TẠP CHÍ KHOA HỌC ĐẠI HỌC HẠ LONG
Scientific Journal of Ha Long University
http://uhl.edu.vn/

PHƯƠNG PHÁP NHẬN DIỆN MẪU SỬ DỤNG MÔ HÌNH TÚI TỪ VÀ MẠNG NƠRON

Nguyễn Toàn Thắng[1*], Đinh Xuân Lâm[1]

[1]Trường Đại học Công nghệ thông tin và Truyền thông, Đại học Thái Nguyên

** Email: thangnt@ictu.edu.vn*

Ngày nhận bài: 20/10/2022 *Ngày nhận bài sửa sau phản biện: 10/11/2022* *Ngày chấp nhận đăng: 14/11/2022*

TÓM TẮT

Mục đích của bài báo là xây dựng một thuật toán nhận diện cử chỉ tay trong các khung hình thu trực tiếp từ camera theo thời gian thực. Thuật toán đề xuất sử dụng mô hình túi từ (bag-of-features, bag-of-words), bộ mô tả đối tượng SURF, phương pháp phân cụm k-means, kết hợp với phương pháp phân lớp bằng mạng nơron. Trong đó, mô hình túi từ kết hợp với SURF và k-means được sử dụng để tạo ra các vectơ đặc trưng làm dữ liệu đầu vào cho mạng nơron. Thuật toán được huấn luyện và thử nghiệm với các bộ dữ liệu ảnh tự tạo. Các thí nghiệm cho thấy, thuật toán đề xuất đảm bảo được tốc độ xử lý cao (dưới 40 ms cho mỗi khung hình) để có thể thực hiện trong thời gian thực với dữ liệu thu trực tiếp từ một camera, có tính bền vững với một số dạng biến đổi của đối tượng (xoay hình, thay đổi kích thước và vị trí trong khung hình), đồng thời đảm bảo độ chính xác nhận diện cao (~ 90%).

Từ khóa: *bộ mô tả đối tượng, mạng nơron, mô hình túi từ, nhận diện cử chỉ, nhận diện mẫu.*

A METHOD FOR PATTERN RECOGNITION USING BAG-OF-WORDS MODEL AND NEURAL NETWORK

ABSTRACT

The purpose of the project is to create an algorithm for real-time hand gesture recognition in video frames captured directly from the camera. The proposed algorithm is based on the bag-of-features (or bag-of-words) model, SURF-descriptor, k-means clustering, and neural network classification method. The bag-of-words model combined with SURF and k-means is used to create feature vectors, which then are fed as input data for the neural network. The algorithm is trained and tested with a self-made image data set. Experiments with various testing data sets demonstrate that the proposed algorithm ensures a high processing speed (less than 40 ms for each frame) to be able to perform in real time with data captured directly from a camera, keeps being invariant to transformations of the object in the video frame (including rotation, scaling and affine transition), and provides high recognition accuracy (~ 90%).

Keywords: *bag-of-words model, gesture recognition, neural network, object descriptor, pattern recognition.*

1. ĐẶT VẤN ĐỀ

Ngày nay, với sự phát triển rộng rãi của các ứng dụng công nghệ thông tin trong cuộc sống, việc tương tác giữa con người và thiết bị ngày càng trở nên quan trọng Trong các lĩnh vực khác cần tới thông tin 3D (như trò chơi máy tính, robot, lĩnh vực thiết kế, v.v.), người ta sử dụng các thiết bị cơ khí như bóng

lăn, cần điều khiển, hay găng tay dữ liệu (Argyros & Lourakis, 2006). Tuy nhiên, con người giao tiếp chủ yếu bằng "nghe" và "nhìn", do đó giao diện người – máy sẽ trực quan hơn nếu con người có thể điều khiển máy tính bằng giọng nói hay cử chỉ giống như khi tương tác giữa người với người trong thế giới thực mà không cần thông qua các thiết bị điều khiển khác như chuột hay bàn phím (Barczak & Dadgostar, 2005). Một ưu điểm khác là người dùng có thể giao tiếp từ xa mà không cần phải có tiếp xúc vật lý với máy tính. So với các hệ thống điều khiển bằng lệnh âm thanh, một hệ thống thị giác sẽ thích hợp hơn trong môi trường ồn ào hoặc trong trường hợp âm thanh bị nhiễu (Bretzner và cs., 2002).

Tương tác người – máy (human – computer interaction, HCI) là một lĩnh vực thu hút nhiều nghiên cứu và đã đạt được nhiều kết quả ấn tượng trong thời gian gần đây. Một trong những bài toán quan trọng của lĩnh vực này là cung cấp khả năng điều khiển máy tính (hoặc thiết bị) từ xa thông qua camera kết nối với máy (Chen và cs., 2007). Bài toán này thường bao gồm các bước: phát hiện đối tượng trong thị trường của camera (ví dụ, tay, mặt, cơ thể người điều khiển hoặc một thiết bị đặc biệt nào đó dùng để điều khiển); theo dõi chuyển động của đối tượng; nhận diện hình dạng và cách thức chuyển động của đối tượng (El-Sawah và cs., 2008). Kết quả nhận diện được sử dụng để tạo ra các lệnh tương ứng cho máy tính.

Nhận dạng các cử động của tay người là cách tự nhiên khi tương tác người – máy. Ngày nay, nhiều nhà nghiên cứu trong các học viện và ngành công nghiệp đang quan tâm đến hướng nghiên cứu này. Nó cho phép con người tương tác với máy rất dễ dàng và thuận tiện mà không cần phải mang thêm bất kỳ thiết bị ngoại vi nào (El-Sawah và cs., 2008).

Mục đích của bài báo là xây dựng một phương pháp nhận diện mẫu trong các khung hình thu trực tiếp từ camera theo thời gian thực để giải quyết bước thứ ba trong bài toán điều khiển máy tính từ xa nêu trên. Phương pháp nhận diện này sử dụng mô hình túi từ (bag-of-features, bag-of-words) (Heap & Hogg, 1996) kết hợp với phương pháp phân lớp bằng mạng nơron (Kolsch & Turk, 2004). Trong đó, mô hình túi từ được sử dụng để tạo ra các vector đặc trưng làm dữ liệu đầu vào cho mạng nơron. Phương pháp nhận diện này cần đảm bảo được tốc độ xử lý cao (để có thể thực hiện trong thời gian thực với dữ liệu thu trực tiếp từ camera) và có tính bền vững với một số dạng biến đổi của đối tượng (xoay hình, thay đổi kích thước và vị trí trong khung hình). Đối tượng nhận diện chính của thuật toán là cử chỉ tay người và một số đồ vật đơn giản.

Mô hình túi từ là một phương pháp biểu diễn đơn giản thường được sử dụng trong xử lý ngôn ngữ tự nhiên (natural language processing), tìm kiếm thông tin (information retrieval) và trong các phương pháp phân lớp văn bản (document classification) (Stenger, 2006).

Mô hình túi từ là mô hình thống kê cho phép sử dụng cùng với các phương pháp học tự động (Stenger và cs., 2001). Theo mô hình túi từ, dữ liệu văn bản không có cấu trúc (độ dài khác nhau) được biểu diễn tần số xuất hiện của từ trong văn bản dưới dạng một vector. Tập các dữ liệu văn bản được chuyển về dạng một bảng có số cột (chiều, từ vựng) rất lớn. Từ bảng dữ liệu này có thể huấn luyện các mô hình học máy tự động. Các mô hình máy học thường được sử dụng bao gồm giải thuật k-means (kNN), Naïve Bayes (NB), cây quyết định (decision tree – DT), support vector machine (SVM), boosting và random forest (RF) (Viola & Jones, 2004).

Mô hình túi từ cho phép biểu diễn tập dữ liệu văn bản về cấu trúc bảng. Bước tiền xử lý này bao gồm việc phân tích từ vựng và tách các từ trong nội dung của tập văn bản, chọn tập hợp các từ có ý nghĩa quan trọng dùng để phân loại, biểu diễn dữ liệu văn bản về dạng bảng để từ đó các giải thuật máy học có thể học để phân loại (Wang & Wang, 2008).

Có thể thấy, tập dữ liệu có thể chứa vài trăm văn bản, bộ từ điển có thể lên đến khoảng vài chục nghìn từ. Khi đó, các mô hình máy học như kNN, NB hay DT có thể xử lý kém hiệu quả (Wagner và cs., 2006).

Để khắc phục, người ta thường thực hiện việc rút gọn chiều dữ liệu. Phương pháp rút gọn có thể là lựa chọn những từ quan trọng nhất giúp phân biệt văn bản này với văn bản khác, hay phương pháp giảm chiều (Zhou & Huang, 2003).

Trong vài năm gần đây, bài toán nhận diện cử chỉ tay người vẫn nhận được sự quan tâm của giới nghiên cứu nhằm ứng dụng trong các phần mềm thực tế ảo hoặc điều khiển từ xa. Các giải pháp nhận diện cử chỉ bàn tay được áp dụng nhiều nhất là các biến thể khác nhau của mạng nơron tích chập, bao gồm mạng nơron tích chập đa luồng (Noreen và cs., 2021), mạng nơron tích chập dựa trên vùng (Soe & Naing, 2019), mạng nơron tích chập sâu (Qi và cs., 2021), mạng nơron tích chập DenseNet (de Oliveira và cs., 2019). Ngoài ra, các nhà nghiên cứu cũng cố gắng đưa các thuật toán nhận diện cử chỉ ra thiết bị nhúng với hiệu suất thấp (Yangüez Cervantes & Zapata-Jaramillo, 2021) để xây dựng hệ thống tương tác người dùng.

2. SỬ DỤNG MÔ HÌNH TÚI TỪ ĐỂ XÂY DỰNG BỘ MÔ TẢ CHO VẬT THỂ VÀ THUẬT TOÁN NHẬN DIỆN VẬT THỂ VỚI MẠNG NƠRON

Để thực hiện phân lớp với mạng nơron, bộ mô tả (descriptor) vật thể thường được biểu diễn bằng một vector có số chiều cố định và bằng số lượng nơron ở lớp đầu vào. Để tạo ra bộ mô tả này, có thể sử dụng nhiều loại đặc trưng: đường viền (contour), góc nghiêng hoặc điểm đặc biệt trên vật thể, vùng đặc biệt trên vật thể, v.v.. Việc lựa chọn đặc trưng này có ý nghĩa quan trọng liên quan đến đặc điểm của vật thể cần nhận diện và phương pháp phân lớp được sử dụng. Đối với bài toán nhận diện vật thể có hình dáng thay đổi như hình bàn tay, bộ mô tả cần có những đặc điểm như: bền vững với biến đổi xoay hình, di chuyển hình và thay đổi độ phóng đại hình vật thể. Ngoài ra, bộ mô tả này cần có kích thước cố định và mang tính đặc trưng cho lớp vật thể cần nhận diện.

Ý tưởng của của phương pháp mô tả vật thể này nằm ở chỗ, một hình vật thể được coi như một tài liệu văn bản, trong đó các đặc trưng được xem như các từ tạo thành văn bản. Tài liệu này được phân lớp dựa trên việc tính toán số lần xuất hiện của một số "từ khóa".

Để đưa ý tưởng này vào nhận diện vật thể, các đặc trưng của vật thể được trích ra từ một tập hợp hình ảnh (tập huấn luyện) và được chia thành các nhóm. Trong mỗi nhóm chọn ra một đặc trưng làm "đại diện" cho toàn bộ nhóm. Mỗi đặc trưng đại diện này sẽ được sử dụng làm một từ khóa. Tập hợp các từ khóa này tạo thành bộ "từ điển". Khi đối chiếu các đặc trưng trích ra từ một bức hình với các từ khóa trong từ điển sẽ thu được một biểu đồ (histogram) tần số của các từ khóa. Biểu đồ này là một vector có kích thước cố định và có thể sử dụng làm vector đầu vào cho các phương pháp nhận diện (vd: mạng nơron).

Một cách tổng quát, ý tưởng của mô hình túi từ khi áp dụng ở đây cho phép tạo ra một bộ mô tả của vật thể là một vector có kích thước cố định và có giá trị tương đối đặc trưng cho lớp vật thể. Bộ mô tả này sẽ được sử dụng làm vector đầu vào cho mạng nơron. Sơ đồ tổng quát của ý tưởng này được thể hiện trong Hình 1.

Để sử dụng bộ mô tả kết hợp với mạng nơron trong việc nhận diện vật thể, bài báo đề xuất một thuật toán thể hiện như trong Hình 2. Thuật toán này bao gồm các giai đoạn sau:

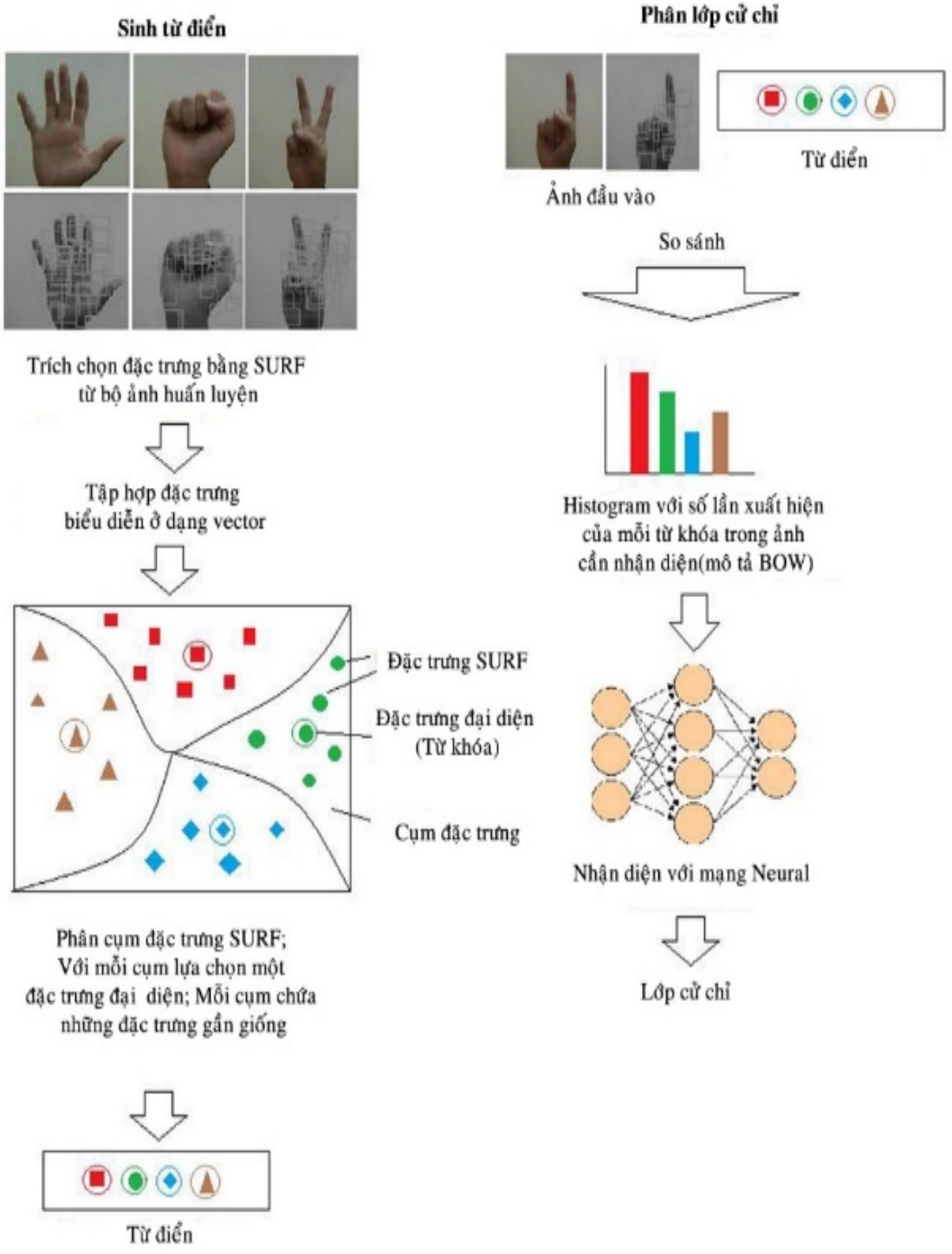
Giai đoạn 1. Huấn luyện

a) Sinh ra bộ từ điển:

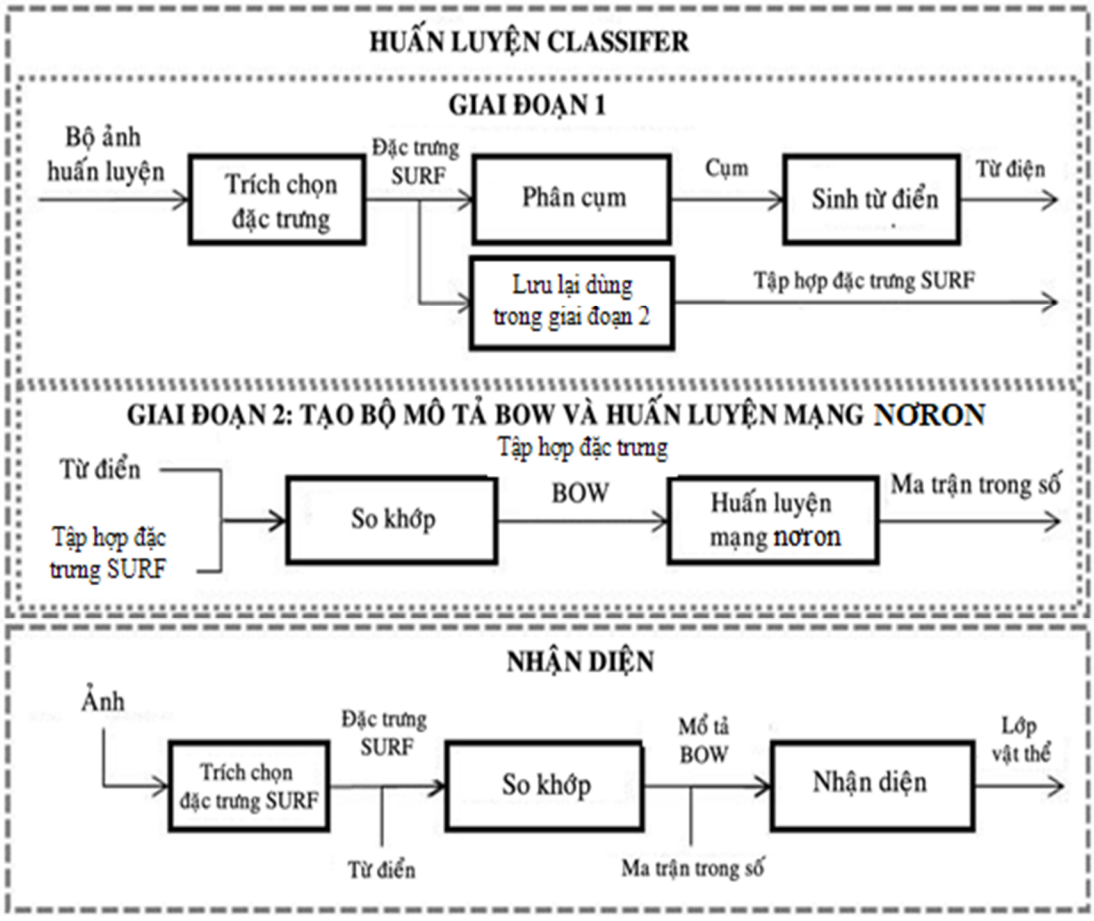
a.i) Trích tất cả đặc trưng từ tất cả các ảnh trong bộ ảnh huấn luyện với phương pháp SURF: Từ hình của mỗi vật thể thu được một số lượng tương đối lớn đặc trưng (ví dụ, đối với hình bàn tay mở có thể thu được từ 20 đến 100 đặc trưng). Mỗi đặc trưng này được mô tả bởi một vector 64 chiều gọi là mô tả SURF (SURF-descriptor).

a.ii) Phân cụm các đặc trưng thu được sử dụng thuật toán k-means: Do số lượng SURF-descriptor khá lớn và không cố định nên không thể trực tiếp sử dụng chúng với mạng nơron. Ở bước này, tất cả các SURF-descriptor được phân thành các cluster. Mỗi cluster chứa một loạt các SURF-descriptor có giá trị gần nhau.

a.iii) Sinh ra từ điển từ các cụm thu được: Đối với mỗi cluster chọn ra một SURF-descriptor đại diện cho cả nhóm. Tập hợp tất cả các SURF-descriptor đại diện và sắp xếp theo một trật tự cố định sẽ tạo ra một "từ điển". Bộ từ điển này sẽ được dùng làm cơ sở để tạo ra một loại mô tả mới phù hợp với các yêu cầu đặt ra. Bài báo sẽ gọi loại mô tả mới này là BOW-descriptor. BOW-descriptor sẽ được sử dụng làm vector đầu vào cho mạng nơron.



Hình 1. Mô tả ý tưởng của thuật toán nhận diện vật thể dựa trên mô hình túi từ



Hình 2. Sơ đồ tổng quát của thuật toán nhận diện đề xuất

b) Sinh các bộ mô tả và huấn luyện mạng nơron:

b.i) Ứng với mỗi ảnh trong bộ ảnh huấn luyện, trích chọn ra các đặc trưng bằng phương pháp SURF: Ở bước a.i. chúng ta đã gộp tất cả các đặc trưng trích được qua phương pháp SURF để phục vụ sinh từ điển. Tại bước này, chúng ta để riêng các SURF-descriptor của từng ảnh vật thể trong bộ huấn luyện nhằm tạo ra BOW-descriptor của ảnh này. Tập hợp các BOW-descriptor của tất cả các ảnh trong bộ huấn luyện sẽ được sử dụng làm dữ liệu huấn luyện của mạng nơron.

b.ii) Sinh ra BOW-descriptor của từng ảnh trong bộ ảnh huấn luyện: Tất cả SURF-descriptor thu được từ một ảnh sẽ được đối chiếu với từ điển. Mỗi SURF-descriptor sẽ được so sánh với một từ trong từ điển để tìm ra từ gần với nó nhất. SURF-descriptor sẽ được thay thế bằng từ tương đương trong từ điển. Khi đếm số lần xuất hiện của mỗi từ, ta sẽ thu được một histogram. Histogram này chính là bộ mô tả của ảnh vật thể chúng ta đang cần tìm (ở trên đã quy ước sẽ gọi là BOW-descriptor).

b.iii) Sử dụng tất cả các BOW-descriptor thu được này làm bộ dữ liệu huấn luyện để dạy cho mạng nơron.

Giai đoạn 2. Nhận diện

– Trích đặc trưng của vật thể dựa trên phương pháp SURF;

– Đối chiếu các đặc trưng thu được với từ điển để thu được BOW-descriptor của ảnh vật thể;

– Sử dụng bộ mô tả này làm dữ liệu đầu vào để nhận diện với mạng nơron đã được huấn luyện ở bước trên.

3. PHƯƠNG PHÁP NGHIÊN CỨU VÀ THÍ NGHIỆM

Phần này sẽ trình bày phương pháp thử nghiệm thuật toán nhận diện hình dạng bàn tay trên các tập dữ liệu khác nhau. Trong các thử nghiệm này, chúng tôi sử dụng các bộ dữ liệu do nhóm tác giả tự xây dựng.

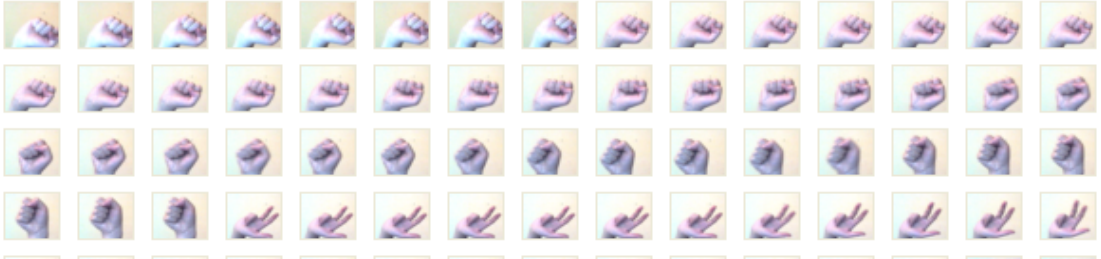
Hình 3. Các lớp vật thể trong bộ dữ liệu

Để huấn luyện thuật toán, hai bộ dữ liệu được tạo ra: một bộ chứa các hình ảnh "sạch" (chỉ chụp hình vật thể, không có nền, không có nhiễu) dùng để sinh ra từ điển, một bộ chứa các hình ảnh với nền và nhiễu dùng để huấn luyện. Trong giai đoạn test sử dụng ba bộ dữ liệu: một bộ với hình vật thể không chứa nhiễu, một bộ có nền đơn giản, một bộ có độ sáng kém với hình nền và các yếu tố nhiễu. Tất cả các bộ dữ liệu này chứa hình bàn tay thuộc bốn lớp: Fist, Open Palm, Palm, V-Shape (Hình 3) được thu từ USB web camera của máy tính.

Bộ dữ liệu "sạch" để sinh từ điển chứa 1160 hình ảnh thuộc bốn lớp, cụ thể như sau: lớp First – 269 hình, lớp Open Palm – 293 hình, lớp Palm – 284 hình, lớp V-Shape – 314 hình. Thuật toán đề xuất cho phép xử lý các hình ảnh với kích thước bất kỳ. Tuy nhiên, để tiện lợi trong việc xây dựng các bộ dữ liệu, tất cả các hình ảnh đều được chụp với cùng kích thước 100×100 pixel. Tất cả các hình trong bộ dữ liệu này chỉ chứa vật thể chụp ở nhiều góc nghiêng và khoảng cách khác nhau, nền trắng và không có nhiễu (Hình 4).

Để huấn luyện mạng nơron, đối với mỗi lớp vật thể trong bộ dữ liệu trên, bổ sung thêm 100 hình được chụp với nền đơn giản và nhiều mức độ chiếu sáng khác nhau để tăng thêm khả năng chịu nhiễu của thuật toán (Hình 5).

Hình 4. Một phần bộ dữ liệu dùng để sinh từ điển



Hình 5. Một phần bộ dữ liệu dùng để sinh từ điển

4. KẾT QUẢ NGHIÊN CỨU

Trong phần này sẽ trình bày các kết quả test sau:

- Test với bộ dữ liệu chứa các ảnh với nền đơn giản ở nhiều góc nghiêng và khoảng cách khác nhau;
- Test với bộ dữ liệu chứa các ảnh có nhiễu;
- Test với bộ dữ liệu chứa các ảnh bị nhiễu nặng.

Tất cả test được thực hiện trên máy tính notebook ASUS ULVT80, hệ điều hành Windows 7 64 bit, 4Gb RAM, camera tích hợp sẵn của máy với tốc độ thu 15 khung hình/giây.

4.1. Test với bộ dữ liệu chứa các ảnh với nền đơn giản ở nhiều góc nghiêng và khoảng cách khác nhau

Thí nghiệm này kiểm tra khả năng của thuật toán nhận diện với các ảnh có nền đơn giản ở các góc nghiêng, khoảng cách và kích thước hình khác nhau. Trong thí nghiệm này sử dụng các bộ dữ liệu test sau đây:

• Bộ dữ liệu gốc kích thước 120×120, mỗi lớp chứa 500 hình (Hình 6);

• Bộ dữ liệu kích thước 100×100, mỗi lớp chứa 500 hình (là các hình của bộ dữ liệu đầu tiên được thu nhỏ kích thước);

• Bộ dữ liệu kích thước 80×80, mỗi lớp chứa 500 hình (là các hình của bộ dữ liệu đầu tiên).

Kết quả thử nghiệm được tổng hợp trong Bảng 1.

Bảng 1. Kết quả thử nghiệm với ảnh mẫu kích thước khác nhau

Kích thước 120×120					
		Lớp			
		Fist	**Open Palm**	**Palm**	**V-Shape**
Nhận diện thành	Fist	493	0	0	0
	Open Palm	0	499	0	0
	Palm	0	0	494	0
	V-Shape	0	0	0	490
Không nhận diện		7	1	6	10
Kết quả					
Thời gian xử lý trung bình (ms)		39	42	39	40
Độ chính xác (%)		98.6	99.8	98.8	98.0
Độ chính xác trung bình		98.8 %			
Kích thước 100×100					
		Lớp			
		Fist	**Open Palm**	**Palm**	**V-Shape**
Nhận diện thành	Fist	492	0	0	0
	Open Palm	0	499	0	0
	Palm	0	0	493	0
	V-Shape	0	0	0	490
Không nhận diện		8	1	7	10
Kết quả					
Thời gian xử lý trung bình (ms)		28	31	29	30
Độ chính xác (%)		98.4	99.8	98.6	98.0
Độ chính xác trung bình		98.7 %			
Kích thước 80×80					
		Lớp			
		Fist	**Open Palm**	**Palm**	**V-Shape**
Nhận diện thành	Fist	490	0	0	0
	Open Palm	0	499	0	0
	Palm	0	0	491	0
	V-Shape	0	0	0	489
Không nhận diện		10	1	9	11
Kết quả					
Thời gian xử lý trung bình (ms)		14	16	15	15
Độ chính xác (%)		98.0	99.8	98.2	97.8
Độ chính xác trung bình		98.5 %			

** Thời gian xử lý bao gồm tổng thời gian trích đặc trưng của SURF, thời gian tính vector BOW, thời gian xử lý trong mạng nơron.*

Kết quả thực nghiệm trên cho thấy sự chênh lệch nhỏ về độ chính xác trung bình. Tuy nhiên, thời gian xử lý trung bình chênh lệch khá lớn (15 ms đối với bộ hình kích thước 80×80, 40 ms với bộ hình kích thước 120×120). Tốc độ xử lý này chấp nhận được để sử dụng trong thời gian thực (với camera có tốc độ thu 15 khung hình/giây thì thời gian xử lý mỗi khung hình không được vượt quá 40 ms, nếu không sẽ tạo ra tình trạng giật hình).

Thuật toán đề xuất đạt được kết quả nhận diện rất cao trong tình huống lý tưởng (một vật thể trên nền trơn) và không phụ thuộc vào khoảng cách chụp hình cũng như góc nghiêng của vật thể trong hình.

Thuật toán hoạt động thiếu hiệu quả khi số lượng đặc trưng SURF thu được quá ít đối với ảnh có kích thước nhỏ. Điều này cũng giúp đưa đến kết luận rằng, nếu một vật thể có bề mặt quá đơn giản (ví dụ: hình quả bóng tròn đồng màu), thuật toán không hoạt động hiệu quả do có quá ít đặc trưng trích ra được từ hình vật thể. Thuật toán hoạt động tốt hơn với các vật thể có hình dạng bề mặt phức tạp.

4.2. Test với bộ dữ liệu chứa các ảnh có nhiễu nhẹ

Bộ dữ liệu chứa các ảnh có nhiễu nhẹ (Hình 7) bao gồm 1000 ảnh cho mỗi lớp. Mỗi ảnh được chụp với độ sáng thấp trên nền đơn giản và có một số vật thể nhỏ khác. Kích thước mỗi ảnh trong bộ dữ liệu này là 100×100 pixel.

Hình 7. Một phần bộ dữ liệu thử nghiệm với nhiễu nhẹ

Kết quả thử nghiệm được tổng hợp trong Bảng 2.

Bảng 2. Kết quả thử nghiệm với ảnh có nhiễu nhẹ

		Lớp			
		Fist	**Open Palm**	**Palm**	**V-Shape**
Nhận diện thành	Fist	947	0	0	0
	Open Palm	0	983	0	1
	Palm	0	0	951	0
	V-Shape	1	0	0	935
Không nhận diện được		52	17	49	64
Kết quả					
Thời gian xử lý trung bình		31	34	30	32
Độ chính xác (%)		94.7	98.3	95.1	93.5
Độ chính xác trung bình		95.8 %			

Trong thử nghiệm này, quan sát thấy rằng, độ chính xác trung bình giảm nhẹ (so với các thử nghiệm trong phần trước), đồng thời tăng thời gian xử lý trung bình của mỗi bức hình. Điều này có thể được giải thích như sau: khi xuất hiện các vật thể khác và hình nền, số lượng đặc trưng SURF tìm thấy tăng lên, do đó làm tăng thời gian xử lý khi xây dựng mô tả BOW; những đặc trưng SURF thu được từ những đối tượng "lạ" (không phải từ vật thể) có ảnh hưởng xấu tới độ chính xác của thuật toán nhận diện.

Kết quả thử nghiệm này cũng đưa đến một kết luận quan trọng: phương pháp biểu diễn đặc trưng BOW có thể hoạt động mà không cần thực hiện phân tách riêng vật thể ra khỏi hình nền.

4.3. Test với bộ dữ liệu chứa ảnh bị nhiễu nặng

Đây là bộ dữ liệu được chụp trong điều kiện thật ở văn phòng với độ sáng không cố định, có lẫn các vật thể lớn khác, với nhiều góc nghiêng và kích thước khác nhau (từ 80×80 tới 120×120). Một phần của bộ dữ liệu này được trình bày ở Hình 8.

Hình 8. Một phần bộ dữ liệu thử nghiệm với nhiễu nặng

Kết quả thử nghiệm được tổng hợp ở Bảng 3.

Bảng 3. Kết quả thử nghiệm với ảnh có nhiễu nặng

		Lớp			
		Fist	**Open Palm**	**Palm**	**V-Shape**
Nhận diện thành	Fist	918	0	1	0
	Open Palm	0	965	0	2
	Palm	1	0	903	1
	V-Shape	1	0	1	918
Không nhận diện được		80	35	95	79
Kết quả					
Thời gian xử lý trung bình (ms)		34	37	34	35
Độ chính xác (%)		91.8	96.5	90.3	91.8
Độ chính xác trung bình		92.6 %			

Trong thử nghiệm này, độ chính xác đã giảm đáng kể và thời gian xử lý tăng lên so với các thử nghiệm trên nhưng nhìn chung, độ chính xác này là chấp nhận được. Nếu thuật toán nhận diện này được sử dụng cùng với một giải pháp theo dõi vật thể (object tracking) sẽ luôn đạt được kết quả tương tự như trong thử nghiệm thứ hai (do phương pháp theo dõi vật thể thường sẽ khoanh vùng được khu vực chỉ chứa vật thể).

5. SO SÁNH KẾT QUẢ NGHIÊN CỨU

Một giải pháp đề xuất ở (Noreen và cs., 2021) đã đạt được độ chính xác rất cao (trên 98%) sử dụng mạng nơron tích chập đa luồng nhận diện 6 loại cử chỉ.

Giải pháp khác cùng sử dụng mạng nơron tích chập dựa trên vùng (Soe & Naing, 2019) có khả năng nhận diện 10 cử chỉ trong thời gian thực được ứng dụng để điều khiển phần mềm VLC.

Một đề xuất sử dụng mạng nơron tích chập sâu (Qi và cs., 2021) được sử dụng để điều khiển robot từ xa thông qua cử chỉ tay.

Sử dụng mạng nơron tích chập DenseNet (de Oliveira và cs., 2019) cho phép người dùng tự định nghĩa các cử chỉ sử dụng trong trò chơi video với độ chính xác rất cao (97.89%).

Trong (Yangüez Cervantes & Zapata-Jaramillo, 2021) công bố một giải pháp nhận diện cử chỉ tay có thể triển khai trên các thiết bị nhúng hiệu suất thấp với độ chính xác tới 95.5%, hoạt động gần trong thời gian thực.

Các so sánh này cho thấy, thuật toán đề xuất đem lại hiệu suất hoạt động tương đương, có tốc độ xử lý cao (hoạt động được trong thời gian thực), không đòi hỏi các thiết bị thu hình đầu vào đắt tiền phức tạp.

6. KẾT LUẬN

Bài báo này đã trình bày một đề xuất về sử dụng mô hình túi từ để xây dựng bộ mô tả cho vật thể trong ảnh và xây dựng thuật toán nhận diện vật thể sử dụng bộ mô tả đề xuất kết hợp với mạng nơron.

Thuật toán đề xuất đạt được kết quả nhận diện rất cao trong tình huống lý tưởng (một vật thể trên nền trơn) và không phụ thuộc vào khoảng cách chụp hình cũng như góc nghiêng của vật thể trong hình.

Thuật toán hoạt động thiếu hiệu quả khi số lượng đặc trưng SURF thu được quá ít đối với ảnh có kích thước nhỏ. Điều này cũng giúp đưa đến kết luận rằng, nếu một vật thể có bề mặt quá đơn giản, thuật toán không hoạt động hiệu quả do có quá ít đặc trưng trích ra được từ hình vật thể. Thuật toán hoạt động tốt hơn với các vật thể có hình dạng bề mặt phức tạp.

Phương pháp biểu diễn đặc trưng BOW có thể hoạt động mà không cần thực hiện phân tách riêng vật thể ra khỏi hình nền. Như vậy môi trường làm việc là một yếu tố ảnh hưởng tới hiệu quả của công việc thử nghiệm thuật toán.

Thời gian xử lý của thuật toán đảm bảo hoạt động được trong thời gian thực. Với tốc độ xử lý trên 15 khung hình mỗi giây, thuật toán đề xuất có thể tích hợp vào các chương trình thu hình từ camera.

TÀI LIỆU THAM KHẢO

Argyros, A. A., & Lourakis, M. I. A. (2006). Vision-Based Interpretation of Hand Gestures for Remote Control of a Computer Mouse. Trong T. S. Huang, N. Sebe, M. S. Lew, V. Pavlović, M. Kölsch, A. Galata, & B. Kisačanin (B.t.v), *Computer Vision in Human-Computer Interaction* (tr 40–51). Springer. https://doi.org/10.1007/11754336_5

Barczak, A. L. C., & Dadgostar, F. (2005). Real-time hand tracking using a set of cooperative classifiers based on Haar-like features. *Research Letters in the Information and Mathematical Sciences*, *7*, 29–42.

Bretzner, L., Laptev, I., & Lindeberg, T. (2002). Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering. *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, 423–428. https://doi.org/10.1109/AFGR-.2002.1004190

Chen, Q., Georganas, N. D., & Petriu, E. M. (2007). Real-time Vision-based Hand

Gesture Recognition Using Haar-like Features. *2007 IEEE Instrumentation & Measurement Technology Conference IMTC 2007*, 1–6. https://doi.org/-10.1109/IMTC.2007.379068

de Oliveira, E., Clua, E. W. G., Vasconcelos, C. N., Marques, B. A. D., Trevisan, D. G., & de Castro Salgado, L. C. (2019). FPVRGame: Deep Learning for Hand Pose Recognition in Real-Time Using Low-End HMD. Trong E. van der Spek, S. Göbel, E. Y.-L. Do, E. Clua, & J. Baalsrud Hauge (B.t.v), *Entertainment Computing and Serious Games* (Vol 11863, tr 70–84). Springer International Publishing. https://doi.org/10.1007/978-3-030-34644-7_6

El-Sawah, A., Georganas, N. D., & Petriu, E. M. (2008). A Prototype for 3-D Hand Tracking and Posture Estimation. *IEEE Transactions on Instrumentation and Measurement*, *57*(8), 1627–1636. https://doi.org/10.1109/TIM.2008.925725

Heap, T., & Hogg, D. (1996). Towards 3D hand tracking using a deformable model. *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, 140–145. https://doi.org/10.1109/AFGR.1996.557255

Kolsch, M., & Turk, M. (2004). Analysis of rotational robustness of hand detection with a Viola-Jones detector. *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, *3*, 107–110. https://doi.org/10.1109-/ICPR.2004.1334480

Noreen, I., Hamid, M., Akram, U., Malik, S., & Saleem, M. (2021). Hand Pose Recognition Using Parallel Multi Stream CNN. *Sensors*, *21*(24), Art. 24. https://doi.org/10.3390/s21248469

Qi, W., Liu, X., Zhang, L., Wu, L., Zang, W., & Su, H. (2021). Adaptive sensor fusion labeling framework for hand pose recognition in robot teleoperation. *Assembly Automation*, *41*(3), 393–400. https://doi.org/10.1108/AA-11-2020-0178

Soe, H. M., & Naing, T. M. (2019). Real-Time Hand Pose Recognition Using Faster Region-Based Convolutional Neural Network. Trong T. T. Zin & J. C.-W. Lin (B.t.v), *Big Data Analysis and Deep Learning Applications* (Vol 744, tr 104–112). Springer. https://doi.org/-10.1007/978-981-13-0869-7_12

Stenger, B. (2006). Template-Based Hand Pose Recognition Using Multiple Cues. Trong P. J. Narayanan, S. K. Nayar, & H.-Y. Shum (B.t.v), *Computer Vision – ACCV 2006* (tr 551–560). Springer. https://doi.org/10.1007/11612704_55

Stenger, B., Mendonca, P. R. S., & Cipolla, R. (2001). Model-based 3D tracking of an articulated hand. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, *2*, II–II. https://doi.org/10.1109/CVPR.2001.990976

Viola, P., & Jones, M. J. (2004). Robust Real-Time Face Detection. *International Journal of Computer Vision*, *57*(2), 137–154. https://doi.org/10.1023/B:VISI.-0000013087.49260.fb

Wagner, S., Alefs, B., & Picus, C. (2006). Framework for a portable gesture interface. *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, 275–280. https://doi.org/10.1109/FGR.2006.54

Wang, C.-C., & Wang, K.-C. (2008). Hand Posture Recognition Using Adaboost with SIFT for Human Robot Interaction. Trong S. Lee, I. H. Suh, & M. S. Kim (B.t.v), *Recent Progress in Robotics: Viable Robotic Service to Human: An Edition of the Selected Papers from the 13th International Conference on Advanced Robotics* (tr 317–329). Springer. https://doi.org/10.1007/978-3-540-76729-9_25

Yangüez Cervantes, N., & Zapata-Jaramillo, C. M. (2021). Artificial Intelligence and Industry 4.0 Across the Continent: How AI and 4.0 are Addressed by Region. Trong D. Burgos & J. W. Branch (B.t.v), *Radical Solutions for Digital Transformation in Latin American Universities: Artificial Intelligence and Technology 4.0 in Higher Education* (tr 157–177). Springer. https://doi.org/-10.1007/978-981-16-3941-8_9

Zhou, H. & Huang. (2003). Tracking articulated hand motion with eigen dynamics analysis. *Proceedings Ninth IEEE International Conference on Computer Vision*, 1102–1109 vol2. https://doi.org/10.1109/ICCV.2003.1238472