộc:

$$y^{(i)}\left(w^T x^{(i)} + b\right) + z^{(i)} \geq 1$$

$$z^{(i)} \geq 0$$

trong đó $z^{(i)}$ là biến lỗi được định nghĩa như là khoảng cách từ phần tử nằm sai phía đến siêu phẳng hỗ trợ của nó và c là hằng số điều chỉnh độ rộng của lề và lỗi. Bài toán đối ngẫu của nó:

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)} x^{(j)} - \sum_{i=1}^{m} \alpha_i$$

với ràng buộc:

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

$$0 \leq \alpha_i \leq c$$

Mô hình SVM cũng có thể được mở rộng để xử lý bài toán phân lớp phi tuyến sử dụng kỹ thuật hàm nhân (kernel function) trên bài toán đối ngẫu. Ta thay tích vô hướng của hai phần tử $x^{(i)}x^{(j)}$ bằng hàm nhân $K\left(x^{(i)}, x^{(j)}\right)$.

Mô hình SVM được cho là một phương pháp tổng quát cho các bài toán của học máy bao gồm: phân lớp, hồi quy và ước lượng mật độ xác suất. Nếu xét riêng về khả năng giải bài toán phân lớp, SVM có tính tổng quát hoá cao (nhờ vào lề lớn) vì thế hiệu quả phân lớp luôn bằng hoặc cao hơn các phương pháp phân lớp khác.

Cần phải nhắc lại rằng bài toán tối ưu của SVM là bài toán quy hoạch toàn phương. Để giải bài toán này, nhiều phương pháp đã được đề xuất và công bố trong đó có thể kể đến mô hình SMO (Platt, 1998). Một số công trình khác biến đổi một ít mô hình SVM để chuyển từ bài toán quy hoạch toàn phương sang bài toán hệ phương trình tuyến tính (Fung và Mangasarian, 2001) hay cải biên bài toán SVM gốc để giải bằng phương pháp lặp Newton (Fung và Mangasarian, 2001). Trong nghiên cứu này chúng tôi sử dụng bản cài đặt libSVM của (Chang và Lin, 2001). Bản cài đặt này được cộng đồng học máy xem như là chuẩn cài đặt của SVM.

4.2 Nhận dạng ngôn ngữ ký hiệu với máy học véc-tơ hỗ trợ

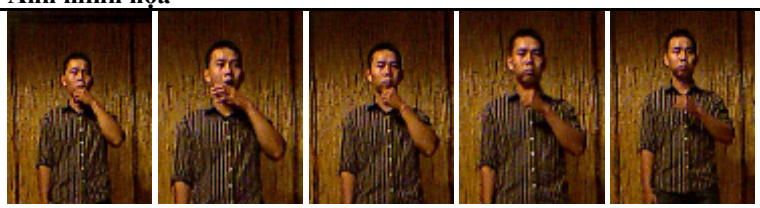
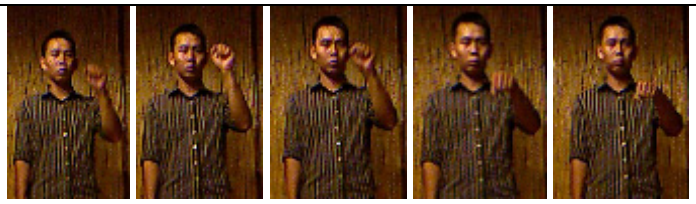
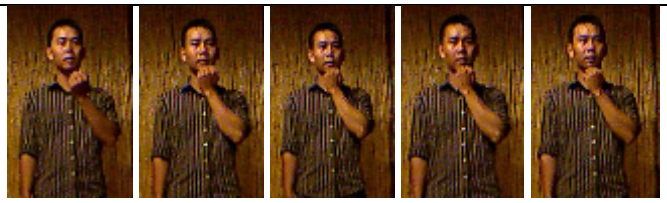
Các đặc trưng được trích ra trong bước trích đặc trưng được dùng để biểu diễn các từ ngôn ngữ dấu hiệu. Như thế, mỗi từ được biểu diễn bằng một véc-tơ đặc trưng có n phần tử. Trong cả ba phương pháp trích đặc trưng chúng tôi đề xuất, n có giá trị khá lớn nên chỉ cần mô hình SVM tuyến tính là có thể phân lớp được dữ liệu.

Mô hình SVM cơ bản chỉ có thể áp dụng để giải quyết bài toán phân lớp nhị phân. Bài toán nhận dạng ngôn ngữ dấu hiệu là bài toán đa lớp: mỗi từ trong ngôn ngữ dấu hiệu tương ứng với một lớp. Trong trường hợp này ta có thể sử dụng một trong hai chiến lược: 1 – tất cả hay 1 – 1 để xử lý dữ liệu đa lớp. Với chiến lược 1 – tất cả, ta cần xây dựng k mô hình SVM ứng với k lớp. Với mô hình thứ i, ta xem một lớp i như là lớp dương và các lớp khác xem như là lớp âm. Để nhận dạng một phần tử dữ liệu mới thuộc lớp nào, ta cho cả k mô hình SVM cùng phân lớp phần tử này, sau đó quyết định lớp của phần tử mới bằng phương pháp bình chọn số đông. Chiến lược 1 – 1 cũng xử lý tương tự. Tuy nhiên, ta cần phải xây dựng $C_2^k = \frac{k(k-1)}{2}$ mô hình tất cả, mỗi mô hình được xây dựng dựa trên dữ liệu của hai lớp.

5 KẾT QUẢ THỰC NGHIỆM

5.1 Dữ liệu

Dữ liệu được thu thập từ 5 học viên, mỗi người đứng cách camera Kinect 2.5m, đứng trực diện với camera Kinect. Mỗi người thực hiện 14 ký hiệu đã được định nghĩa trước và thực hiện 4 lần với mỗi ký hiệu. Camera Kinect sẽ tiến hành thu lại tất cả dữ liệu bao gồm ảnh màu, ảnh độ sâu, tọa độ 20 khớp xương và lưu lại vào tập tin có định dạng .xed. Hình 2 minh hoạ một số ngôn ngữ ký hiệu do chúng tôi thu thập từ các học viên của trường dạy trẻ khuyết tật thành phố Cần Thơ.

STT	Ký hiệu	Ảnh minh họa
1	Ông	
2	Bà	
3	Đúng	
4	Sai	
5	Hiểu	
6	Nghĩ	

Hình 2: Một số từ trong ngôn ngữ dấu hiệu

5.2 Kết quả nhận dạng với các phương pháp trích đặc trưng khác nhau

Bảng 1 trình bày kết quả phân lớp (độ chính xác phân lớp tổng thể) đối với các phương pháp trích đặc trưng khác nhau. Phương pháp 4 cho kết quả cao nhất so với các phương pháp 1, 2 và 3. Ngoài ra, cũng cần phải chú ý rằng thông tin về hình dáng bàn đóng một vai trò khá quan trọng trong việc phân biệt từ này và từ khác. Chỉ riêng thông tin về hình dáng của bàn tay đã cho kết quả 80%. Việc kết hợp đặc trưng quỹ đạo chuyển động và đặc trưng về hình dáng bàn tay cho kết quả cao nhất, đạt 90%. Điều này cho thấy rằng thông tin về hình dáng bàn tay đóng vai trò rất quan trọng trong việc nhận dạng ngôn ngữ ký hiệu.

Bảng 1: so sánh kết quả phân lớp đối với các phương pháp trích đặc trưng

TT	Đặc trưng	Phương pháp 1	Phương pháp 2	Phương pháp 3 (k=4)	Phương pháp 4 (k=4)
1	Quỹ đạo	19.64%	30.45%	37.14%	38.93%
2	Hình dáng bàn tay			80.02%	
3	Quỹ đạo + hình dáng bàn tay	86.93%	87.93%	88.93%	**90%**

6 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi đã trình bày một phương pháp mới trong nhận dạng ngôn ngữ dấu hiệu với dữ liệu thu thập từ camera Kinect và đặc trưng GIST. Các thực nghiệm được thực nghiệm trên tập dữ liệu thật thu từ các em học viên trường khuyết tật thành phố Cần Thơ. Kết quả thực nghiệm cho thấy rằng phương pháp trích đặc trưng quỹ đạo bằng cách kết hợp chia đoạn và tổng hợp thông tin theo thời gian cho kết quả cao nhất. Việc kết hợp đặc trưng quỹ đạo kết hợp với đặc trưng hình dáng bàn tay đã cải thiện đáng kể hiệu quả nhận dạng. Với mô hình máy học véc-tơ hỗ trợ, độ chính xác phân lớp đạt 90%. Kết quả này có thể so sánh được với các phương hiện nay trong lĩnh vực nhận dạng ngôn ngữ dấu hiệu như mô hình Markov ẩn.

Với kết quả khả quan như thế, chúng tôi sẽ tiếp tục nghiên cứu, thực nghiệm với số lượng từ nhiều hơn, phức tạp hơn (được thực hiện bằng 2 tay và có thể kết hợp với các bộ phận khác của cơ thể). Một hướng phát triển khác là nghiên cứu phương pháp nhận dạng các từ liên tục nhằm xây dựng thành một hệ thống có khả năng giao tiếp với người khiếm thính. Chúng dự định thực hiện điều này trong các nghiên cứu tiếp theo.

Cũng cần phải chú ý rằng, các thực nghiệm trong bài báo này được thực hiện với camera Kinect của Microsoft phiên bản v1. Phiên bản v2 của Kinect có khả năng định vị các khớp tốt hơn và cho phép chụp ảnh màu, ảnh độ sâu rõ hơn. Chúng tôi hi vọng rằng kết quả nhận dạng sẽ tốt hơn nếu thực nghiệm trên dữ liệu thu nhận với Kinect v2.

LỜI CẢM TẠ

Nhóm tác giả xin chân thành cảm ơn sự hỗ trợ kinh phí của Trường Đại học Cần Thơ thông qua đề tài cấp cơ sở T2015-29. Chân thành cảm ơn sự hỗ trợ của các em học viên trường khuyết tật thành phố Cần Thơ trong việc thu thập dữ liệu.

TÀI LIỆU THAM KHẢO

1. Agarwal, A. and M.K., Thakur, 2013. In proceedings of the 6th International Conference on Contemporary Computing (IC3), 181 – 185.
2. Chang, C. C. and C. J. Lin, 2001, Libsvm – a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.
3. Duong Van Hieu, Supot Nitsuwat, Sign Language recognition for hearing-impaired people using trajectory feature based on the Fuzzy Hidden Markov Models, Hội thảo quốc gia lần thứ 12, một số vấn đề chọn lọc của công nghệ thông tin và truyền thông: chủ đề phát hiện tri thức từ dữ liệu, Biên Hòa, 2009.
4. Fung, G. and O. L. 2001. Mangasarian, Proximal Support Vector Machine Classifiers, in Proceedings of Conference on Knowledge Discovery and Data Mining, August 26-29, 2001, San Francisco, CA. 77 – 86.
5. Fung, G. and O. L. Mangasarian. 2002. Finite Newton Method for Lagrangian Support Vector Machine Classification. Technial report, Data Mining Institute, Computer Sciences Department, University of Wisconsin.
6. Hussein, Mohamed E., Marwan Torki, Mohammad A. Gowayyed, Motaz El-Saban, 2013, Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations, in Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, 2466 – 2472.
7. Huỳnh Hữu Hưng, Nguyễn Trọng Nguyên, Võ Đức Hoàng, Hồ Viết Hà, Nhận dạng ngôn ngữ ký hiệu tiếng Việt sử dụng mạng Neuron nhân tạo, Tạp chí Khoa học và Công nghệ, Đại học Đà Nẵng, Số: 12 (61); pp: 75-80, 2012.
8. Le Thi-Lan, Minh-Quoc Nguyen, Thi-Thanh-Mai Nguyen, 2013, Human posture recognition using human skeleton provided by Kinect, in Proceedinds of the International Conference on Computing, Management and Telecommunications, 340 – 345.
9. Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2):91–110.

10. Platt, J. 1998. Sequential minimal optimization: a fast algorithm for training support vector machines. Microsoft research technical report MSR-TR-98-14.
11. Shotton, J., A. Fitzgibbon and M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. 2011. Real-time human pose recognition in parts from single depth images," In Proceedings of IEEE Conference on CVPR, 1297-1304.
12. Vapnik, V. 1995. The nature of statistical learning theory, Springer-Verlag, New York.
13. Wang Jiagn, Zicheng Liu, Ying Wu, Junsong Yuan, 2012. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1290 – 1297.