P-ISSN 1859-3585 | E-ISSN 2615-9619

NHẬN DIỆN KHUÔN MẶT SỬ DỤNG MẠNG NƠRON TÍCH CHẬP XẾP CHỒNG VÀ MÔ HÌNH FACENET

A FACE RECOGNITION SYSTEM USING MULTI-TASK CASCADED CONVOLUTIONAL NETWORKS AND FACENET MODEL

Trần Hồng Việt[1,*], Đỗ Đình Tiến[1], Nguyễn Thị Trà[1], Trần Lâm Quân[2]

TÓM TẮT

Mạng nơ-ron tích chập (CNN) là một trong những mô hình học sâu hiệu quả nhất trong lĩnh vực nhận diện khuôn mặt, các vùng hình ảnh khác nhau luôn được sử dụng đồng thời khi trích xuất các đặc trưng hình ảnh, nhưng trong thực tế, các phần của khuôn mặt đóng những vai trò khác nhau trong việc nhận diện. Trong bài báo này, chúng tôi sử dụng mối tương quan giữa phát hiện và hiệu chỉnh để nâng cao hiệu suất trong một mạng nơ-ron tích chập xếp chồng (MTCNN). Ngoài ra, chúng tôi sử dụng framework FaceNet của Google để tìm hiểu cách ánh xạ từ hình ảnh khuôn mặt đến không gian Euclide, nơi khoảng cách tương ứng trực tiếp với độ đo độ tương tự khuôn mặt để trích xuất hiệu suất của các thuật toán đặc trưng khuôn mặt. Thuật toán gộp trung bình có trọng số được áp dụng cho mạng FaceNet và thuật toán nhận dạng khuôn mặt dựa trên mô hình FaceNet cải tiến được đề xuất. Thực nghiệm và ứng dụng thử nghiệm cho thấy thuật toán nhận dạng khuôn mặt được đề xuất có độ chính xác nhận dạng cao sử dụng phương pháp nhận dạng khuôn mặt dựa trên học sâu.

***Từ khóa**: Nhận diện khuôn mặt, học sâu, FaceNet, mạng nơ-ron tích chập, mạng nơ-ron tích chập xếp chồng.*

ABSTRACT

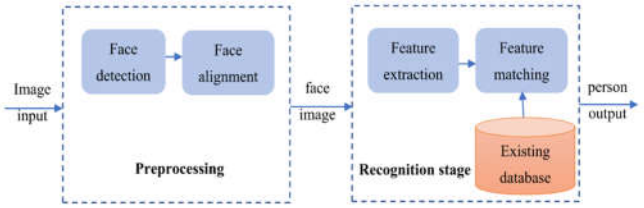
The convolutional neural networks (CNN) is one of the most successful deep learning model in the field of face recognition, the different image regions are always treated equally when extracting image features, but in fact different parts of the face play different roles in face recognition. In this paper, we use the inherent correlation between detection and calibration to enhance their performance in a deep multi-task cascaded convolutional neural network (MTCNN). In addition, we utilize Google's FaceNet framework to learn a mapping from face images to a compact Euclidean space, where distances directly correspond to a measure of face similarity to extract the performance of facial feature algorithms. The weighted average pooling algorithm is applied to the FaceNet network, and a face recognition algorithm based on the improved FaceNet model is proposed. The experiments and apply system show that the proposed face recognition algorithm has high recognition accuracy using face recognition method based on deep learning.

***Keywords**: Face recognition, deep learning, faceNet, convolutional neural networks, multi-task cascaded convolutional neural network.*

[1]Khoa Công nghệ thông tin, Trường Đại học Kinh tế Kỹ thuật Công nghiệp
[2]Trung tâm ứng dụng khoa học hàng không, VietnamAirlines
[*]Email: thviet79@gmail.com
Ngày nhận bài: 12/4/2021
Ngày nhận bài sửa sau phản biện: 20/5/2021
Ngày chấp nhận đăng: 25/6/2021

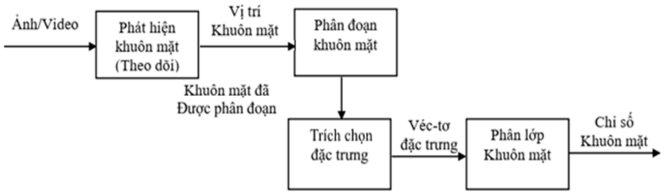
1. GIỚI THIỆU

Mạng nơ-ron tích chập (CNN) [1, 2, 3] là một trong những mô hình học sâu thành công nhất trong lĩnh vực nhận dạng khuôn mặt, các vùng hình ảnh khác nhau luôn được sử dụng đồng thời khi trích xuất các đặc trưng hình ảnh, nhưng trong thực tế, các phần khác nhau của khuôn mặt đóng những vai trò khác nhau trong nhận diện khuôn mặt. Mỗi khuôn mặt của mỗi người có sự độc đáo và nét đặc trưng riêng biệt.



Hình 1. Quá trình nhận dạng khuôn mặt
(Nguồn: https://core.ac.uk/download/pdf/208977767.pdf)

Hình 1 mô tả quá trình nhận diện khuôn mặt. Trường hợp ảnh đầu vào (image input) gồm cả không gian có chứa khuôn mặt người muốn định danh thì ta cần phát hiện vùng ảnh chỉ chứa khuôn mặt của người đó (face detection). Đây cũng là một bài toán được tập trung nghiên cứu [4, 5]. Ảnh khuôn mặt có thể được tiền xử lý (căn chỉnh chẳng hạn - face alignment) nhằm đảm bảo chất lượng cho nhận diện. Khuôn mặt của mỗi người được trích chọn và biểu diễn thông qua một véc-tơ đặc trưng (feature extraction) nhằm mô tả những đặc điểm riêng biệt của khuôn mặt người đó và để so sánh với các khuôn mặt khác. Việc so sánh khuôn mặt đầu vào với cơ sở dữ liệu các khuôn mặt đã được lưu trữ (existing database) trở thành việc tính toán mức độ gần nhau giữa các véc-tơ đặc trưng (feature matching), từ đó tìm ra khuôn mặt giống nhất trong cơ sở dữ liệu.



Hình 2. Các bước chính trong một hệ thống nhận dạng khuôn mặt

Một hệ thống nhận diện mặt người thông thường bao gồm bốn bước xử lý sau:

1. Phát hiện khuôn mặt (Face Detection).

2. Phân đoạn khuôn mặt (Face Alignment hay Segmentation).

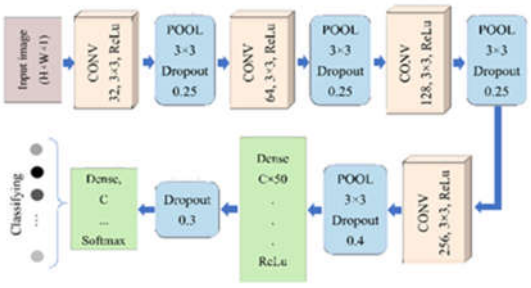
3. Trích chọn đặc trưng (Feature Extraction).

4. Nhận diện (Recognition) hay Phân lớp khuôn mặt (Face Classification).

Bên cạnh những bước chính nêu trên, ta còn có thể áp dụng thêm một số bước khác như tiền xử lý, hậu xử lý nhằm làm tăng độ chính xác cho hệ thống. Sau bước phát hiện khuôn mặt, ta có thể thực hiện bước tiền xử lý (Preprocessing) bao gồm các bước căn chỉnh ảnh (face image alignment) và chuẩn hóa ánh sáng (illumination normalization). Do một vài thông số như: tư thế khuôn mặt, độ sáng, điều kiện ánh sáng,… phát hiện khuôn mặt được đánh giá là bước khó và quan trọng nhất so với các bước còn lại của hệ thống.

Trong nghiên cứu này, chúng tôi không tập trung tìm hiểu bước phát hiện khuôn mặt mà chỉ tập trung chủ yếu vào bước nhận diện khuôn mặt qua việc đề xuất sử dụng một phương pháp nhận diện khuôn mặt với mạng đa tích chập xếp chồng và xây dựng ứng dụng nhận diện khuôn mặt cho các đối tượng sinh viên và các cán bộ giảng viên trên một số dữ liệu thu thập được. Cách tiếp cận của chúng tôi là sử dụng mô hình đạt độ chính xác cao trong nhận diện khuôn mặt dựa trên mạng nơron tích chập với cơ chế học sâu, kiểm tra mô hình trên bộ dữ liệu mẫu và thử nghiệm với bài toán nhận diện khuôn mặt. Bài báo được cấu trúc gồm: Phần 1 giới thiệu về bài toán nhận diện khuôn mặt; Phần 2 trình bày các nghiên cứu liên quan; Phần 3 giới thiệu phương pháp đề xuất và nêu bật một số ưu điểm và hạn chế; Phần 4 thực nghiệm và phân tích kết quả; Phần 5 kết luận và một số định hướng nghiên cứu.

2. CÁC NGHIÊN CỨU LIÊN QUAN



Hình 3. Một minh họa về kiến trúc dạng khối của mô hình CNN (Nguồn: Researchgate.net)

Những năm gần đây, sự phát triển mạnh mẽ của công nghệ học sâu (deep learning) với mạng nơron tích chập (convolutional neural network - CNN) và được ứng dụng thành công trong nhiều bài toán thực tế [3, 5]. CNN là một cấu trúc mạng nơron nhân tạo gồm ba loại lớp nơron (hình 3): lớp nơron tích chập (convolution layer), lớp nơron gộp chung (pooling layer) và lớp nơron kết nối đầy đủ (fully connected layer). Hai lớp nơron đầu (tích chập và gộp chung) thực hiện vai trò trích chọn đặc trưng của ảnh khuôn mặt, trong khi lớp thứ ba (kết nối đầy đủ) thực hiện vai trò ánh xạ các đặc trưng được trích chọn thành đầu ra cuối cùng, tức là định danh của người được nhận diện. Lớp nơron tích chập đóng vai trò quan trọng trong CNN, bao gồm một chồng các phép toán tích chập, là một loại phép tuyến tính chuyên biệt. Lớp nơron gộp chung đóng vai trò làm giảm số chiều của không gian đặc trưng được trích chọn (hay còn gọi là subsampling) nhằm tăng tốc độ xử lý của quá trình nhận diện. Quá trình học mạng nơron là điều chỉnh các tham số học của mạng (trainable parameters) gồm các trọng số liên kết của lớp nơron tích chập và lớp nơron kết nối đầy đủ. Thuật toán học điển hình của mạng nơron dạng này là lan truyền ngược sai số với mục tiêu giảm thiểu sai số kết quả nhận diện của mạng. Ngoài ra, mạng còn có các tham số cần phải thiết lập trước khi áp dụng như kích thước của nhân trong phép tích chập, độ trượt của phép tích chập, hàm kích hoạt, phương pháp tính của lớp nơron gộp chung và các tham số của mạng.

Nhiều nghiên cứu ứng dụng CNN trong nhận diện khuôn mặt với các cải tiến ngày một hiệu quả và chất lượng cao hơn, ứng dụng đa dạng vào các bài toán thực tế. Nghiên cứu [6] phân tích tính hiệu quả của CNN so với các phương pháp nhận diện gồm: phân tích thành phần chính (PCA), mô hình biểu đồ mẫu nhị phân cục bộ (LBPH) và láng giềng gần nhất (KNN). Thử nghiệm trên cơ sở dữ liệu ORL cho thấy LBPH đạt kết quả tốt hơn PCA và KNN, nhưng đối với CNN được đề xuất cho độ chính xác nhận diện tốt nhất (98,3% so với ba phương pháp kia chưa đến 90%). Qua đây phần nào khẳng định phương pháp dựa trên CNN hiệu quả hơn các phương pháp khác.

Nghiên cứu [7] đã phân tích đánh giá với các kiến trúc CNN cải tiến khác nhau cho nhận diện khuôn mặt. Thứ nhất là kiến trúc chứa 22 lớp nơron với 140 triệu tham số học và cần 1,6 tỷ FLOPS (floating-point operations per second) cho mỗi ảnh. Dạng kiến trúc thứ hai dựa trên mô hình mạng Interception của GoogleNet gồm các phiên bản với kích thước đầu vào khác nhau nhằm làm giảm không gian tham số học của mạng. Các kiến trúc này được ứng dụng vào các phạm vi khác nhau, trong khi kiến trúc CNN có kích thước lớn cho kết quả cao và phù hợp với ứng dụng trên các máy tính lớn thì với CNN nhỏ hoặc rất nhỏ sẽ phù hợp với các ứng dụng trên thiết bị di động cầm tay nhưng vẫn đảm bảo kết quả chấp nhận được. Nhằm tăng hiệu quả cao hơn, nghiên cứu [8] đề xuất một kiến trúc CNN với quy mô "rất sâu" gồm 11 khối với 37 lớp nơron, 8 khối đầu đóng vai trò trích chọn đặc trưng và 3 khối sau thực hiện chức năng phân lớp để nhận diện. Kiến trúc CNN này được chạy trên quy mô dữ liệu học mạng rất lớn (LFW và YTF với hàng nghìn định danh và hàng triệu bức ảnh) và cho kết quả (98,95% trên LFW và 97,3% trên YTF) tốt hơn so với các mô hình CNN khác.

Nghiên cứu [4, 14] đã đề xuất một hệ thống mạng nơron tích chập cho nhận diện khuôn mặt với sự cải tiến dựa trên kiến trúc CNN của VGG (Visual Geometry Group -

University of Oxford). Đó là sử dụng mô-đun CReLu (hàm kích hoạt của nơron) thay cho mô-đun hàm kích hoạt (ReLu) thông thường, mô-đun CReLu thực hiện ghép nối một ReLu chỉ chọn phần dương với một ReLu chỉ chọn phần âm của sự kích hoạt. Ở đây chính là điểm gấp đôi mức độ phi tuyến của hàm kích hoạt trong CNN và đã được xác định cho chất lượng kết quả tốt hơn. Dựa trên mô hình đề xuất này, xây dựng một hệ thống nhận diện khuôn mặt theo thời gian thực với một mạng nơron tích chập nhiều lớp ("rất sâu") và phân tích thử nghiệm cho kết quả tốt hơn so với kết quả thu được khi sử dụng mô hình ban đầu.

Nghiên cứu [8] cải tiến chất lượng nhận diện cho mô hình dựa trên CNN bằng cách áp dụng phương pháp học mạng nơron với kỹ thuật "triplet loss". Một số nghiên cứu khác tập trung vào vấn đề nhận diện biểu cảm khuôn mặt với các kỹ thuật được đề xuất. Nghiên cứu [9] đã sử dụng mô hình CNN để thiết kế hệ thống nhận diện 6 loại biểu cảm khuôn mặt khác nhau với việc đưa vào tiền xử lý hình ảnh trước khi nhận diện. Nghiên cứu [10] sử dụng kết hợp mô hình nhị phân cục bộ (LBP) và mô hình CNN để nhận diện biểu cảm khuôn mặt. Theo đó, hình ảnh của khuôn mặt được chuyển thành bản đồ đặc trưng bằng LBP, sau đó bản đồ đặc trưng LBP này được sử dụng làm đầu vào của CNN để huấn luyện mạng và nhận diện.

3. PHƯƠNG PHÁP ĐỀ XUẤT

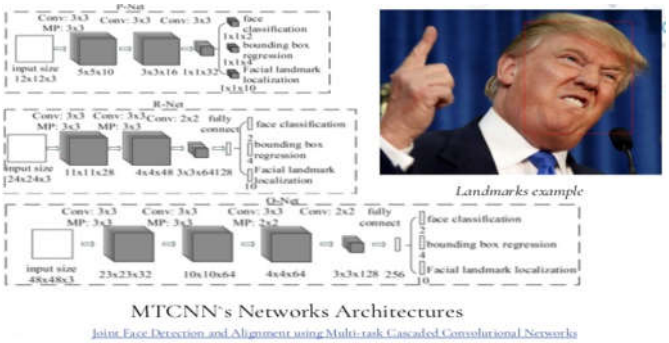
Trong phần này, chúng tôi thiết kế mô hình nhận diện khuôn mặt tập trung chủ yếu vào bước nhận diện khuôn mặt qua việc đề xuất sử dụng phương pháp nhận diện khuôn mặt với mạng đa tích chập xếp chồng sử dụng mô hình Facenet và mô hình VGG16.

3.1. Tiền xử lý ảnh đầu vào

Phần này áp dụng một số phương pháp tiền xử lý trên hình ảnh đầu vào, bao gồm phát hiện và cắt xén để lấy vùng ảnh chứa khuôn mặt, cải thiện chất lượng ảnh. Trong thực tế ứng dụng, ảnh đầu vào thường được trích xuất từ camera nên bao gồm cả không gian nền, do đó, ta phải thực hiện giai đoạn tìm kiếm và phát khuôn mặt (gọi là face detection) nhằm xác định vùng ảnh chứa đúng khuôn mặt cần xử lý và cắt bỏ không gian nền của ảnh. Để thực hiện điều này, các tác giả sử dụng phương pháp phát hiện vùng ảnh có chứa khuôn mặt dựa vào MTCNN.

Khi ảnh khuôn mặt được phát hiện, thực hiện cắt vùng ảnh khuôn mặt đó từ nền, cải thiện chất lượng ảnh khuôn mặt này bằng việc chuyển đổi hình ảnh đầu vào thành hình ảnh đa cấp độ xám và áp dụng phép cân bằng mức xám, co giãn về kích thước đúng với đầu vào của mạng nơron đã thiết kế để thực hiện trích chọn đặc trưng và phân lớp.

MTCNN (Multi-task Cascaded Convolutional Networks) gồm 3 mạng CNN (Convolution, Relu, Max Pooling, Fully Connected Layers) xếp chồng và đồng thời hoạt động khi detect khuôn mặt. Kiến trúc của MTCNN thể hiện như hình 4. Mỗi mạng có cấu trúc khác nhau và đảm nhiệm vai trò khác nhau trong task. MTCNN hoạt động theo ba bước, mỗi bước dùng một mạng noron riêng lần lượt là: mạng đề xuất P-Net (Proposal Network) nhằm dự đoán các vùng trong ảnh ví dụ là vùng chứa khuôn mặt, mạng tinh chế R-Net (Refine Network) sử dụng đầu ra của P-Net để loại bỏ các vùng không phải khuôn mặt và mạng đầu ra (Output Network): sử dụng đầu ra R-Net để đưa ra kết quả cuối cùng với 5 điểm đánh dấu khuôn mặt: 2 điểm mắt, 1 điểm mũi và 2 điểm khóe miệng. Facenet là sản phẩm nghiên cứu của Google giới thiệu năm 2015, với model này đầu vào là ảnh đúng kích thước cho đầu ra là một vector 128 features cho 1 khuôn mặt. Sau đó dùng SVM để phân nhóm các vector đó vào các nhóm để biết véc-tơ đó là mặt của ai.



Hình 4. Kiến trúc dạng khối của mô hình MTCNN (Nguồn: https://tinhte.vn/thread/mi-ai-nhan-dien-khuon-mat-trong-video-bang-mtcnn-va-facenet.3013864/)

3.2. Trích chọn đặc trưng (FaceNet)

Facenet là một hệ thống nhúng cho việc nhận dạng và phân cụm khuôn mặt được đề xuất bởi nhóm tác giả làm việc tại Google[13] dựa trên việc nhúng mỗi ảnh vào không gian Euclide bằng cách sử dụng mạng CNN. Thuật toán nhận diện khuôn mặt trước facenet đều tìm cách biểu diễn khuôn mặt bằng một vector embedding (là vector chuyển dữ liệu chữ viết thô thành dữ liệu số thực) thông qua một layer bottleneck (nút thắt cổ chai có tác dụng giảm chiều dữ liệu. Trong facenet, quá trình encoding của mạng convolutional neural network đã giúp ta mã hóa bức ảnh về 128 chiều. Sau đó những vector này sẽ làm đầu vào cho hàm loss function đánh giá khoảng cách giữa các vector. Để áp dụng triple loss, quá trình học được thực hiện với mỗi bộ ba mẫu học gồm, trong đó là hình ảnh của một người cụ thể (gọi là ảnh neo - anchor), là ảnh khác của cùng một người với ảnh (gọi là ảnh dương - positive) và là hình ảnh của bất kỳ một người khác (gọi là ảnh âm - negative). Mục tiêu ở đây là học mạng nơron (điều chỉnh trọng số mạng) sao cho phản hồi của mạng nơron với cặp mẫu là gần nhau hơn.

Hàm triplet loss luôn lấy 3 bức ảnh làm input và trong mọi trường hợp kì vọng:

$$d(A, P) < d(A, N) \quad (1)$$

Để làm cho khoảng cách giữa vế trái và vế phải lớn hơn, ta sẽ cộng thêm vào vế trái một hệ số α không âm rất nhỏ. Khi đó (1) trở thành:

$$d(A, P) + \alpha \le d(A, N)$$

$$\rightarrow \|f(A) - f(P)\|_2^2 + \alpha \le \|f(A) - f(N)\|_2^2$$

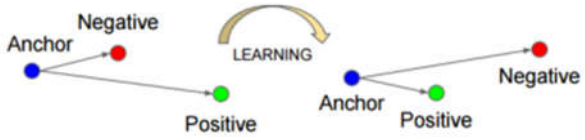
$$\rightarrow \|f(A) - f(P)\|_2^2 - \|f(A) - f(N)\|_2^2 + \alpha \le 0$$

Như vậy hàm loss function sẽ là:

$$\mathcal{L}(A, P, N) = \sum_{i=0}^{n} \|f(A) - f(P)\|_2^2 - \|f(A) - f(N)\|_2^2 + \alpha \quad (2)$$

Trong đó n là số lượng các bộ 3 hình ảnh được đưa vào huấn luyện.

Mục tiêu của hàm loss function là tối thiểu hóa khoảng cách giữa 2 ảnh khi chúng là negative và tối đa hóa khoảng cách khi chúng là positive.



Hình 5. Sai số bộ ba tối thiểu hóa khoảng cách giữa ảnh (Anchor) và ảnh (Positive) và tối đa hóa khoảng cách giữa ảnh (Anchor) và ảnh (Negative)

Do đó để loại bỏ ảnh hưởng của các trường hợp nhận diện đúng Negative và Positive lên hàm loss function. Ta sẽ điều chỉnh giá trị đóng góp của nó vào hàm loss function về 0.

Tức là nếu:

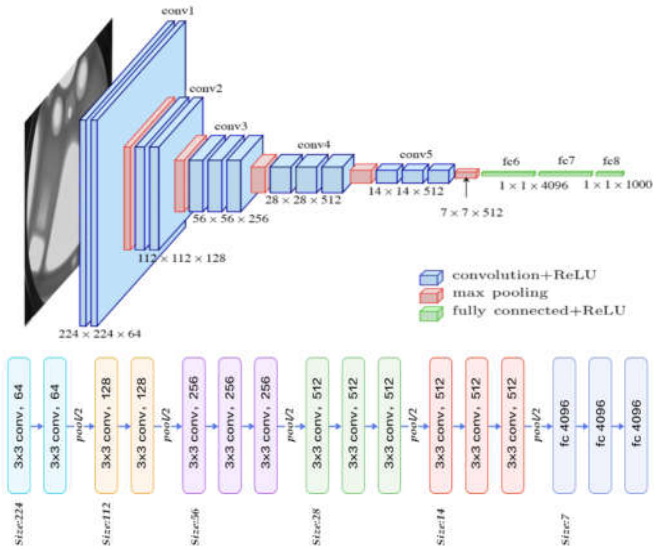
$$\|f(A) - f(P)\|_2^2 - \|f(A) - f(N)\|_2^2 + \alpha \leq 0 \quad (3)$$

sẽ được điều chỉnh về 0. Khi đó hàm loss function trở thành:

$$\mathcal{L}(A, P, N) = \sum_{i=0}^{n} \max(\|f(A_i) - f(P_i)\|_2^2 - \|f(A_i) - f(N_i)\|_2^2 + \alpha, 0) \quad (4)$$

Như vậy khi áp dụng Triple loss vào các mô hình convolutional neural network ta có thể tạo ra các biểu diễn vector tốt nhất cho mỗi một bức ảnh. Những biểu diễn véc-tơ này sẽ phân biệt tốt các ảnh Negative rất giống ảnh Positive. Và đồng thời các bức ảnh thuộc cùng một label sẽ trở nên gần nhau hơn trong không gian chiếu Euclidean.

3.3. Nhận dạng và phân lớp (VGG16)



Hình 6. Kiến trúc của VGG16 (Nguồn: https://nttuan8.com/bai-6-convolutional-neural-network)

VGG16 là mạng CNN được đề xuất bởi K. Simonyan and A. Zisserman, University of Oxford [13]. Mô hình này giành vị trí nhất về phát hiện đối tượng và vị trí haivề phân loại ảnh trong cuộc thi ILSVRC 2014., sau khi train bởi mạng VGG16 đạt độ chính xác cao nằm trong top-5 test trong dữ liệu ImageNet gồm 14 triệu hình ảnh thuộc 1000 lớp khác nhau. Kiến trúc của VGG16 mô tả trong hình 6.

Kiến trúc bao gồm 13 lớp tích chập, 5 lớp max-pooling và 3 lớp kết nối đầy đủ. Số lớp có các tham số có thể điều chỉnh là 16 (13 lớp tích chập và 3 lớp kết nối đầy đủ). Số lượng bộ lọc trong khối đầu tiên là 64, con số này được nhân đôi trong các khối tiếp theo đó cho đến khi đạt 512. Mô hình này được hoàn thiện bởi hai lớp ẩn kết nối đầy đủ và một lớp đầu ra. Hai lớp kết nối đầy đủ có cùng số nơ-ron là 4096. Lớp đầu ra bao gồm 1000 nơ-ron tương ứng với số loại của tập dữ liệu Imagenet.

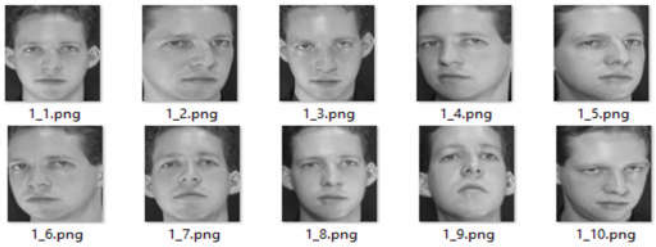
Trong bước nhận dạng hay phân lớp tức là xác định danh tính (identity) hay nhãn của ảnh (label) - đó là ảnh của ai, ở bước nhận dạng/phân lớp, chúng tôi sử dụng phương pháp VGG16. VGG16 sẽ tiến hành phân lớp ảnh trong tập huấn luyện, khi đưa ảnh vào nhận dạng sẽ được so sánh, tìm ra ảnh đó thuộc vào lớp nào.

Thực nghiệm trên các bộ dữ liệu được cộng đồng đánh giá sử dụng (trong phần 4.1). Từ đó xây dựng ứng dụng nhận diện khuôn mặt cho đối tượng sinh viên và các cán bộ giảng viên trên một số dữ liệu thu thập được. Cách tiếp cận của chúng tôi là sử dụng mô hình đạt độ chính xác cao và hiệu quả trong việc nhận diện khuôn mặt. Để kết hợp các phương pháp nhận dạng khuôn mặt nhằm đạt hiệu quả cao, trong phương pháp đề xuất, chúng tôi đưa ra phương pháp phát hiện khuôn mặt (MTCNN), trích chọn đặc trưng (FaceNet), phân lớp và nhận diện khuôn mặt (VGG16).

4. THỰC NGHIỆM VÀ KẾT QUẢ

4.1. Mô tả dữ liệu

Bộ dữ liệu ngoài có bộ dữ liệu mẫu AT&T và Yale được công bố và sử dụng khá rộng rãi cho bài toán nhận diện khuôn mặt [6, 16, 17] gồm có thêm bộ dữ liệu thu thập thêm gồm ảnh của 15 sinh viên, 36 giảng viên trong khoa CNTT và 455 giảng viên Trường Đại học Kinh tế Kỹ thuật - Công nghiệp Hà Nội.



Hình 7. Các ảnh của đối tượng "s1" trong dữ liệu AT&T

Hình 8. Ảnh của 5 người đầu tiên trong dữ liệu AT&T

Bộ dữ liệu khuôn mặt AT&T (hay còn gọi là dữ liệu ORL) được tạo bởi Phòng thí nghiệm AT&T thuộc Đại học

Cambridge, năm 2002. Dữ liệu gồm 400 hình ảnh của 40 người với 10 biểu cảm khuôn mặt khác nhau cho mỗi người, mỗi biểu cảm tương ứng một hình ảnh. Tất cả các hình ảnh được chụp trên nền đồng nhất tối màu với các đối tượng trong tư thế thẳng đứng, chụp từ phía trước và một số trường hợp có hơi nghiêng sang trái hoặc phải, lên trên hoặc xuống dưới. Ảnh khuôn mặt mọi người đều quan sát được, tức không bị che mất những đặc trưng liên quan. Tất cả các hình ảnh đều là ảnh đa mức xám với kích thước 112(cao) × 92(rộng) pixel. Hình 7, 8 minh hoạ các hình ảnh với các trạng thái khác nhau của một người trong tập dữ liệu này.

Bộ dữ liệu khuôn mặt Yale được tạo bởi Trung tâm điều khiển và thị giác máy tính tại Đại học Yale, New Haven. Tập dữ liệu này gồm 165 hình ảnh khuôn mặt chụp từ phía trước và dưới dạng đa cấp xám của 15 người khác nhau. Có 11 hình ảnh cho mỗi người mô tả các biểu cảm khuôn mặt và điều kiện khác nhau như ánh sáng (ánh sáng phía bên phải, ánh sáng ở trung tâm và ánh sáng phía bên trái), trạng thái nét mặt (bình thường, buồn, vui, ngạc nhiên, buồn ngủ và nháy mắt), gồm cả ảnh có đeo kính hoặc không đeo kính. Kích thước của tệp tin hình ảnh tất cả đều là 243(cao) × 320(rộng). Hình 9, 10 minh hoạ các hình ảnh với độ sáng, trạng thái khác nhau của một người trong tập dữ liệu này.

Tập dữ liệu ORL và Yale được chia làm 2 tập chính là tập luyện (processed) và tập thử nghiệm (raw) và được nhận diện thông qua ảnh Các tập dữ liệu còn lại cũng được chia làm 2 tập chính như bộ ORL và Yale nhưng theo tỷ lệ tập thử nghiệm bằng tập luyện và được nhận diện thông qua webcam.

Hình 9. Một phần của tập thử nghiệm trong tập dữ liệu ORL

Hình 10. Các ảnh của đối tượng "subject01" trong dữ liệu Yale (Nguồn:https://colab.research.google.com/drive/10TSK9mJdtpuzArCTsEKh5eIq50m0yq2o)

Bộ dữ liệu thu thập để chạy ứng dụng gồm 15 sinh viên, 36 giảng viên khoa CNTT và 455 cán bộ giảng viên khác của trường Đại học Kinh tế Kỹ thuật Công nghiệp. Dữ liệu gồm ảnh của các đối tượng trong tư thế đứng, chụp từ phía trước và một số trường hợp có hơi nghiêng sang trái hoặc phải, lên trên hoặc xuống dưới, độ tương phản, ánh sáng khác nhau, gồm cả ảnh có đeo kính hoặc không đeo kính, có mũ hoặc không mũ. Ảnh khuôn mặt mọi người đều quan sát được, tức không bị che mất những đặc trưng liên quan.

Hình 11 đưa ra minh họa một phần tập ảnh được huấn luyện (processed) sau khi đã tìm và cắt khuôn mặt.

Hình 11. Minh họa một phần của ảnh huấn luyện

Hình 12 đưa ra một mẫu về tập ảnh thử nghiệm (raw).

Hình 12. Minh họa về ảnh thử nghiệm ban đầu

4.2. Kết quả thử nghiệm

Quá trình thử nghiệm được thực hiện trên hệ thống máy chủ với cấu hình bộ xử lý GPU, giới bộ nhớ 16Gb RAM và 16Gb GPU. Hệ thống được cài đặt môi trường Python, các frameworks và thư viện cơ bản cho học máy (machine learning) như numpy, matplotlib, tensorflow, keras,... thuận lợi cho việc tổ chức dữ liệu phục vụ chạy thử nghiệm và lưu trữ kết quả. Theo đó, chương trình thử nghiệm của chúng tôi được xây dựng trên môi trường Python và sử dụng frameworks của tensorflow với giao diện thư viện keras, đây là thư viện cung cấp các tính năng khá mạnh mẽ cho xử lý ảnh và cho mô hình Neural network.

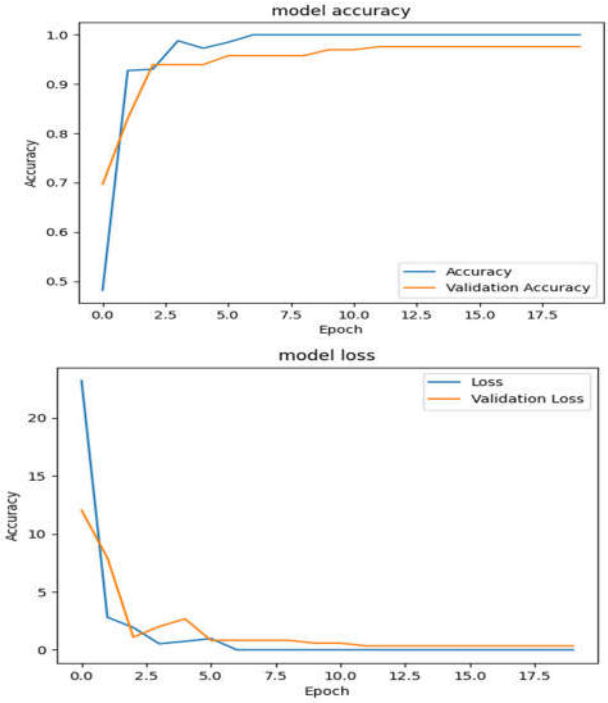
Kết quả trên bộ dữ liệu AT&T và Yale khi so sánh giữa 3 phương pháp, được trình bày ở bảng 1. Để có thể so sánh giữa các mô hình tôi sử dụng phương pháp CNN làm tiêu chuẩn phương pháp này cho tỉ lệ accuracy đạt 95%. Đề xuất ban đầu dùng phương pháp MTCNN, FaceNet và sử dụng SVM để phân lớp tuy nhiên tỷ lệ này đạt 95,1% và tương đương với CNN. Chúng tôi đã đưa ra đề xuất cải tiến dùng phương pháp như trên nhưng thay vì sử dụng SVM chúng tôi sử dụng VGG16 và phương pháp này cho kết quả tốt hơn đạt 97,0% cao hơn so với hai phương pháp trước.

Bảng 1. Kết quả thử nghiệm giữa các mô hình của tập dữ liệu Yale, AT&T

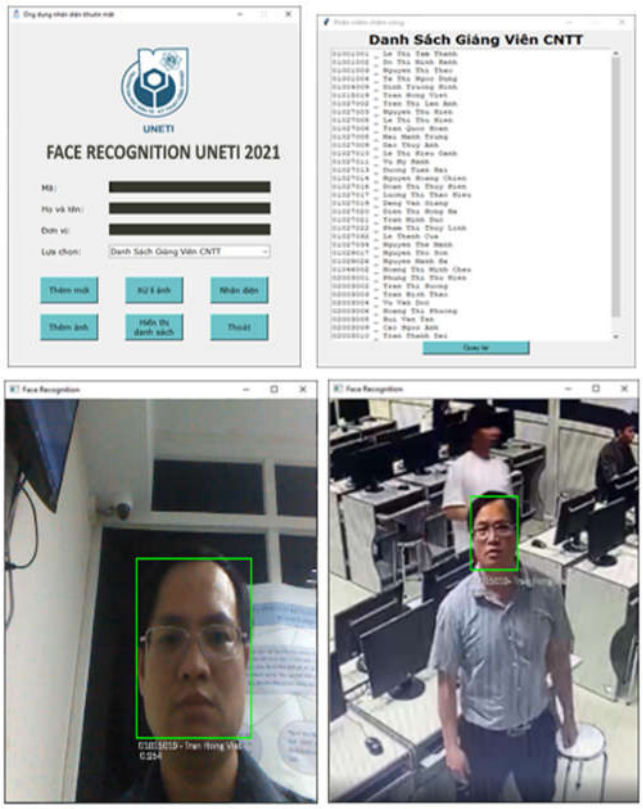
Phương pháp	Tỉ lệ train model
CNN	95,0%
MTCNN, FaceNet sử dụng SVM (*)	95,1%
MTCNN, FaceNet sử dụng VGG16 (**)	97,0%

Kết quả quá trình huấn luyện VGG16 trên hai tập dữ liệu AT&T và Yale được thể hiện trong hình 13. Đồ thị thể hiện tỉ lệ khi huấn luyện dữ liệu sử dụng VGG16 trên tập AT&T và Yale. Kết quả thể hiện được hiệu quả trên số epoch và độ chính xác trong quá trình huấn luyện. Đây là kết quả 10 lần chạy thử nghiệm. Kết quả trên cả hai tập dữ liệu đều cho kết quả độ chính xác phân lớp (accuracy) tốt.

Từ kết quả trên, chúng tôi xây dựng ứng dụng cho việc nhận diện khuôn mặt cho dữ liệu thu thập là các sinh viên và cán bộ giảng viên như mô tả trong mục 4.1. Ứng dụng được thực hiện trên PyCharm Community Edition, sử dụng Python 3.7 virtualenv.



Hình 13. Đồ thị tỉ lệ khi huấn luyện dữ liệu sử dụng VGG16 trên tập AT&T, Yale



Hình 14. Minh họa ứng dụng trên bộ dữ liệu cán bộ giảng viên Khoa Công nghệ thông tin

5. KẾT LUẬN

Trong bài báo này, chúng tôi đã đề xuất mô hình dựa trên mạng nơron tích chập xếp chồng (MTCNN) để nhận diện khuôn mặt con người. Mô hình này bao gồm 3 mạng CNN xếp chồng và đồng thời hoạt động khi detect khuôn mặt. Mỗi mạng có cấu trúc khác nhau và đảm nhiệm vai trò khác nhau trong task. Đầu ra của mô hình là vị trí khuôn mặt và các điểm trên mặt như: mắt, mũi, miệng. Trong mô hình này phát hiện khuôn mặt bằng MTCNN, trích xuất đặc trưng bằng Facenet và dùng SVM để phân lớp và nhận diện mặt. Bên cạnh đó chúng tôi cải tiến MTCNN, trích xuất đặc trưng bằng Facenet kết hợp việc phân lớp của mô hình VGG16. Giải pháp này có nhiều ưu điểm như: nhận diện được mặt ở nhiều góc khác nhau, không cần nhìn thẳng, nhận diện chính xác hơn, trích xuất được nhiều đặc trưng khuôn mặt hơn. Chúng tôi cũng tiến hành thực nghiệm bằng MTCNN sử dụng mô hình Facenet và MTCNN sử dụng mô hình VGG16 để so sánh độ chính xác. Các mô hình đảm bảo độ chính xác cao trong việc nhận diện mặt ở nhiều góc độ và đảm bảo đầu ra các đặc trưng khuôn mặt. Dựa trên các mô hình này chúng tôi tiến hành xây dựng ứng dụng nhận diện khuôn mặt với tập dữ liệu là các cán bộ giảng viên và sinh viên.

Thời gian tới, chúng tôi sẽ nghiên cứu cải thiện hiệu quả phân lớp trong các mô hình hiện nay đạt kết quả cao của AlexNet, VGG, Inception [17, 18, 19] , phân tích điều chỉnh một số lớp CONV bằng lớp nơron. Bên cạnh đó, chúng tôi thiết kế hệ thống thu thập dữ liệu hình ảnh để tạo bộ dữ liệu huấn luyện cho mô hình. Việc xây dựng ứng dụng nhận diện khuôn mặt cho đối tượng sinh viên, giảng viên bước đầu đã đạt được kết quả tốt và có tính thực tiễn cao. Đây là cơ sở đó chúng tôi phát triển tiếp và xây dựng ứng dụng cho bài toán thực tiễn như hệ thống điểm danh khuôn mặt các sinh viên trong lớp học, hệ thống nhận diện cán bộ tại cơ quan, hệ thống chấm công, hệ thống nhận diện cư dân…

TÀI LIỆU THAM KHẢO

[1]. Jing C., Song T., Zhuang L., Liu G., Wang L., Liu K., 2018. *A survey of face recognition technology based on deep convolutional neural networks.* Comput. Appl. Softw. 35(1), 223-231. https://doi.org/10.3969/j.issn.1000-386x.2018.01.039

[2]. Mao Y., 2017. *Research on Face Recognition Algorithms Based on Deep Neural Networks.* Master, Zhejiang University.

[3]. Y. LeCun, Y. Bengio, 1995. *Convolutional networks for images, speech, and time-series.* In M. A. Arbib, editor, The Handbook of Brain Theory and Neural Networks. MIT Press.

[4]. Lionel Landry S. De o, Elie T. Fute, Emmanuel Tonye, 2018. *CNNSFR: A Convolutional Neural Network System for Face Detection and Recognition.* International Journal of Advanced Computer Science and Applications, Vol. 9, No. 12, pp.240-244.

[5]. Mei Wang, Weihong Deng, 2021. *Deep face recognition: A survey.* Neuro computing Volume 429, Pages 215-244. https://doi.org/10.1016/j.neucom.2020.10.081

[6]. Patrik Kamencay, Miroslav Benco, Tomas Mizdos, Roman Radil, 2017. *A New Method for Face Recognition Using Convolutional Neural Network*. Digital Image Processing and Computer Graphics, Vol. 15, No. 4, pp.663-672.

[7]. James Philbin, Florian Schro, Dmitry Kalenichenko, 2015. *FaceNet: A Unified Embedding for Face Recognition and Clustering*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[8]. Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, 2015. *Deep Face Recognition*. University of Oxford.

[9]. Kevin Santoso, Gede Putra Kusuma, Kevin Santoso, Gede Putra Kusuma, 2018. *Face Recognition Using Modified OpenFace*. 3rd International Conference on Computer Science and Computational Intelligence, Procedia Computer Science, No.135, pp.510–517.

[10]. Sonali Sawardekar, Sowmiya Raksha Naik, 2018. *Facial Expression Recognition using Efficient LBP and CNN*. International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395-0056, Volume: 05, Issue: 06, p-ISSN: 2395-0072, pp.2273-2277.

[11]. Andre Teixeira Lopes, Edilson de Aguiar, Thiago Oliveira-Santos, 2015. *A Facial Expression Recognition System Using Convolutional Networks.* 28th SIBGRAPI on Conference Graphics, Patterns and Images.

[12]. Ekberjan Derman and Albert Ali Salah, 2018. *Continuous Real-Time Vehicle Driver Authentication Using Convolutional Neural Network Based Face Recognition.* 13th IEEE International Conference on Automatic Face & Gesture Recognition.

[13]. F. Schroff, D. Kalenichenko, J. Philbin, 2015. *FaceNet: A unified embedding for face recognition and clustering.* 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815-823, doi: 10.1109/CVPR.2015.7298682.

[14]. K. Simonyan, Andrew Zisserman, 2015. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. In Journal CoRR, volume abs/1409.1556.

[15]. Hoda Mohammadzade, Amirhossein Sayyafan, Benyamin Ghojogh, 2018. *Pixel-Level Alignment of Facial Images for High Accuracy Recognition Using Ensemble of Patches.* Journal of the Optical Society of America A 35(7).

[16]. M. A. Abuzneid, A. Mahmood, 2018. *Enhanced Human Face Recognition Using LBPH Descriptor, Multi-KNN, and BPNN*. IEEE Access, Vol. 6, pp.20641-20651.

[17]. Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton, 2012. *Imagenet classification with deep convolutional neural networks.* In F.Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 1097–1105. Curran Associates, Inc.

[18]. Karen Simonyan, Andrew Zisserman, 2014. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv e-prints, page arXiv:1409.1556.

[19]. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, 2015. *Going deeper with convolutions*. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–9.

AUTHORS INFORMATION

Tran Hong Viet[1], Do Dinh Tien[1], Nguyen Thi Tra[1], Tran Lam Quan[2]

[1]Faculty of Information Technology, University of Economics - Technical for Industries

[2]Center of Aviation Science Application, VietnamAirlines