

NHẬN DẠNG CỬ CHỈ TAY BẰNG RA-ĐA FMCW DỰA TRÊN MẠNG NƠ-RON TÍCH CHẬP KẾT NỐI CHÉO

Lê Hải^{1*}, Hoàng Văn Phúc¹, Đoàn Văn Sáng², Lê Thị Trang³, Lê Đại Phong¹

Tóm tắt: Nghiên cứu này đề xuất sử dụng mạng nơ-ron tích chập kết nối chéo (Cross-CNN: Cross-connection Convolutional Neural Network) để nhận dạng cử chỉ tay dựa trên dữ liệu phân tích phổ Doppler vi mô (micro-Doppler) của ra-đa FMCW (Frequency Modulated Continuous Wave). Thêm vào đó, các mức nhiễu khác nhau được thêm vào dữ liệu để đánh giá đầy đủ hơn đối với mô hình đề xuất. Kết quả thử nghiệm cho thấy, mô hình được huấn luyện với dữ liệu có nhiễu cho chất lượng nhận dạng tốt hơn mô hình được huấn luyện với dữ liệu không có nhiễu. Sau đó, mô hình Cross-CNN được phân tích, đánh giá khi thay đổi các tham số cấu trúc, từ đó chọn ra bộ tham số phù hợp nhất cho bài toán đề ra. Cuối cùng, mô hình Cross-CNN được đánh giá, so sánh với các mô hình hiện có khác khi thực hiện trên cùng một tập dữ liệu. Kết quả là, mạng Cross-CNN cho khả năng nhận dạng cử chỉ tay tốt hơn các mô hình khác nhờ vào các kết nối chéo cho phép sử dụng kết hợp các đặc tính cũ với đặc tính mới trong quá trình huấn luyện mạng.

Từ khóa: Mạng nơ-ron tích chập; Nhận dạng cử chỉ tay; Ra-đa FMCW; Phổ micro-Doppler.

1. GIỚI THIỆU

Hiện nay, với sự phát triển bùng nổ của các hệ thống và thiết bị thông minh, việc điều khiển không tiếp xúc các thiết bị điện tử, gia dụng đang trở nên phổ biến. Theo đó, sự chuyển động của tay, chân, cơ thể hoặc sự thay đổi nét mặt sẽ là những dấu hiệu để các mô-đun giao tiếp người-máy (HMI: human-machine interface) thực hiện việc nhận dạng hoặc phân loại [1]. Các mô-đun HMI thường sử dụng camera quang học để nhận dạng là chủ yếu. Nhưng tính bảo mật riêng tư của dữ liệu hình ảnh lại trở thành vấn đề thách thức đối với các nhà phát triển thiết bị. Một giải pháp thay thế cho camera quang học, đó là ra-đa FMCW (Frequency Modulated Continuous Wave), vì nó có tính bảo mật, xử lý nhanh, và tiết kiệm năng lượng [2]. Năm bắt được lợi thế đó, hiện nay, các hãng sản xuất điện thoại đã bắt đầu sử dụng ra-đa FMCW để nhận biết cử chỉ của con người nhằm thực hiện một số tác vụ trên điện thoại mà không cần tiếp xúc. Điển hình như công nghệ Soli được phát triển bởi Google cho đồng hồ thông minh và điện thoại [3].

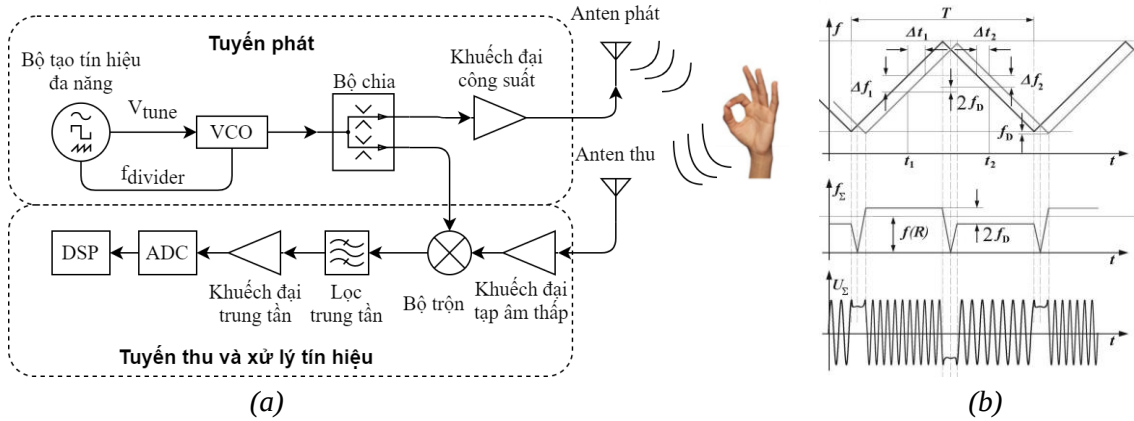
Gần đây, có nhiều công bố sử dụng thuật toán trí tuệ nhân tạo (AI: Artificial Intelligence), đặc biệt là các mô hình học máy (ML: Machine Learning) và học sâu (DL: Deep Learning), để nhận dạng cử chỉ tay dựa vào các đặc trưng phổ micro-Doppler thu được từ ra-đa FMCW. Điển hình như Scherer và cộng sự [4] đã sử dụng kết hợp các mạng nơ-ron tích chập (CNN: Convolutional Neural Network) không gian và thời gian để nhận dạng cử chỉ bằng ra-đa cự ly ngắn. Mô hình này đã đạt được độ chính xác 86,6% trên tập dữ liệu 26 người dùng với 11 cử chỉ, và 92,4% với 11 người dùng. Đặc biệt, mô hình CNN được thiết kế rất nhẹ, chỉ chiếm 92 kB bộ nhớ. Trong một nghiên cứu khác, Amin và cộng sự [5] đã giới thiệu một kỹ thuật đơn giản nhưng hiệu quả cho bài toán nhận dạng cử chỉ tay bằng cảm biến ra-đa. Theo đó, kỹ thuật đề xuất có thể phân loại các cử chỉ tay dựa trên đường bao của dấu hiệu micro-Doppler. Bằng cách này, mô hình đề xuất trong [5] đạt độ chính xác nhận dạng lên đến 96%. Bên cạnh những nghiên cứu về mô hình thuật toán, Ritchie và cộng sự [6] đã công bố một bộ dữ liệu về 4 cử chỉ tay được thu thập bằng ra-đa FMCW ở tần số 24 GHz, độ rộng băng thông 750 MHz. Tập dữ liệu này được kiểm tra thử với các thuật toán ML. Kết quả cho thấy, kỹ thuật SVM bậc hai (Support Vector Machine Quadratic) đạt tỉ lệ nhận dạng chính xác cao nhất (74,2%). Tập dữ liệu này được chia sẽ công khai và làm cơ sở để các nhà nghiên cứu đề xuất các mô hình cạnh tranh khác.

Trên cơ sở nghiên cứu khảo sát, trong bài báo này, nhóm tác giả đã đề xuất một mô hình CNN mới, tên là Cross-CNN, nhằm nâng cao hiệu năng nhận dạng cử chỉ tay bằng ra-đa FMCW.

Mô hình đề xuất sử dụng các cấu trúc kết nối chéo (Cross-connection) và kết nối bỏ qua (Skip-connection) để sử dụng lại các đặc tính cũ trong quá trình huấn luyện, nhờ vào đó, hiệu năng của mô hình được cải thiện đáng kể. Mô hình Cross-CNN được huấn luyện và kiểm chứng trên tập dữ liệu Dop-Net được công bố trong [6]. Khi so sánh với một số mô hình khác, mô hình Cross-CNN mà chúng tôi đề xuất đạt hiệu năng vượt trội hơn các mô hình khác cả về độ chính xác nhận dạng, thời gian thực thi và kích thước bộ nhớ.

2. RA-ĐA FMCW VÀ PHÂN TÍCH PHỔ MICRO-DOPPLER

Ra-đa FMCW là thiết bị được sử dụng phổ biến cho mục đích đo cự ly và vận tốc chuyển động của mục tiêu tầm gần. Ở phần này, nhóm tác giả sẽ làm rõ nguyên lý hoạt động của ra-đa FMCW làm cơ sở để phân tích phổ micro-Doppler. Sơ đồ nguyên lý hoạt động của ra-đa FMCW được minh họa như trong hình 1(a). Theo đó, tại tuyến phát, bộ tạo tín hiệu đa năng sẽ tạo tín hiệu điều khiển cho bộ VCO (Voltage Controlled Oscillator) phát ra tín hiệu điều chế tần số liên tục (FMCW). Tín hiệu này sẽ được chia làm hai nhánh: nhánh thứ nhất đi qua bộ khuếch đại công suất để đến Anten phát, từ đó phát xạ tín hiệu vào không gian; nhánh thứ hai được đưa đến bộ trộn ở tuyến thu. Tín hiệu FMCW truyền đi trong không gian, nếu gặp vật cản (mục tiêu) nó sẽ phản xạ trở lại và đưa tới Anten thu. Tín hiệu từ Anten sẽ được khuếch đại bằng khối khuếch đại tạp âm thấp, rồi đi đến bộ trộn. Tại đây, bằng việc trộn tần giữa tín hiệu thu và bản sao của tín hiệu phát, đầu ra của bộ trộn sẽ là tín hiệu ở trung tần. Tín hiệu trung tần này tiếp tục được khuếch đại và biến đổi từ tương tự sang số nhờ bộ ADC (Analog to Digital Converter). Cuối cùng, dữ liệu được thu thập và xử lý tại bộ xử lý tín hiệu số DSP (Digital Signal Processor).



Hình 1. Nguyên lý hoạt động của một ra-đa FMCW: (a) Sơ đồ nguyên lý;

(b) Tương quan tín hiệu thu so với tín hiệu phát.

Hình 1(b) minh họa mối tương quan giữa tần số và thời gian của tín hiệu thu so với tín hiệu phát. Trường hợp tín hiệu FMCW có dạng tam giác thì cự ly của mục tiêu được tính theo công thức sau:

$$R = \frac{c \cdot f(R)}{2 \left(\frac{df}{dt} \right)} = \frac{c \cdot (\Delta f_1 + \Delta f_2)}{4 \left(\frac{df}{dt} \right)}, \quad (1)$$

trong đó, c là vận tốc của sóng điện từ trong không gian, df/dt là độ biến thiên tần số theo thời gian của tín hiệu. Tương tự, độ dịch tần Doppler cũng được xác định như sau:

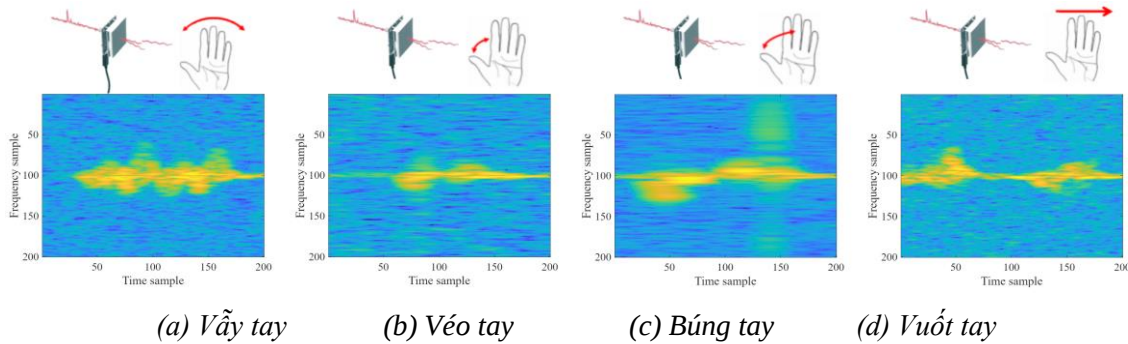
$$f_D = \frac{|\Delta f_1 - \Delta f_2|}{2} \quad (2)$$

Xem xét phân tích phổ micro-Doppler, giả sử vận tốc xuyên tâm của mục tiêu thay đổi theo

một hàm phụ thuộc vào thời gian, khi đó, mục tiêu sẽ gây ra tần số Doppler thay đổi theo thời gian. Đặt $v = v_0 + a(t)t$, với $a(t)$ là gia tốc biến thiên theo thời gian, v_0 là vận tốc ban đầu, thì tần số Doppler là:

$$f_D = \frac{2(v_0 + a(t)t)}{\lambda} = \frac{2v_0}{\lambda} + \frac{2a(t)t}{\lambda} = f_{D0} + f_D(t). \quad (3)$$

Có thể thấy, thành phần thứ nhất của biểu thức (4) là tần số Doppler do chuyển động tịnh tiến gây ra, còn thành phần thứ hai là tần số Doppler do chuyển động có gia tốc gây ra.



Hình 2. Ảnh phổ micro-Doppler của bốn cử chỉ tay trong tập dữ liệu Dop-Net.

Xét đến cử chỉ của tay, mỗi cử chỉ có chuyển động khác nhau sẽ tạo ra phổ Doppler khác nhau, như chỉ tra trong hình 2. Các phổ này là những dấu hiệu hữu ích để một mạng nơ-ron học sâu có thể tự động nhận dạng. Theo đó, bằng cách sử dụng biến đổi Fourier thời gian ngắn (STFT: Short Time Fourier Transform) của tín hiệu trung tần ở máy thu có thể tạo được hình ảnh micro-Doppler và từ đó phân loại được chúng [7]. Phương trình tổng quát của phép biến đổi STFT được mô tả như sau:

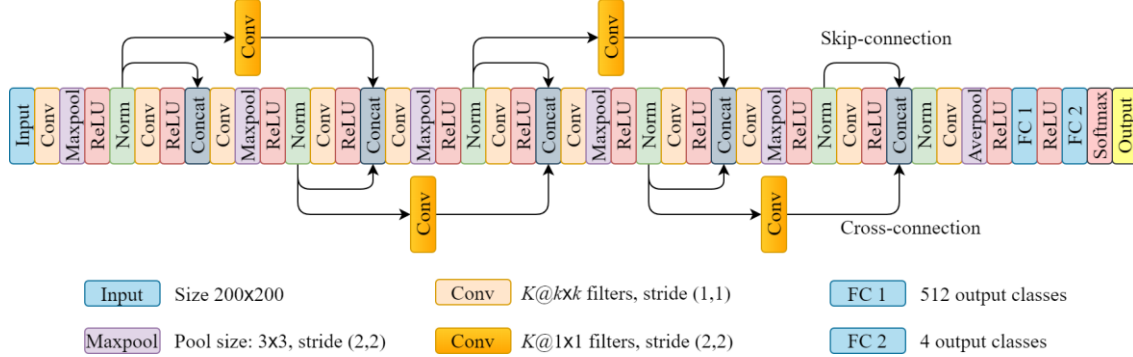
$$X(\tau, \omega) = \text{STFT}\{x(t)\} = \int_{-\infty}^{+\infty} x(t)w(t-\tau)e^{-j\omega t} dt, \quad (4)$$

trong đó, $x(t)$ là tín hiệu đầu vào cho biến đổi STFT $\{\cdot\}$, $w(t-\tau)$ là một hàm cửa sổ (window function).

3. MÔ HÌNH CROSS-CNN

Để nhận dạng các cử chỉ tay, nghiên cứu này đề xuất một mô hình CNN sử dụng cách kết nối Cross-connection và Skip-connection, tên là Cross-CNN, như được thể hiện như trong hình 3. Từ hình vẽ có thể thấy, mô hình Cross-CNN có một mạch kết nối chính từ lớp đầu vào (Input) đến lớp đầu ra (Output) thông qua các lớp được kết nối liên tục gồm lớp tích chập (Conv), lớp gộp cực đại (Maxpool), lớp kích hoạt (ReLU), lớp chuẩn hóa (Norm), và lớp ghép theo chiều sâu (Concat). Bên cạnh mạch chính, chúng tôi thiết kế thêm hai loại kết nối khác gồm Skip-connection và Cross-connection. Trong đó, Skip-connection được sử dụng để lấy bản đồ đặc tính từ lớp Norm để ghép theo độ sâu kênh với đầu ra của lớp ReLU tại lớp Concat. Như thể hiện trong hình 3, sau khi bản đồ đặc tính được chuẩn hóa bởi lớp Norm, nó sẽ tiếp tục được trích chọn (lọc) bởi lớp Conv và lớp ReLU. Trong quá trình lọc, một số đặc tính hữu ích có thể bị triệt tiêu, vì vậy sử dụng Skip-connection nhằm mục đích lấy lại đặc tính cũ cho lần trích chọn tiếp theo. Tương tự như Skip-connection, Cross-connection cũng được sử dụng với mục đích sử dụng lại các đặc tính cũ, cái mà có thể bị triệt tiêu trong quá trình trích chọn của mạng nơ-ron. Tuy nhiên, Cross-connection thực hiện kết nối xa hơn so với Skip-connection và thực hiện đan chéo nhau. Vì phải nhảy qua lớp Maxpool mà tại đó thực hiện giảm bản đồ đặc trưng đi 2 lần theo mỗi chiều (rộng và dài), nên Cross-connection sử dụng một lớp Conv có K bộ lọc kích thước 1×1 ,

bước nhảy (2, 2) nhằm tạo ra kích thước bản đồ đặc tính bằng với kích thước đầu ra của lớp ReLU mà không làm thay đổi đặc tính cũ.



Hình 3. Sơ đồ mô hình Cross-CNN cho bài toán nhận dạng chữ chỉ tay.

4. KẾT QUẢ ĐÁNH GIÁ THỬ NGHIỆM

Mô hình Cross-CNN được đánh giá thử nghiệm trên tập dữ liệu Dop-Net [6] gồm bốn cử chỉ tay được thu bởi một ra-đa FMCW. Dữ liệu được xử lý để tạo ảnh phổ Doppler với kích thước 200×200 , như thể hiện trong hình 2. Vì vậy, lớp đầu vào (Input) của Cross-CNN cũng phải được chỉ định kích thước 200×200 . Do tập dữ liệu có bốn cử chỉ tay nên lớp FC2 (lớp kết nối đầy đủ) được thiết kế với 4 phân lớp đầu ra. Ngoài ra, để tăng thách thức cho mạng Cross-CNN, chúng tôi đã thêm nhiễu Gauss cho tập dữ liệu với các giá trị SNR từ -20 dB đến 20 dB với bước cách 2 dB. Quá trình huấn luyện mạng Cross-CNN là quá trình cập nhật trọng số của mạng gồm hai tiến trình truyền thẳng (Forward propagation) và truyền ngược (Backward propagation) được lặp lại nhiều lần sau mỗi vòng lặp [8]. Quá trình truyền thẳng được coi là giai đoạn kiểm thử, còn quá trình truyền ngược là giai đoạn cập nhật trọng số. Trong sự truyền thẳng, đầu ra của mỗi lớp có thể được khái quát bằng công thức sau:

$$\mathbf{y}_l = f_l(\mathbf{w}_l * \mathbf{x}_l + \mathbf{b}_l), \quad (5)$$

trong đó, $\mathbf{x}_l$, $\mathbf{y}_l$, $\mathbf{w}_l$, $\mathbf{b}_l$, và $f_l(\cdot)$ lần lượt là giá trị đầu vào, đầu ra, trọng số, bias và hàm phi tuyến của lớp thứ l . Ở lớp cuối cùng, đầu ra $\mathbf{y}_{out}$ sẽ được so sánh với nhãn đúng $\mathbf{y}_{true}$ để tính hàm mất mát (Loss function). Nghiên cứu này thực hiện phân loại bốn cử chỉ tay, vì vậy, hàm mất mát Cross-entropy kết hợp điều chỉnh L2 (L2 regularization) được lựa chọn [9]. Trong trường hợp này, hàm mất mát được mô tả như sau:

$$L_{CE} = -\sum_i^C y_{i,true} \log(y_{i,out}) + \lambda \sum_i^C w_i^2 \quad (6)$$

trong đó, λ gọi là tham số điều chỉnh, C là số lượng phân lớp đầu ra, w là trọng số. Sau khi đã xác định được hàm mất mát, quá trình truyền ngược sẽ được thực hiện để cập nhật trọng số. Trong quá trình này, thuật toán SGD (Stochastic Gradient Descent) [10] được sử dụng như sau:

$$\mathbf{w}_{updated} = \mathbf{w}_{old} - \eta \frac{dL_{CE}}{d\mathbf{w}} \quad (7)$$

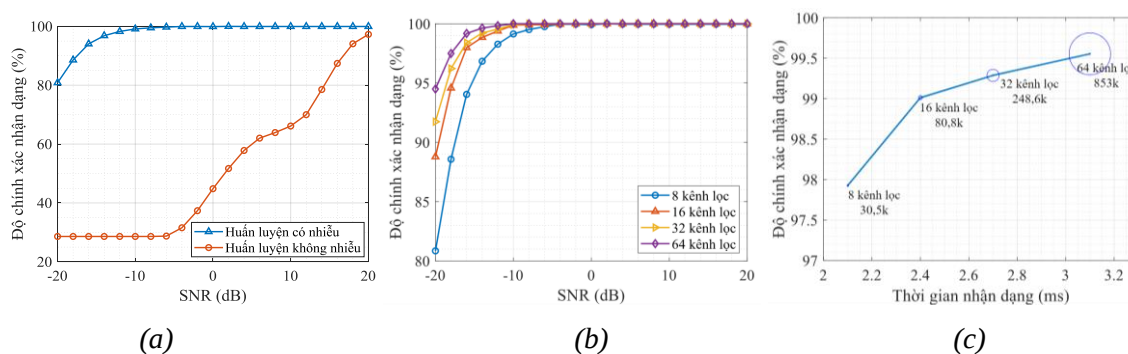
trong đó, η là tỉ lệ học.

Mô hình Cross-CNN được huấn luyện trong 10 epoches; tốc độ học khởi tạo là 0,01; tốc độ này sẽ giảm đi 10 lần sau bốn epoches; tham số điều chỉnh L2 là $\lambda = 1 \times 10^{-4}$. Quá trình huấn luyện được thực hiện với GPU của máy tính Laptop có cấu hình: CPU Core i5 9300H, RAM 16GB - 2667 MHz, GPU GTX1660ti 8GB. Tập dữ liệu được chia thành 80% cho huấn luyện và 20% cho kiểm chứng mô hình.

4.1. Đánh giá hiệu năng của Cross-CNN khi huấn luyện với các tập dữ liệu khác nhau

Mô hình CNN là phương pháp hướng dữ liệu, nghĩa là dữ liệu càng phong phú thì hiệu năng nhận dạng càng tốt. Trong phép thử đầu tiên, mô hình Cross-CNN sẽ được huấn luyện với hai tập dữ liệu: tập thứ nhất không có nhiễu và tập thứ 2 có nhiễu với SNR từ -20 dB đến 20 dB, bước cách 2 dB. Sau khi được huấn luyện trong 10 epoches. Hai mô hình được kiểm tra với dữ liệu có nhiễu. Cần lưu ý ở đây rằng, tập dữ liệu dùng để kiểm tra không được đưa vào huấn luyện. Cả hai mô hình đều có cấu hình với 8 kênh lọc kích thước 3×3 , tức là, hai mô hình có cấu hình hoàn toàn giống nhau, chỉ khác là huấn luyện trên tập có nhiễu và tập không có nhiễu.

Trong quá trình huấn luyện, mô hình huấn luyện với tập dữ liệu không có nhiễu nhanh chóng đạt được tiệm cận với độ chính xác cao hơn so với khi huấn luyện với tập dữ liệu có nhiễu. Tuy nhiên, khi kiểm tra hiệu năng của hai mô hình này với tập dữ liệu kiểm tra có nhiễu thì mô hình được huấn luyện với tập dữ liệu có nhiễu cho hiệu năng cao hơn so với mô hình khi huấn luyện với tập không có nhiễu. Kết quả thể hiện trong hình 4(a) cho thấy, mô hình huấn luyện có nhiễu đạt độ chính xác trên 99% khi $\text{SNR} \geq -10$ dB, trong khi đó, mô hình huấn luyện không có nhiễu chỉ đạt độ chính xác trên 80% khi $\text{SNR} > 12$ dB. Như vậy, có thể khẳng định dữ liệu càng đa dạng thì mô hình học sâu càng cho hiệu năng tốt hơn.



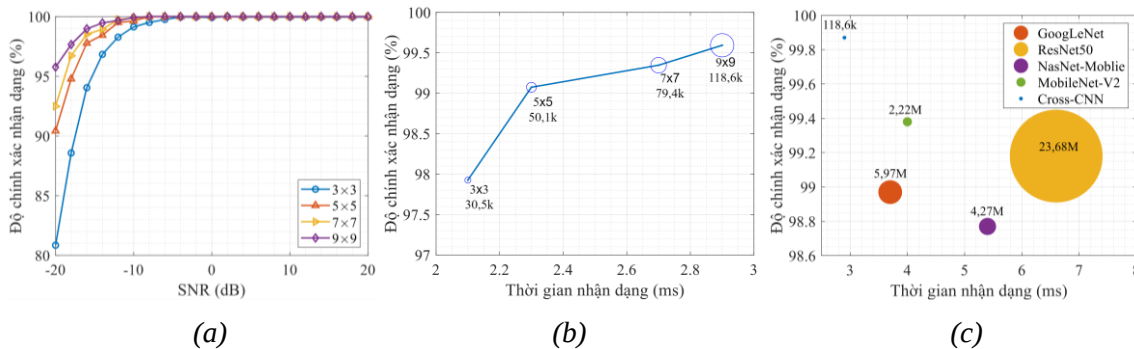
Hình 4. Đánh giá hiệu năng của mô hình Cross-CNN khi: (a) Thay đổi dữ liệu huấn luyện; (b) Thay đổi số lượng kênh lọc (độ chính xác theo SNR); (c) Độ chính xác trung bình theo thời gian nhận dạng và kích thước mạng.

4.2. Đánh giá hiệu năng của Cross-CNN khi thay đổi số lượng kênh lọc trong lớp Conv

Ở phần này, nhóm tác giả đánh giá hiệu năng của Cross-CNN khi thay đổi số lượng các kênh lọc lần lượt là 8, 16, 32 và 64. Các kênh lọc đều được chỉ định với kích thước 3×3 . Mô hình được huấn luyện trên tập dữ liệu có nhiễu và giữ cố định các tham số khác. Kết quả kiểm tra hiệu năng nhận dạng cử chỉ tay được thể hiện như trong hình 4(b) và (c). Trong đó, hình 4(b) thể hiện mối quan hệ giữa độ chính xác nhận dạng với các giá trị SNR, hình 4(c) thể hiện mối liên hệ giữa độ chính xác, thời gian thực thi và kích thước của mô hình (thể hiện bằng độ lớn hình tròn). Hình 4(b) cho thấy, càng tăng số lượng kênh lọc thì tỉ lệ nhận dạng đúng càng tăng. Đặc biệt, sự khác biệt hiệu năng được thể hiện rõ rệt với các giá trị SNR thấp từ -20 dB đến -4 dB. Trong khi đó, với $\text{SNR} > -4$ dB thì sự khác biệt độ chính xác của các mô hình là không nhiều. Mặc dù tăng số kênh lọc có thể cải thiện độ chính xác nhận dạng cử chỉ tay nhưng nó cũng làm tăng kích thước mô hình và thời gian nhận dạng, như thể hiện trong hình 4(c). Thật vậy, khi tăng số kênh lọc từ 8 lên 16 thì độ chính xác nhận dạng trung bình tăng đáng kể từ $\approx 97,9\%$ lên đến $\approx 99\%$, kích thước mạng tăng khoảng 50,3 nghìn trọng số. Trong khi đó, sự chênh lệch độ chính xác trung bình giữa các mô hình với 16, 32 và 64 kênh lọc là không nhiều (tăng khoảng 0,28%), nhưng kích thước mô hình lại tăng đáng kể, lần lượt là 167,8 nghìn và 604,4 nghìn trọng số. Vì vậy, việc lựa chọn mô hình phù hợp để cân bằng hiệu năng giữa các chỉ tiêu về độ chính xác, kích thước mô hình và thời gian thực thi là rất cần thiết.

4.3. Đánh giá hiệu năng của Cross-CNN khi thay đổi kích thước bộ lọc trong lớp Conv

Tiếp theo, mô hình Cross-CNN với 8 kênh lọc trong các lớp Conv được lựa chọn cho phần đánh giá này khi thay đổi kích thước của bộ lọc lần lượt là 3×3 , 5×5 , 7×7 và 9×9 . Đây cũng là những tham số làm ảnh hưởng đến hiệu năng nhận dạng cử chỉ tay của mô hình đề xuất. Chú ý rằng, các tham số huấn luyện và kiểm tra là như nhau cho tất cả các trường hợp. Kết quả đánh giá được chỉ ra trong hình 5. Từ hình 5(a) cho thấy, mặc dù số lượng kênh lọc không đổi (8 kênh) nhưng khi tăng kích thước bộ lọc thì tỉ lệ nhận dạng đúng cử chỉ tay cũng tăng theo. Sự khác biệt độ chính xác của các kích thước thể hiện rõ khi $\text{SNR} < -6 \text{ dB}$, và không có nhiều sự khác biệt khi $\text{SNR} \geq -6 \text{ dB}$. Đặc biệt các bộ lọc 5×5 , 7×7 và 9×9 có độ chênh lệch độ chính xác không đáng kể, và cao hơn đáng kể so với bộ lọc 3×3 . Hình 5(b) cho thấy, việc tăng kích thước bộ lọc cũng làm gia tăng kích thước mô hình và đồng thời cải thiện độ chính xác. Tuy nhiên, khi so sánh với hình 4(c) có thể thấy, tỉ lệ tăng độ chính xác nhận dạng tương đương nhau, nhưng số lượng trọng số của mô hình thay đổi kích thước bộ lọc tăng chậm hơn so với số lượng trọng số của mô hình khi thay đổi số lượng kênh lọc. Cụ thể, mô hình 8 kênh lọc kích thước 9×9 có độ chính xác trung bình 99,59% nhưng chỉ có 118,6 nghìn trọng số, trong khi đó, mô hình 64 kênh lọc kích thước 3×3 chỉ đạt độ chính xác trung bình 99,55% mà số lượng trọng số là 853 nghìn, cao gấp 7,8 lần so với mô hình 8 kênh lọc kích thước 9×9 . Bên cạnh đó, tốc độ xử lý của mô hình 8 kênh lọc kích thước 9×9 (2,9 ms) nhanh hơn mô hình 64 kênh lọc kích thước 3×3 (3,1 ms). Như vậy, có thể thấy rằng, tăng kích thước bộ lọc có thể mang lại hiệu quả cao hơn so với tăng số lượng kênh lọc. Mặc dù vậy, cần phải cân đối giữa hai tham số này của mạng để phù hợp với phần cứng sử dụng và nhu cầu ứng dụng của mô hình.



Hình 5. Đánh giá hiệu năng của mô hình Cross-CNN khi thay đổi kích thước bộ lọc và so sánh với các mô hình khác: (a) Độ chính xác theo SNR; (b) Độ chính xác trung bình theo thời gian nhận dạng và kích thước mạng; (c) So sánh các mô hình.

4.4. So sánh hiệu năng của Cross-CNN với các mô hình hiện có khác

Ở phần cuối cùng, mô hình Cross-CNN 8 kênh lọc kích thước 9×9 được chọn để so sánh với một số mô hình CNN hiện có khác như GoogLeNet [11], ResNet50 [12], NasNet-Mobile [13] và MobileNet-V2 [14]. Để có sự cạnh tranh công bằng, các mô hình được huấn luyện trên cùng một tập dữ liệu Dop-Net có thêm nhiễu. Để làm được điều đó, đầu vào của bốn mô hình nêu trên được điều chỉnh với kích thước 200×200 , và đầu ra được chỉ định thành bốn phân lớp. Các chỉ tiêu so sánh gồm độ chính xác trung bình, kích thước mạng, và thời gian thực thi trung bình cho một lần nhận dạng cử chỉ tay. Kết quả so sánh thể hiện trong hình 5(c) cho thấy, mô hình Cross-CNN có số lượng trọng số ít nhất (118,6 nghìn trọng số), độ chính xác nhận dạng trung bình mà nó đạt được cao nhất (99,59%), và thời gian thực thi cũng nhanh nhất (2,9 ms). Các mô hình còn lại đều có sử dụng Skip-connection, cái mà chỉ sử dụng lại đặc tính một lần, nên trong một số trường hợp, các đặc tính cũ vẫn có thể bị triệt tiêu trong các khối trích chọn tiếp theo, điều này làm cho hiệu quả phân loại chưa tối ưu. Mô hình Cross-CNN mặc dù số lượng trọng số ít hơn

nhưng lại đạt được độ chính xác cao hơn là do nó đã được khéo léo sử dụng các Cross-connection kết hợp với Skip-connection để tăng cường sử dụng các đặc tính cũ nhiều lần và chéo nhau, cái mà vẫn còn hữu ích trong quá trình truyền thẳng của mạng nơ-ron. Hơn nữa, chúng tôi cũng xem xét đến sự phù hợp giữa mạng nơ-ron Cross-CNN với tập dữ liệu Dop-Net, rằng số nhân của dữ liệu này ít (bốn nhân), vì vậy, những mô hình có nhiều trọng số là không quá cần thiết để phân loại cho tập dữ liệu này.

5. KẾT LUẬN

Tóm lại, bài báo đã đề xuất mô hình Cross-CNN nhằm nâng cao hiệu năng nhận dạng cử chỉ tay bằng FMCW dựa vào các dấu hiệu phổ micro-Doppler. Qua khảo sát đánh giá bằng cách thay đổi tập dữ liệu khi huấn luyện, thay đổi các tham số cấu trúc có thể thấy rằng, mô hình Cross-CNN đạt được hiệu năng tốt nhất khi thiết kế các lớp tích chập với 8 kênh lọc kích thước 9×9 . Khi so sánh với một số mạng nơ-ron hiện có khác trên cùng một tập dữ liệu, Cross-CNN đạt được hiệu năng vượt trội so với những mô hình đó. Qua nghiên cứu có thể rút ra một số đánh giá như sau: (1) mô hình CNN cần được huấn luyện với tập dữ liệu phong phú, đa dạng để nâng cao hiệu năng phân loại/nhận dạng khi được ứng dụng thực thi; (2) mô hình CNN càng lớn thì độ chính xác phân loại càng cao, tuy nhiên, nó sẽ tăng chi phí bộ nhớ và thời gian thực thi, vì vậy, cần phải lựa chọn và đánh đổi giữa các chỉ tiêu để đạt được hiệu quả tốt nhất tùy thuộc vào mục đích, nhu cầu sử dụng. Trong tương lai, nhóm nghiên cứu tiếp tục tối ưu hóa mô hình Cross-CNN nhằm nâng cải thiện độ chính xác cho các tập dữ liệu lớn hơn, nhiều cử chỉ tay phức tạp hơn, đồng thời tiến hành đo thực nghiệm kiểm chứng và đưa vào ứng dụng chạy thời gian thực trên thiết bị.

TÀI LIỆU THAM KHẢO

- [1]. X. Zabulis, H. Baltzakis, and A. A. Argyros, "Vision-based hand gesture recognition for human-computer interaction," in *The Universal Access Handbook*. Boca Raton, FL, USA: CRC Press, 2009.
- [2]. S. Ahmed, K. D. Kallu, S. Ahmed, and S. H. Cho, "Hand Gestures Recognition Using Radar Sensors for Human-Computer-Interaction: A Review," *Remote Sensing*, vol. 13, no. 3, p. 527, Feb. 2021.
- [3]. [Online] Available: <https://atap.google.com/soli>.
- [4]. M. Scherer, M. Magno, J. Erb, P. Mayer, M. Eggimann and L. Benini, "TinyRadarNN: Combining Spatial and Temporal Convolutional Neural Networks for Embedded Gesture Recognition With Short Range Radars," in *IEEE Internet of Things Journal*, vol. 8, no. 13, pp. 10336-10346, 1 July 2021.
- [5]. M. G. Amin, Z. Zeng and T. Shan, "Hand Gesture Recognition based on Radar Micro-Doppler Signature Envelopes," 2019 IEEE Radar Conference (RadarConf), 2019, pp. 1-6.
- [6]. M. Ritchie, R. Capraru, and F. Fioranelli, "Dop-NET: a micro-Doppler radar data challenge," *Electronics Letters*, vol. 56, no. 11, pp. 568-570, May 2020.
- [7]. Q. Zhang, "Micro-Doppler Characteristics of Radar Targets", Kidlington, United Kingdom: Butterworth-Heinemann, 2016.
- [8]. A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Dive into Deep Learning". 2019.
- [9]. K. Janocha and W. M. Czarnecki, "On Loss Functions for Deep Neural Networks in Classification," *Schedae Informaticae*, vol. 1/2016, 2017.
- [10]. S. Ruder, "An overview of gradient descent optimization algorithms," arXiv:1609.04747v2, online [Available] <https://arxiv.org/abs/1609.04747>, 2016.
- [11]. C. Szegedy et al., "Going Deeper with Convolutions," arXiv:1409.4842v1 17 Sep 2014. [Online] Available: <https://arxiv.org/pdf/1409.4842.pdf>.
- [12]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.
- [13]. B. Zoph, V. Vasudevan, J. Shlens and Q. V. Le, "Learning Transferable Architectures for Scalable Image Recognition," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8697-8710.

- [14]. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510-4520.

ABSTRACT

HAND GESTURE RECOGNITION USING FMCW RADAR BASED ON CROSS-CONNECTION CONVOLUTIONAL NEURAL NETWORK

This study proposes a cross-connection convolutional neural network, namely Cross-CNN, to recognize hand gestures based on micro-Doppler spectrum data of FMCW (Frequency Modulated Continuous Wave) radar. In addition, different noise levels are added to the dataset for improving the recognition accuracy of the proposed model when predicting gestures in different noise conditions. The experimental results show that the model trained on the dataset with noise gives better recognition accuracy than the model trained on the dataset without noise does. Afterwards, the Cross-CNN model is investigated in changing the structural superparameters for selecting the most suitable parameter set for the proposed problem. Finally, the chosen Cross-CNN model is compared with other existing models in the same dataset and training conditions. As a result, the Cross-CNN network outperforms other models in terms of recognition accuracy, time-consumption and structural capacity thanks to using cross-connections which allow to combine new features with former ones in training process of the network.

Keywords: Convolutional neural network; Hand gesture recognition; FMCW radar micro-Doppler spectrum.

Nhận bài ngày 15 tháng 7 năm 2021

Hoàn thiện ngày 08 tháng 9 năm 2021

Chấp nhận đăng ngày 10 tháng 10 năm 2021

Địa chỉ: ¹Viện Tích hợp hệ thống, Học viện Kỹ thuật quân sự;
²Khoa Thông tin - Ra đa, Học viện Hải quân;
³Khoa Điện tử, Trường Đại học Công nghiệp Hà Nội.
*Email: haidvt@gmail.com.