NHẬN DẠNG CỬ CHỈ BÀN TAY DÙNG MẠNG NƠ-RON CHẬP

HAND GESTURE RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK

Lê Minh Thành[1], Lê Minh[1], Phan Văn Ca[1], Đặng Phước Hải Trang[1], Đỗ Duy Tân[1], Trương Ngọc Sơn[1*]

[1]Trường Đại học Sư phạm Kỹ thuật TP. Hồ Chí Minh

[*]Tác giả liên hệ: sontn@hcmute.edu.vn

(Nhận bài: 19/02/2021; Chấp nhận đăng: 15/4/2021)

Tóm tắt - Giao diện người – máy sẽ trực quan hơn nếu con người có thể điều khiển máy tính bằng giọng nói hay cử chỉ mà không cần dùng các thiết bị điều khiển như chuột hay bàn phím. Đặc biệt, hệ thống thị giác sẽ thích hợp hơn trong môi trường ồn ào hoặc có âm thanh bị nhiễu. Bên cạnh đó, mạng nơ-ron chập được áp dụng khá nhiều trong các bài toán nhận dạng với độ chính xác cao như nhận dạng gương mặt, nhận dạng số viết tay, xử lý ngôn ngữ tự nhiên. Bài báo này thiết lập một mạng nơ-ron chập với 14 lớp và ứng dụng vào hệ thống nhận dạng 6 cử chỉ bàn tay phải, với đối tượng đầu vào là các ảnh tĩnh thu được từ camera điện thoại. Tập dữ liệu huấn luyện được tạo ra từ các cử chỉ tay của 7 người. Kết quả mô phỏng trên matlab cho thấy hệ thống có tỷ lệ chính xác 98,6% đối với các ảnh bàn tay được chụp chính diện, có độ sáng và độ mở của các ngón tay thích hợp.

Từ khóa - Nhận dạng cử chỉ bàn tay; nơ-ron chập; CNN

Abstract - The human-machine interfaces will be more efficient when operated with voices or gestures without any hardware, such as mouse or keyboards. In particular, vision-based systems will be more appropriate in loud environments or environments with noises. In addition, the convolutional neural network has been applied more and more frequently in recognition problems with high accuracy such as face recognition, handwritten digits recognition, natural language processing. In this paper, we employed a convolutional neural network with 14 layers for the hand gesture recognition system with 6 different gestures of the right hand, and the input images were taken by a phone camera. The training data set was collected from the hand gesture of 7 people. The simulation results obtained using Matlab show that the system has an accuracy of 98.6% for hand images taken from the front with the appropriate brightness and suitable finger distance.

Key words - hand gesture recognition; convolutional neural network; CNN

1. Giới thiệu

Ngày nay, tự động hóa đã và đang dần thay thế các hoạt động của con người trong nhiều lĩnh vực. Với các yêu cầu thiết kế đòi hỏi độ chính xác cao, nhiều hệ thống đã có thể giúp con người tiếp cận đến những nơi mà tưởng chừng con người không đến được. Điều này thách thức một giao diện người – máy không những phải đạt hiệu quả cao về sự thông hiểu nhau mà còn phải đạt tốc độ xử lý nhanh chóng.

Giao diện người – máy cơ bản nhất được sử dụng thông qua bàn phím và chuột bị giới hạn bởi khoảng cách giữa người dùng với đối tượng cần tương tác [1]. Một số tương tác qua giọng nói đã đem lại nhiều tiện ích cho người dùng như điều khiển thiết bị thông qua giọng nói trong ngôi nhà thông minh [2], các vấn đề nhận dạng đối tượng cần thiết trong an ninh [3]... Tuy nhiên, các giao diện này bị giới hạn bởi các đặc trưng giọng nói theo vùng miền, từ đó dẫn đến việc thiết kế hệ thống phức tạp và khó được sử dụng phổ biến [4].

Nhận diện các cử chỉ bàn tay là phương pháp để xây dựng giao diện người dùng thân thiện giữa máy và người sử dụng. Trong tương lai gần, công nghệ nhận dạng cử chỉ bàn tay cho phép các máy phức hợp và các thiết bị thông minh hoạt động dựa trên tư thế bàn tay, ngón tay và sự di chuyển của bàn tay, loại bỏ việc giao tiếp vật lý giữa người và máy. Ngày nay, với sự phát triển của các thư viện mã nguồn mở trong lĩnh vực thị giác máy tính đã cho phép thiết kế các ứng dụng nhận dạng cử chỉ bàn tay dễ dàng hơn và có thể áp dụng các ứng dụng này rộng rãi vào nhiều lĩnh vực như y học [5], nhận dạng ngôn ngữ cử chỉ [6], điều khiển robot [7], thực tế ảo [8], điều khiển các thiết bị trong nhà [9] và các ứng dụng giải trí [10]. Giải thuật nhận dạng cử chỉ bàn tay được phát triển ban đầu dựa trên kỹ thuật xử lý ảnh và thị giác máy tính. Các giải thuật này chủ yếu dựa vào việc phân đoạn và tách đặt trưng của bàn tay dựa vào một số đặc trưng như màu da, khung xương, độ xâu, mô hình 3 chiều, hoặc nhận dạng dựa vào chuyển động [11]-[13]. Trong những năm gần đây, trí tuệ nhân tạo trong đó cụ thể là các mạng học sâu (Deep neural network) trở nên hiệu quả và được áp dụng trong nhiều ứng dụng như nhận dạng, phân loại ảnh, xử lý ngôn ngữ tự nhiên. Một trong những yếu tố chính là việc phát triển của công nghệ vi mạch cho phép các hệ thống máy tính có cấu hình mạnh ra đời đã tạo điều kiện cho việc thực thi các mạng nhiều lớp trở nên hiệu quả hơn trước. Song song với việc phát triển phần cứng cũng như các mạng học sâu, sự phát triển của các thư viện mã nguồn mở cho phép thiết kế cách mạng học sâu cho các ứng dụng cũng đa dạng và đơn giản hơn. Trong bài báo này, nhóm tác giả trình bày một thiết kế mạng nơ-ron tích chập cho bài toán nhận dạng cử chỉ bàn tay. Mạng nơ-ron tích chập được huấn luyện trên tập mẫu được nhóm tác giả tự tạo bao gồm 27,600 mẫu với 6 lớp khác nhau bao gồm "năm ngón tay khép kín", "năm ngón tay mở", "cử chỉ bốn ngón tay mở", "bàn tay nắm", "cử chỉ có ba ngón tay mở" và "cử chỉ có hai ngón tay mở", được đặt tên tương ứng từ class1 đến class6. Quá trình thực nghiệm cho thấy hệ thống có thể nhận dạng đạt độ chính xác lên đến 98,6%

[1] Ho Chi Minh City University of Technology and Education (Le Minh Thanh, Le Minh, Phan Van Ca, Dang Phuoc Hai Trang, Do Duy Tan, Son Ngoc Truong)

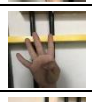
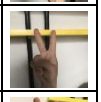
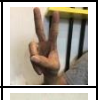
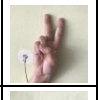
2. Thiết kế hệ thống nhận dạng cử chỉ bàn tay

2.1. Chuẩn bị tập dữ liệu huấn luyện

Hệ thống nhận dạng ở bài báo này được xây dựng để phân biệt được 6 loại cử chỉ bàn tay phải. Tập dữ liệu đầu vào cho quá trình huấn luyện được tạo dựa theo tập dữ liệu Cambride-Gestture Data Base [14] với 27,600 hình ảnh có kích thước 3024×3024 bao gồm các ảnh được chụp từ 7 người ở các điều kiện không quá sáng, không quá tối và ảnh nền khác nhau. Các bàn tay được chụp ở vị trí và tư thế khác nhau: Thẳng, nghiêng trái, nghiêng phải, gần và xa. Tập dữ liệu trên được chia thành 2 tập dữ liệu con là tập huấn luyện và tập kiểm tra với tỉ lệ tương ứng là 80% và 20%. Trong tập dữ liệu huấn luyện và tập kiểm tra có tất cả các trường hợp về tư thế và vị trí của các cử chỉ có trong tập dữ liệu, các tệp trong cả hai tập dữ liệu huấn luyện và kiểm tra không trùng nhau.

Ảnh đầu vào được giảm kích thước xuống còn 227×227 để phù hợp với mạng nơ-ron chập để tối ưu về thời gian và tài nguyên. Mẫu bàn tay của 7 người trong tập dữ liệu (tương ứng với số thứ tự từ 1 đến 7) được trình bày trong Bảng 1

Bảng 1. *Tập dữ liệu ngõ vào*

STT	Class1	Class2	Class3	Class4	Class5	Class6
1						
2						
3						
4						
5						
6						
7						

Bảng 1 liệt kê mẫu bàn tay của 7 người với 6 lớp cử chỉ khác nhau được chụp từ điện thoại để tạo tập dữ liệu cho quá trình huấn luyện.

2.2. Thiết kế kiến trúc mạng nơ-ron chập

Một mô hình mạng nơ-ron chập (Convolutional neural network) bao gồm 28 lớp với thông số chi tiết được trình bày ở Bảng 2 được thiết kế cho ứng dụng nhận dạng cử chỉ bàn tay.

Bảng 2. *Các thông số của mạng nơ-ron được đề xuất*

STT	Kiểu lớp	Thông số
1	Ảnh đầu vào	227×227×3
2	Lớp chập	96 bộ lọc kích thước 11×11×3 với bước trượt [4 4]
3	ReLU	Hàm kích hoạt
4	Cross Channel Normalization	Lớp chuẩn hóa
5	Max Pooling	Lớp gộp, cửa sổ 3×3
6	Lớp chập	256 bộ lọc kích thước 5×5×48 với bước trượt [1 1]
7	ReLU	Hàm kích hoạt
8	Cross Channel Normalization	Lớp chuẩn hóa
9	Max Pooling	Lớp gộp, của sổ 3×3 với bước trượt [2 2]
10	Chập	384 bộ lọc kích thước 3×3×256 với bước trượt [1 1]
11	ReLU	Hàm kích hoạt
12	Chập	384 bộ lọc với kích thước 3×3×192 với bước trượt [1 1]
13	ReLU	Hàm kích hoạt
14	Chập	256 Bộ lọc với kích thước 3×3×192 với bước trượt [1 1]
15	ReLU	Hàm kích hoạt
16	Max Pooling	kernel 3x3 với bước trượt [2 2]
17	Lớp kết nối đầy đủ	4096 nơ-ron
18	ReLU	Hàm kích hoạt
19	Dropout	50%
20	Lớp kết nối đầy đủ	4096 nơ-ron
21	ReLU	Hàm kích hoạt
22	Dropout	50%
23	Lớp kết nối đầy đủ	1000 nơ-ron
24	ReLU	Lớp hiệu chỉnh
25	Dropout	50%
26	Lớp kết nối đầy đủ	6 nơ-ron
27	Softmax	Phân bố xác suất ngõ ra
28	Ngõ ra nhãn	

Đầu vào của hệ thống là các ảnh màu 3 kênh R, G, B với kích thước 227×227×3 từ tập dữ liệu. Sau đó, ảnh đầu vào sẽ được nhân chập lần đầu tiên với 96 bộ lọc 3 chiều có kích thước 11×11×3 với bước trượt [4 4]. Các ảnh đặc trưng đầu ra (96 ảnh với kích thước 55×55×3) tiếp tục được đưa qua hàm kích hoạt ReLU và thực hiện chuẩn hóa chéo kênh (5 kênh/phần tử). Sau đó, các ảnh tiếp tục được đưa qua lớp gộp sử dụng hàm Max Pooling với bộ lọc 3×3 và bước trượt [2 2] và thu được các ảnh có kích thước 27×27×3. Các thao tác bao gồm nhân chập, kích hoạt với hàm ReLu và chuẩn hóa chéo kênh (5 kênh/phần tử) tiếp tục được thi 1 lần nữa trên các ảnh để thu được 256 ảnh đặc trưng với kích thước 11×11×3. Các ảnh này tiếp tục được nhân chập và kích hoạt bằng hàm ReLU thêm 3 lần nữa trước khi đi qua lớp Max Pooling lần cuối cùng. Kết quả thu được lúc này là 256 ảnh với kích thước 2×2×3. Sau đó, một lớp kết nối đầy đủ được sử dụng, theo sau là hàm kích hoạt ReLU và Dropout với tỉ lệ 50% để tránh hiện tượng quá khớp. Thao tác này sẽ được thực thi 3 lần với số lượng nơ-ron ở các lớp kết nối đầy đủ lần lượt là 4096, 4096 và 1000 nơ-ron. Cuối cùng, một lớp kết nối đầy đủ với 6 nơ-ron được sử dụng, theo sau

là hàm Softmax để cho ra ngõ ra cuối cùng của hệ thống để tạo ra phân bố xác xuất. Ngõ ra cuối cùng gồm 6 nhãn, tương ứng với 6 cử chỉ tay cần nhận dạng (class1 – class6).

3. Kết quả thực nghiệm

3.1. Kết quả quá trình huấn luyện mạng

Độ chính xác và giá trị mất mát trong quá trình huấn luyện và kiểm tra với số lần lặp lại từ 100 đến 882 lần được trình bày trong Bảng 3

Bảng 3. *Quá trình huấn luyện và kiểm tra*

Số lần lặp lại	Thời gian (s)	Giá trị mất mát trên tập huấn luyện	Giá trị mất mát trên tập kiểm tra	Độ chính xác trên tập huấn luyện (%)	Độ chính xác trên tập kiểm tra (%)
1	1,97	2,3371	2,3371	10	17,30
100	238,99	0,2921	0,4770	88	81,18
200	465,30	0,0508	0,4093	100	83,97
300	692,01	0,1436	0,4549	96	81,47
400	923,21	0,0058	0,4239	100	84,17
500	1157,13	0,0088	0,3034	100	88,51
600	1391,19	0,0015	0,3487	100	87,64
700	1620,36	0,0127	0,2780	100	89,08
800	1852,25	0,0015	0,4406	100	85,89
882	2052,19	0,0002	0,2561	100	99,17

Dựa vào kết quả liệt kê trong Bảng 3 có thể thấy, sau 882 lần huấn luyện tỉ lệ nhận dạng chính xác trên tập kiểm tra đã cải thiện từ 17,3% ở lần huấn luyện đầu tiên lên tới 99,17% ở lần huấn luyện thứ 882. Các thông số mô hình ở lần huấn luyện cuối cùng sẽ được sử dụng để thực nghiệm trên hệ thống cho việc nhận dạng trong thời gian thực.

3.2. Kết quả nhận dạng các ảnh trong tập kiểm tra

Kết quả nhận dạng các ảnh trong tập kiểm tra theo từng cử chỉ với tổng số 920 tệp được trình bày trong ma trận tương quan ở Bảng 4.

Bảng 4. *Kết quả nhận dạng theo từng cử chỉ*

	Class1	Class2	Class3	Class4	Class5	Class6	Tỷ lệ (%)
Class1	920	0	0	0	0	0	100
Class2	0	914	1	0	0	0	99,34
Class3	0	0	909	0	0	0	98,8
Class4	0	50	0	920	0	0	100
Class5	0	0	0	0	860	22	93,47
Class6	0	0	0	0	0	920	100

Với tập kiểm tra gồm 920 ảnh, tỉ lệ nhận dạng đúng cao nhất là 100% xảy ra ở 3 trạng thái ngõ ra là Class1 (năm ngón tay khép kín), Class4 (bàn tay nắm) và Class 6 (cử chỉ có hai ngón tay mở), tỉ lệ nhận dạng đúng thấp nhất ở ngõ ra Class5 (cử chỉ có ba ngón tay mở) với độ chính xác là 93,47%. Tính trung bình, tỉ lệ nhận dạng chính xác cho cả 6 trạng thái ngõ ra đối với tập dữ liệu kiểm tra là 98.6%.

Để đánh giá độ tin cậy của hệ thống, nhóm tác giả tiến hành kiểm tra quá trình nhận dạng của hệ thống đối với các ảnh cử chỉ bàn tay có độ mở của các ngón tay khác nhau và không trùng với các ảnh đã có trong cơ sở dữ liệu. Việc kiểm tra được thực hiện với ảnh chụp trực tiếp từ camera và được thử hiện 100 lần. Kết quả được trình bày trong Bảng 5.

Bảng 5. *Kết quả nhận dạng các độ mở khác nhau của cử chỉ*

Mẫu			
Kết quả	Class1	Class1	Class1
Tỷ lệ (%)	100	99,99	100
Mẫu			
Kết quả	Class2	Class2	Class2
Tỷ lệ (%)	100	99,42	100
Mẫu			
Kết quả	Class3	Class3	Class3
Tỷ lệ (%)	99,97	99,8	99,06
Mẫu			
Kết quả	Class3	Class5	Class5
Tỷ lệ (%)	91	97,88	99,74
Mẫu			
Kết quả	Class6	Class6	Class6
Tỷ lệ (%)	98,99	99,22	99,24

Như vậy, với ngón tay có các độ mở khác nhau, kết quả nhận dạng của hệ thống là chính xác nhất ở các trạng thái ngõ ra Class1 (99,99%), Class2 (99,8%), và Class6 (99,15%).

4. Kết luận

Trong bài báo này, nhóm tác giả đã đề xuất một mô hình mạng nơ-ron chập ứng dụng cho việc nhận dạng 6 cử chỉ bàn tay với ảnh đầu vào được chụp trực tiếp từ camera. Tập dữ liệu được nhóm tạo ra với 6 lớp cử chỉ bàn tay khác nhau. Kết quả kiểm chứng cho thấy, hệ thống có thể nhận dạng tốt, có tỉ lệ đúng trung bình lên tới 98,6%, với các ảnh đầu vào có điều kiện ánh sáng, góc chụp và độ mở ngón tay thích hợp. Mô hình mạng nơ-ron tích chập đề xuất cho ứng dụng nhận dạng cử chỉ bàn tay có thể được ứng dụng trong các hệ thống điều khiển không tiếp xúc, ứng dụng chuyển đổi ngôn ngữ cử chỉ sang văn bản hoặc trong các ứng dụng điều khiển thông minh khác.

Lời cảm ơn: Bài báo là sản phẩm của đề tài cấp trường trọng điểm mã số T2020 – 44TĐ được hỗ trợ bởi trường Đại học Sư phạm Kỹ thuật TP.HCM.

TÀI LIỆU THAM KHẢO

[1] Oyebade Oyedotun and Adnan Khashman, "Deep learning in vision-based static hand gesture recognition", *Neural Computing and Applications*, vol. 28, Apr. 2016.

[2] Deval G. Patel, "Point Pattern Matching Algorithm for Recognition of 36 ASL Gestures", *International Journal of Science and Modern Engineering (IJISME),* vol. 1, no. 7, June 2013.

[3] Dennis Núñez Fernández and Bogdan Kwolek, *"Hand Posture Recognition Using Convolutional Neural Network*", Polish National Science Center (CNN), Dec. 2014.

[4] Aashni Haria, Archanasri Subramanian, Nivedhitha Asokkumar, Shristi Poddar, and Jyothi Nayak, "Hand Gesture Recognition for Human Computer Interaction", *Procedia Computer Science*, vol. 115, pp. 367-374, Dec. 2017.

[5] J. P. Wachs, M. Kölsch, H. Stern, Y. Edan, "Vision-based hand-gesture applications" *Commun. ACM 2011*, vol. 54, no. 2, pp. 60–71, 2011.

[6] J.R. Pansare, S. H. Gawande, M. Ingle, "Real-time static hand gesture recognition for American Sign Language (ASL) in complex background", *Journal of Signal and Information Processing*, vol. 3, no. 2, Aug. 2012.

[7] M. Van den Bergh, D. Carton, R. De Nijs, N. Mitsou, C. Landsiedel, K. Kuehnlenz, D. Wollherr, L. Van Gool, M. Buss, "Real-time 3D hand gesture interaction with a robot for understanding directions from humans", *Proceedings of the 2011 Ro-Man, Atlanta, GA, USA,* 31 July–3 August, pp. 357–362, 2011.

[8] R.Y. Wang, J. Popovi´c, "Real-time hand-tracking with a color glove", *ACM Trans. Graph*., vol. 28, pp. 1–8, 2009.

[9] S. Desai, A. Desai, "Human Computer Interaction through hand gestures for home automation using Microsoft Kinect", *Proceedings of the International Conference on Communication and Networks*, Xi'an, China, 10–12 October, pp. 19–29, 2017.

[10] H. Kaur, J. Rani, "A review: Study of various techniques of Hand gesture recognition". *Proceedings of the 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, Delhi, India, pp. 1–5, Jul. 2016.

[11] G. R. S. Murthy, R. S. Jadon, "A review of vision based hand gestures recognition", *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, pp. 405–410, 2009.

[12] R. Z. Khan, N. A. Ibraheem, "Hand gesture recognition: A literature review". *Int. J. Artif. Intell. Appl.*, vol. 3, pp. 161-174, 2012.

[13] J. Suarez and R. R. Murphy, "Hand gesture recognition with depth images: A review", *The 21st IEEE International Symposium on Robot and Human Interactive Communication*, Paris, France, pp. 411-417, 2012.

[14] T-K. Kim, S-F. Wong and R. Cipolla, Tensor Canonical Correlation Analysis for Action Classification, *In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Minneapolis, MN, 2007.