Phương pháp nhận dạng khuôn mặt người từ webcam

Nguyễn Thị Thanh Tân[1]
Khoa Công nghệ thông tin, Trường Đại học Điện lực
Hà Nội, Việt Nam
tanntt@epu.edu.vn

Huỳnh Văn Huy[2],
Trường Đại học Bà Rịa Vũng Tàu
Bà Rịa, Vũng Tàu
huynhvanhuy@gmail.com

Ngô Quốc Tạo[3]
Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam
Hà Nội, Việt Nam
nqtao@ioit.ac.vn

Tóm tắt: Bài báo này đề xuất một mô hình hiệu quả để giải quyết bài toán nhận dạng khuôn mặt trực tiếp từ hệ thống camera/webcam. Trong đó, bài báo tập trung chính vào hai công đoạn: Phát hiện và nhận dạng khuôn mặt từ khung hình webcam. Phương pháp phát hiện khuôn mặt được đề xuất sử dụng các đặc trưng HOG và bộ phân lớp tuyến tính SVM. Mô hình nhận dạng khuôn mặt được đề xuất trên cơ sở mô hình mạng neural học sâu FaceNet để tự động trích chọn đặc trưng khuôn mặt và bộ phân lớp SVM. Hiệu quả của mô hình nhận dạng được kiểm nghiệm trên các tập cơ sở dữ liệu chuẩn, đã được cộng đồng nghiên cứu nhận dạng khuôn mặt người trên thế giới bao gồm cơ sở dữ liệu UOF, FEI, JAFFE và LZW. Các kết quả thực nghiệm cho thấy mô hình đề xuất đạt độ chính xác cao và ổn định trên các tập dữ liệu thử nghiệm được thu thập từ môi trường thực tế.

Từ khóa: khuôn mặt; khung hình (frame); nhận dạng; mạng neural học sâu; tiền xử lý, căn chỉnh khuôn mặt; phát hiện khuôn mặt; trích chọn đặc đặc trưng; phân lớp; tập mẫu

I. ĐẶT VẤN ĐỀ

Trên thế giới, bài toán nhận dạng sinh trắc học nói chung và nhận dạng khuôn mặt nói riêng đã được đầu tư nghiên cứu từ vài chục năm về trước và đã được nhiều kết quả về lý thuyết lẫn ứng dụng thực tiễn. Hiện nay các công nghệ nhận dạng sinh trắc học không chỉ dùng để xác thực nhân thân mà còn được dùng trong rất nhiều bài toán thực tiễn như kiểm soát vào/ra, kiểm soát truy cập mạng, đảm bảo mức độ an ninh cần thiết tại các khu vực quan trọng như nhà ga, sân bay, ngân hàng, hỗ trợ tự động hóa chấm công,v.v.

Tại Việt Nam, công nghệ nhận dạng sinh trắc học cũng đã được ứng dụng rất phổ biến, điển hình là các hệ thống chấm công tự động dựa trên nhận dạng vân tay, mặt người, các hệ thống giám sát an ninh, phát hiện đối tượng, phát hiện đột nhập, phát hiện và cảnh báo sự cố, bất thường. Tuy nhiên, theo tìm hiểu của nhóm tác giả, hầu hết các sản phẩm nhận dạng sinh trắc học hiện có tại Việt Nam đều được nhập khẩu từ nước ngoài.

Trong bài báo này, chúng tôi đề xuất một giải pháp tổng thể để giải quyết bài toán bài toán nhận dạng khuôn mặt người trực tiếp từ các thiết bị camera/webcam, hướng tới mục tiêu ứng dụng xây dựng các hệ thống camera giám sát, kiểm soát vào/ra, phát hiện đột nhập, phát hiện đối tượng lạ mặt, chấm công tự động, v.v. Trong đó, việc cải thiện chất lượng nhận dạng được tập trung ở hai công đoạn chính là phát hiện khuôn mặt trực tiếp từ các khung hình và nhận dạng các khuôn mặt người đã được phát hiện. Mô hình phát hiện khuôn mặt được đề xuất sử dụng các đặc trưng HOG và bộ phân lớp tuyến tính SVM [11]. Mô hình nhận dạng khuôn mặt được đề xuất sử trên cơ sở kết hợp mô hình mạng neural học sâu FaceNet [5] có khả năng tự động trích chọn đặc trưng khuôn mặt người và bộ phân lớp SVM.

Trong phần 2, bài báo đề cập đến các hướng tiếp cận liên quan trong nhận dạng khuôn mặt người. Phần 3 đề xuất một giải pháp tổng thể để nhận dạng khuôn mặt người với độ chính xác cao, đáp ứng được tính thời gian thực, phù hợp với bài toán nhận dạng khuôn mặt trực tiếp từ camera/webcam. Các kết quả thực nghiệm, đánh giá hiệu quả của mô hình được trình bày trong phần 4. Cuối cùng phần kết luận sẽ tổng kết lại những kết quả hiện đã đạt được và một số đề xuất cho hướng phát triển tiếp theo.

II. CÁC HƯỚNG TIẾP CẬN LIÊN QUAN

Nhận dạng mặt khuôn mặt người là quá trình xác định danh tính tự động cho từng đối tượng người trong ảnh/video dựa vào nội dung. Rất nhiều hướng tiếp cận đã được đề xuất để giải quyết bài toán này [7], [9], [15], [8]. Nhìn chung, quy trình giải quyết bài toán thường bao gồm các công đoạn cơ bản như: (i) Thu nhận hình ảnh; (ii) Tiền xử lý, tăng cường chất lượng hình ảnh; (iii) Phát hiện, căn chỉnh, crop ảnh khuôn mặt; (iv) Nhận dạng (trích chọn đặc trưng và phân lớp) khuôn mặt.

Các hướng tiếp cận trước đây chủ yếu dựa trên đặc trưng (feature-based) và luôn cố gắng đưa ra các định nghĩa tường minh để biểu diễn khuôn mặt dựa

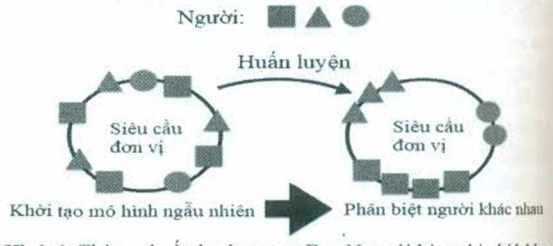
trên các tỷ lệ khoảng cách, diện tích và góc [15]. Một biểu diễn khuôn mặt được định nghĩa tường minh hướng tới mục tiêu xây dựng một không gian đặc trưng trực quan. Tuy nhiên, trong thực tế các biểu diễn được định nghĩa tường minh thường không chính xác. Để khắc phục điều đó, các hướng tiếp cận sau này được đề xuất dựa trên ý tưởng sử dụng các mô hình học máy thống kê có khả năng học để lựa chọn các đặc trưng khuôn mặt từ một tập mẫu cho trước, điển hình như phương pháp PCA (Principal Component Analysis), trong đó mỗi khuôn mặt được biểu diễn dưới dạng tổ hợp các eigenvectors, eigenfaces và fisherfaces [10], [17], phương pháp sử dụng các mô hình mạng neural tích chập CNN (Convolutional Neural Network) [16].

Hiện tại, hiệu quả của các mô hình nhận dạng khuôn mặt đã được cải thiện đáng kể dựa trên việc kết hợp sử dụng các mô hình học sâu để tự động phát hiện các đặc trưng trên khuôn mặt và các kỹ thuật phân lớp thống kê. Trong [20], [21], [22] các tác giả đã đề xuất một mô hình nhận dạng phức tạp, nhiều công đoạn dựa trên việc kết hợp đầu ra của một mạng neural tích chập học sâu D-CNN (Deep Convolutional Neural Network) với PCA để giảm chiều dữ liệu và bộ phân lớp SVM.

Zhenyao và cộng sự [22] xây dựng một mạng neural học sâu để căn chỉnh các khuôn mặt theo hướng nhìn trực diện sau đó huấn luyện một mạng CNN để phân lớp và xác định danh tính cho mỗi khuôn mặt. Y. Taigman và cộng sự [21] đề xuất mô hình DeepFace dựa trên ý tưởng kết hợp nhiều công đoạn (multi-stage): trước tiên sử dụng một mô hình khuôn mặt 3 chiều để chuẩn hóa các ảnh đầu vào (đã được thu thập với các tư thế, góc cạnh khác nhau) về tư thế nhìn thẳng (trực diện), sau đó xây dựng một kiến trúc mạng neural học sâu DNN (Deep Neural Net) với 120 triệu tham số, có khả năng học từ một tập dữ liệu khổng lồ với trên 4.4 triệu khuôn mặt đã được gán nhãn. Trong kiến trúc mạng DNN DeepFace, lớp mạng cuối cùng được loại bỏ và đầu ra của lớp mạng trước đó được sử dụng như một biểu diễn thấp chiều của khuôn mặt. Các kết quả thực nghiệm cho thấy mô hình này đạt độ chính xác trên 97.35% đối với tập dữ liệu LFW [6].

Nhìn chung, các ứng dụng nhận dạng khuôn mặt thường mong muốn tìm được một biểu diễn ít chiều, có khả năng tổng quát hóa tốt đối với những khuôn mặt mới mà mạng chưa được huấn luyện bao giờ. Mục tiêu của DeepFace cũng nhằm giải quyết bài toán đó, tuy nhiên để có được sự biểu diễn này cần phải huấn luyện mạng trên một tập dữ liệu lớn. Đó cũng chính là điểm hạn chế của DeepFace.

Trong [5], Florian Schroff và cộng sự đã đề xuất kiến trúc mạng học sâu FaceNet với hàm chi phí bộ ba (triplet loss function) được định nghĩa trực tiếp trên các biểu diễn. Hình 1 mô tả quá trình huấn luyện mạng FaceNet với hàm chi phí bộ ba để học cách phân cụm các biểu diễn khuôn mặt của cùng một người. Một siêu cầu đơn vị (unit hypersphere) là một siêu cầu có số chiều lớn sao cho khoảng cách từ tất cả các điểm tới tâm của siêu cầu bằng 1.

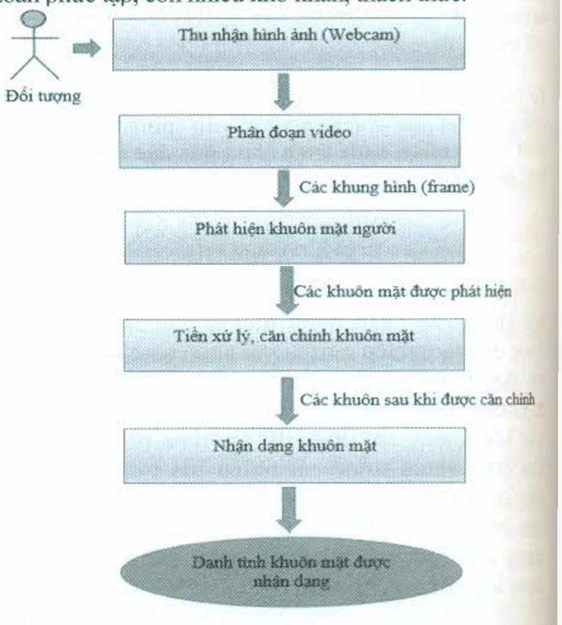


Hình 1. Thủ tục huấn luyện mạng FaceNet với hàm chi phí bộ ba

Các cải tiến quan trọng của FaceNet bao gồm: (i) Đề xuất hàm chi phí bộ ba; (ii) thủ tục lựa chọn các bộ ba trong khi huấn luyện; (iii) cho phép học từ các tập dữ liệu khổng lồ để tìm ra kiến trúc mạng thích hợp.

III. ĐỀ XUẤT MÔ HÌNH NHẬN DẠNG KHUÔN MẶT NGƯỜI TỪ WEBCAM

Thực tế cho thấy, việc nhận dạng đối tượng nói chung và nhận dạng khuôn mặt nói riêng trực tiếp từ hệ thống camera giám sát hoặc webcam hiện vẫn là một bài toán phức tạp, còn nhiều khó khăn, thách thức.



Hình 2. Phương pháp nhận dạng khuôn mặt người từ webcam

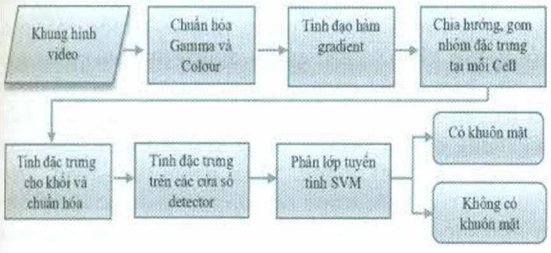
Một trong những thách thức điển hình của bài toán này là hình ảnh khuôn mặt của đối tượng chuyển động và thay đổi liên tục với nhiều tư thế, góc nghiêng/xoay và trạng thái khác nhau. Điều này đòi hỏi các thuật toán nhận dạng phải có khả năng tổng quát hóa, không bị ảnh hưởng nhiều bởi độ

nghiêng/xoay, và dịch chuyển của đối tượng. Ngoài ra, việc nhận dạng trực tiếp từ camera/webcam luôn đòi hỏi phải đáp ứng được tính thời gian thực (real time). Mô hình nhận dạng khuôn mặt người trực tiếp từ webcam hoặc camera được đề xuất cụ thể trên Hình 2.

Từ tín hiệu video đầu vào, bước xử lý đầu tiên sẽ tiến hành phân đoạn video thành các khung hình (frame) riêng biệt. Việc phân đoạn video ở đây được tiến hành theo thời gian (ngưỡng được chọn hiện tại là 24 khung hình trên giây). Mỗi khung hình có thể không chứa, chứa một phần hoặc chứa toàn bộ khuôn mặt. Vì vậy, trong bước xử lý đầu tiên, thuật toán sẽ tiến hành phát hiện (face detection) và xác định vị trí của các khuôn mặt (nếu có) trên ảnh. Các khuôn mặt phát hiện được sau đó sẽ tiếp tục được tiền xử lý nhằm tăng cường chất lượng hình ảnh (loại nhiễu, khử bóng/mờ), chuẩn hóa kích thước và độ phân giải ảnh, căn chỉnh khuôn mặt về hướng trực diện (nhìn thẳng). Các khuôn mặt sau khi đã tiền xử lý sẽ được sử dụng làm đầu vào cho một mô hình mạng neral học sâu (DNN-Deep Neural Network). Mô hình này sẽ tự động học và trích chọn ra các đặc trưng để nhận dạng (phân lớp) khuôn mặt. Bước xử lý cuối của thuật toán sẽ tiến hành phân lớp (nhận diện) các khuôn mặt. Bản chất của việc phân lớp khuôn mặt là tìm kiếm đối tượng người có mẫu khuôn mặt giống với khuôn mặt cần nhận dạng nhất. Để thực hiện được điều này, các mô hình phân lớp cần phải được huấn luyện với một tập mẫu cho trước. Trong đó, mỗi mẫu khuôn mặt được thể hiện bằng tập đặc trưng thu được từ các mô hình phát hiện đặc trưng DNN ở bước trên.

A. Phát hiện khuôn mặt trên khung hình

Như đã đề cập ở trên, bản chất của việc phát hiện khuôn mặt là quá trình tìm kiếm và định vị khuôn mặt trên frame ảnh bất kỳ. Phương pháp phát hiện khuôn mặt ở đây được đề xuất sử dụng các đặc trưng HOG (Histograms of Oriented Gradients) và bộ phân lớp tuyến tính SVM (Support Vector Machines)[11].



Hình 3. Phương pháp phát hiện khuôn mặt

Ý tưởng chính của đặc trưng HOG là hình dạng và trạng thái của vật có thể được đặc trưng bởi sự phân bố về gradient và hướng của cạnh. Đặc trưng này được phát triển dựa trên các đặc trưng SIFT (Scale-Invariant Feature Transform), đặc trưng HOG được tính trên cả một vùng. Do sự biến thiên màu sắc trong các vùng khác nhau nên mỗi vùng sẽ cho ta một vector đặc trưng của nó. Vì vậy để có được đặc trưng của toàn bộ cửa sổ (window) ta phải kết hợp nhiều vùng liên tiếp lại với nhau. Các bước cơ bản trong quy trình phát hiện khuôn mặt người trên các khung hình được mô tả cụ thể trên Hình 3.

Đầu vào của thuật toán là một frame ảnh bất kỳ thu được từ bước phân đoạn video. Bước xử lý đầu tiên sẽ tiến hành chuyển đổi ảnh trong không gian RGB (ảnh màu) sang ảnh đa cấp xám (gray scale), sau đó tiến hành cân bằng histogram trên ảnh gray scale để giảm sự nhạy cảm với nguồn sáng. Bước xử lý tiếp theo sẽ tính sự biến thiên màu sắc tại tất cả các pixel của ảnh gray scale theo chiều X[-1, 0, 1] và theo chiều Y $\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$, thu được 2 ảnh gradient-x (đạo hàm theo trục x) và gradient-y (đạo hàm theo trục y) có kích thước bằng kích thước ảnh gray scale. Hai ảnh thu được cho thấy sự biến thiên màu sắc nói trên. Tiếp theo tiến hành tính góc và hướng biến thiên màu sắc từ 2 ảnh gradient-x và gradient-y.

Việc lưu trữ chính xác từng giá trị góc (orientation) của từng điểm ảnh (x,y) tốn nhiều chi phí và không mang lại nhiều kết quả, do vậy ta sẽ chia không gian góc ra thành các bin. Việc phân chia bin càng nhỏ sẽ càng làm tăng độ chính xác, các kết quả thực nghiệm trong [18] cho thấy kích thước bin khoảng 200 cho kết quả tốt nhất đối với việc phát hiện khuôn mặt người. Do đó, với không gian hướng biến thiên trong miền từ 0^0 - 180^0 sẽ được chia thành 9 bin như sau: [0^0 – 20^0], [21^0 – 40^0], [41^0 - 60^0], [61^0 - 80^0], [81^0 - 100^0], [101^0 - 120^0], [121^0 - 140^0], [141^0 - 160^0], [161^0 - 180^0]. Ứng với mỗi bin trên, tiến hành thống kê biên độ (magnitude) tại từng vị trí. Với mỗi bin, tại vị trí (x,y) nếu góc (orientation) thuộc về bin đó thì giá trị của bin đó tại vị trí (x,y) bằng giá trị biên độ, ngược lại giá trị bin tại vị trí (x,y) bằng 0. Bước tiếp theo tiến hành tính toán vector đặc trưng cho từng cell (mỗi cell thường được chọn với kích thước 8×8 pixel). Vector đặc trưng của mỗi cell sẽ gồm 9 thành phần tương ứng với 9 bin và giá trị tại thành phần i bằng tổng giá trị của các điểm trong bin i mà có tọa độ nằm trong cell đó. Tiếp theo, tính toán vector đặc trưng cho từng khối (block), mỗi khối thường được chọn với kích thước 2×2 cells (16×16 pixel). Vector đặc trưng của khối được tính bằng cách ghép vector đặc trưng của từng cell trong block lại với nhau. Số thành phần của vector đặc trưng tại mỗi khối được tính theo công thức:

$$Size_{feature/block} = n_{cell} \times Size_{feature/cell}$$

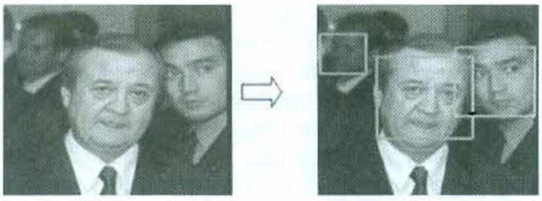
Trong đó: $Size_{feature/block}$ là đặc trưng trong block, n_{cell} là số cell trong một block, $Size_{feature/cell}$ là số feature trong một cell.

Với giả thiết mỗi cell có kích thước 8×8 pixels, mỗi block có kích thước 2×2 cells (16×16 pixels), không gian hướng biến thiên xét trong miền miền từ 0^0 - 180^0 và được chia thành 9 bin thì số đặc trưng trong mỗi khối sẽ được tính bằng 4×9 = 36 thành phần. Từ đó, tiến hành tính toán vector đặc trưng các các cửa sổ trên toàn bộ ảnh đầu vào. Trong đó, một cửa sổ (Window) được tạo bởi các khối xếp gối nhau – overlapping. Đặc trưng của một cửa sổ sẽ được tính bằng cách ghép các vector đặc trưng của từng block tạo lên cửa sổ đó. Số thành phần đặc trưng của mỗi cửa sổ được xác định như sau:

$$n_{block/window} = \left(\frac{W_{window} - W_{block} \times W_{cell}}{W_{cell}} + 1 \right) \times \left(\frac{H_{window} - H_{block} \times H_{cell}}{H_{cell}} + 1 \right)$$

$$Size_{feature/window} = n_{block/window} \times Size_{feature/block}$$

Trong đó: W_{window}, W_{Block}, W_{cell} lần lượt là chiều rộng của window, block và cell (tính theo đơn vị pixel); H_{window}, H_{Block}, H_{cell} lần lượt là chiều cao của window, block và cell (tính theo đơn vị pixel); $n_{Block/Window}$ là số block trong một cửa sổ, $Size_{Feature/Window}$ là số đặc trưng trong một cửa sổ.



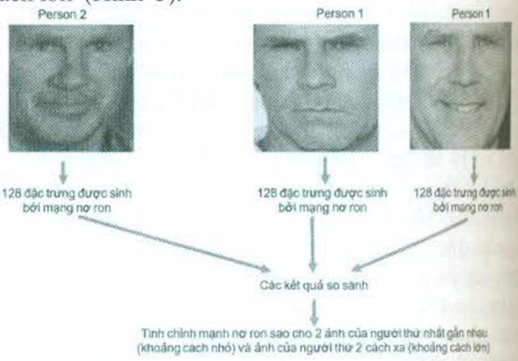
(a) Khung hình đầu vào (b) Các khuôn mặt được phát hiện

Hình 4. Kết quả phát hiện khuôn mặt

Ở bước xử lý cuối cùng, toàn bộ vector đặc trưng thu được trên mỗi cửa sổ sẽ được sử dụng làm đầu vào của bộ phân lớp tuyến tính SVM[12]. Bộ phân lớp có nhiệm vụ xác định lớp mẫu (có chứa khuôn mặt hay không chứa khuôn mặt) đối với mỗi ảnh đầu vào dựa trên các tri thức mà thuật toán đã được huấn luyện. Hình 4-b thể hiện kết quả của thuật toán phát hiện khuôn mặt người trên một ảnh đầu vào cụ thể (Hình 4-a).

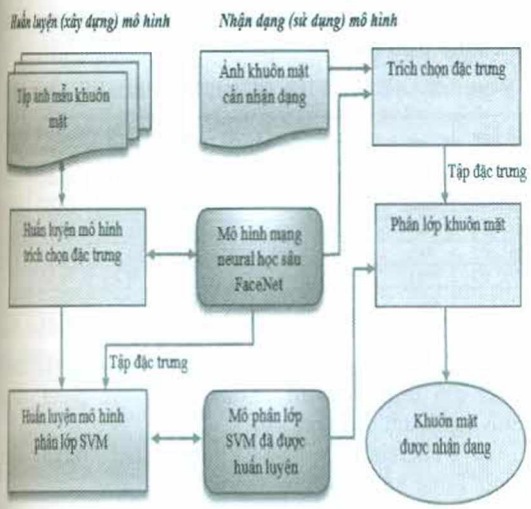
B. Nhận dạng khuôn mặt người

Công đoạn nhận dạng thường gồm 2 bước xử lý chính là trích chọn đặc trưng và phân lớp khuôn mặt. Phương pháp trích chọn đặc trưng ở đây được đề xuất sử dụng các lớp mạng neural học sâu FaceNet, được Florian Schroff và cộng sự đã đề xuất năm 2015 [5]. Đây là mô hình có khả năng học từ một tập mẫu cho trước nhằm tự động phát hiện các đặc trưng quan trọng nhất để nhận dạng đối tượng. Ý tưởng chính của hướng tiếp cận này dựa trên việc học một không gian Euclidean nhúng trong mỗi ảnh sử dụng một cấu hình mạng neural tích chập học sâu (deep convolutional network). Mạng được huấn luyện sao cho khoảng cách L2 bình phương trong không gian nhúng là tương ứng trực tiếp với độ tương tự của khuôn mặt. Cụ thể là các khuôn mặt của cùng một người sẽ có khoảng cách nhỏ và các khuôn mặt của các người khác nhau sẽ có khoảng cách lớn (Hình 5).



Hình 5. Huấn luyện mạng neural, tự động trích rút đặc trưng

Mạng được huấn một cách trực tiếp để đầu ra của nó trở thành một vector đặc trưng 128 chiều sử dụng hàm chi phí bộ ba (tripletbased loss function). Một bộ ba (triplet) được định nghĩa bao gồm hai khuôn mặt của cùng một người - positive và một khuôn mặt của người khác – negative. Mục tiêu của hàm chi phí là phân tách cặp khuôn mặt positive ra khỏi khuôn mặt negative sử dụng một lề khoảng cách - distance margin. Từ các độ đo thu được, thuật toán sẽ ước lượng giá trị của hàm chi phí dựa trên việc so sánh khoảng cách giữa 2 tập đặc trưng, được sinh ra từ 2 ảnh khuôn mặt khác nhau của cùng một người (được gọi là người thứ nhất) và tập đặc trưng thứ 3 được sinh ra từ ảnh khuôn mặt của một người khác (được gọi là người thứ hai). Các giá trị ước lượng của hàm chi phí sau khi tính sẽ được lan truyền ngược từ lớp cuối cùng đến lớp đầu tiên của mạng để tinh chỉnh trọng số (cập nhật lại trọng số) trên các lớp mạng. Quá trình tính toán, ước lượng và cập nhật trọng số của mạng được lặp đi lặp lại liên tục cho đến khi giá trị của hàm chi phí thỏa mãn điều kiện đã cho. Lặp lại các bước trên đối với toàn bộ tập dữ liệu huấn luyện cho đến khi thuật toán huấn luyện mạng hội tụ. Mô hình nhận dạng khuôn mặt được mô tả cụ thể trên Hình 6.



Hình 6. Nhận dạng khuôn mặt người

Kết quả thực nghiệm cho thấy việc sử dụng các lớp mạng học sâu để trích chọn đặc trưng cho độ chính xác cao. Do thuật toán được huấn luyện với tập dữ liệu lớn, đa dạng nên các đặc trưng phát hiện được thường ít bị ảnh hưởng bởi nhiễu và các tính chất nghiêng, xoay của ảnh. Tuy nhiên, do mạng được kiến trúc nhiều lớp và số liên kết giữa các lớp mạng rất lớn nên việc tính toán trên mạng thường mất nhiều thời gian. Điều này dẫn tới tốc độ tổng thể của thuật toán nhận dạng sẽ bị ảnh hưởng. Vì vậy, để đảm bảo thuật toán có thể đáp ứng tính thời gian thực (real-time) trong quá trình nhận dạng, chúng tôi đã tận dụng khả năng tính toán GPU (Graphic Proccessing Unit), cho phép việc tính toán trên các lớp mạng thực hiện theo cơ chế song song.

IV. ĐÁNH GIÁ THỰC NGHIỆM

❖ Môi trường thực nghiệm

Chương trình thực nghiệm được cài đặt trong môi trường python, sử dụng các thư viện NumPy [24] cho việc biểu diễn, lưu trữ và thao tác dữ liệu, thư viện opencv [23] để thực hiện các thai tác xử lý ảnh cơ bản, thư viện Scikit-Learn [25] cho việc thử nghiệm các mô hình học máy (mạng neural, mô hình svm, v.v.). Chương trình được thử nghiệm trên hệ điều hành Windows 10, máy PC tốc độ 2.4GHz, bộ nhớ 6GB.

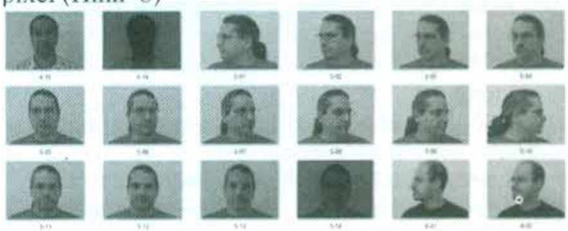
❖ Dữ liệu thử nghiệm

Hiệu quả của mô hình nhận dạng được đánh giá trên các bộ cơ sở dữ liệu chuẩn (chứa các khung hình được thu nhận từ các thiết bị camera, webcam khác nhau), được công bố dùng chung cho các nhóm nghiên cứu trên thế giới, được cung cấp tại [26]. Đây là các CSDL dùng chung cho các nhóm nghiên cứu. Thông tin của CSDL mẫu được mô tả cụ thể như sau:

- **Cơ sở dữ liệu UOF**: Được cung cấp bởi trường đại học Essex của Anh (University of Essex, UK), bao gồm 4 tập dữ liệu: faces94, faces95, faces96 và grimace. Ảnh trong cơ sở dữ liệu là ảnh màu 24 bit định dạng dạng JPEG. Tập dữ liệu chứa một tập hợp các hình ảnh khuôn mặt gồm 395 cá nhân (cả nam và nữ) với 20 ảnh cho mỗi cá nhân, tổng cộng có 7900 hình ảnh. Tất cả khuôn mặt chủ yếu được thực hiện bởi các sinh viên đại học năm đầu tiên có độ tuổi từ 18 đến 20 và một số người lớn tuổi, một số cá nhân đeo kính và có râu, thuộc nhiều chủng tộc khác nhau (Hình 7).

Hình 7. Cơ sở dữ liệu UOF

- **Cơ sở dữ liệu FEI**: Bao gồm các tập dữ liệu: Fei_P1, Fei_P2 và Fei_P3, với các ảnh khuôn mặt được chụp từ tháng 6 năm 2005 đến tháng 3 năm 2006 tại Phòng thí nghiệm Trí tuệ nhân tạo FEI ở Paulo, Brazil. Bao gồm 200 cá nhân (100 nam, 100 nữ), với 14 ảnh cho mỗi cá nhân, tổng cộng 2800 hình ảnh. Tất cả khuôn mặt chủ yếu được thực hiện bởi các sinh viên và nhân viên của FEI, có độ tuổi từ 19 đến 40, với ngoại hình, kiểu tóc và đồ trang điểm khác biệt, đều được chụp trên nền ảnh màu trắng, ở vị trí đứng thẳng đứng và quay vòng lần lượt tới 1800. Kích thước của mỗi ảnh là 640x480 pixel (Hình 8)



Hình 8. Cơ sở dữ liệu mẫu FEI

- **Cơ sở dữ liệu JAFFE**: Chứa các khuôn mặt nữ Nhật Bản, được chụp tại khoa tâm lý học của Đại học Kyushu, Nhật Bản, bao gồm 213 hình ảnh của 7 biểu hiện khuôn mặt (6 biểu hiện cảm xúc cơ bản trên khuôn mặt + 1 trung tính), được chụp bởi 10 người phụ nữ Nhật Bản (Hình 9).

Hình 9. Cơ sở dữ liệu JAFFE

• **Cơ sở dữ liệu LFW**: Bao gồm những khuôn mặt được gắn nhãn trong tự nhiên. Bộ dữ liệu gồm 13233 hình ảnh khuôn mặt của 5749 người được thu thập từ web. Mỗi khuôn mặt được gắn nhãn với tên của người đó, trong đó 1680 người có từ 2 hình ảnh khác biệt trở lên (Hình 10).

Hình 10: Cơ sở dữ liệu LZW

❖ Kết quả thực nghiệm

Quá trình đánh giá thực nghiệm được chia thành 2 công đoạn: Đánh giá hiệu quả của mô hình phát hiện khuôn mặt người trên khung hình webcam và đánh giá độ chính xác nhận dạng. Hiệu quả của mô hình phát hiện khuôn mặt được đánh giá dựa trên các độ đo được định nghĩa cụ thể trong phần sau đây:

▪ Độ chính xác phát hiện khuôn mặt **DP** (Detection Precision):

DP = Số vùng khuôn mặt phát hiện đúng trên tổng số khuôn mặt cần phát hiện

▪ Khả năng tìm hết **DR** (Detection Recall):

DR = Số vùng khuôn mặt phát hiện đúng /(Số vùng khuôn mặt phát hiện đúng + Số vùng không được phát hiện)

▪ Độ trung bình điều hòa **DM** (Dectection F-Measure):

DM = (2×FDP*FDR)/ (FDP+FDR)

Bên cạnh đó, để các kết quả thực nghiệm chính xác và trực quan, trong qúa trình thử nghiệm, chúng tôi đã so sánh hiệu quả của mô hình phát hiện khuôn mặt đề xuất với mô hình phát hiện khuôn mặt sử dụng đặc trưng Haar-Like và bộ phân lớp AdaBoost (được quy ước gọi tên là phương pháp Haar-Like AdaBoost) [19]. Các kết quả thực nghiệm được mô tả cụ thể trên **Bảng 1**.

Bảng 1. ĐÁNH GIÁ HIỆU QUẢ PHÁT HIỆN KHUÔN MẶT

dữ liệu thử nghiệm	Số mẫu	Phương pháp đề xuất			Phương pháp Haar-Like AdaBoost		
		DP	*DR*	*DM*	*DP*	*DR*	*DM*
Faces96	3040	98.42	99.2	98.81	93.75	94.50	94.12
FEI_P1	700	98.43	98.43	98.43	80.71	80.71	80.71
FEI_P2	700	99.14	99.14	99.14	83	83.00	83.00
FEI_P3	700	97.43	97.43	97.43	79.43	79.43	79.43
JAFFE	213	100	100	100	100	100	100
LFW	13233	99.74	99.74	99.74	93.27	93.27	93.27

Hiệu quả của mô hình nhận dạng tổng thể được đánh giá dựa trên độ chính xác nhận dạng, được định nghĩa cụ thể như sau:

R_Precision = Số khuôn mặt nhận dạng đúng/Tổng số khuôn mặt cần nhận dạng.

Quá trình đánh giá thực nghiệm được thực hiện lần lượt trên từng tập dữ liệu. Mỗi tập dữ liệu được chia ngẫu nhiên thành 2 tập training và testing theo tỷ lệ 90/10 (90% số mẫu để huấn luyện mô hình và 10% số mẫu còn lại để kiểm thử).

Việc huấn luyện mô hình gồm 2 công đoạn: Huấn luyện bộ trích chọn đặc trưng (mô hình mạng neural học sâu FaceNet) và huấn luyện bộ phân lớp SVM (xem Hình 6). Quy trình huấn luyện được tiến hành cụ thể như sau: Từ tập mẫu huấn luyện đầu vào, trước tiên bộ phát hiện khuôn mặt sẽ tiến hành tìm kiếm, định vị và crop vùng ảnh khuôn mặt trên mỗi khung hình. Toàn bộ tập ảnh khuôn mặt crop sau đó sẽ được sử dụng làm đầu vào để huấn luyện mô hình trích chọn đặc trưng. Tập đặc trưng đầu ra của mô hình trích chọn đặc trưng sẽ được sử dụng làm đầu vào để huấn luyện mô hình phân lớp SVM.

Các kết quả thực nghiệm được mô tả cụ thể trên **Bảng 2**. Trong đó, hiệu quả của mô hình đề xuất được so sánh với phương pháp phân lớp sử dụng đặc trưng PCA và bộ phân lớp Eigenface (được quy ước gọi tên là phương pháp PCA- Eigenface).

Bảng 2. ĐÁNH GIÁ ĐỘ CHÍNH XÁC NHẬN DẠNG

Tập dữ liệu thử nghiệm	Số khuôn mặt cần nhận dạng	R_Precision (%)	
		PP đề xuất	*PCA- Eigenface*
Faces96	3040	98.02	83.23
FEI_P1	700	98.16	82.12
FEI_P2	700	98.74	83.62
FEI_P3	700	97.55	75.43
JAFFE	213	99.02	95.17
LFW	13233	95.26	78.13

Từ các kết quả thực nghiệm cho thấy phương pháp đề xuất đạt được độ chính xác cao (trên 95%) trên tất cả các tập dữ liệu thử nghiệm. Trong khi đó độ chính xác của Phương pháp PCA- Eigenface bị ảnh hưởng nhiều bởi độ sáng và độ dịch chuyển của ảnh đầu vào.

V. KẾT LUẬN

Bài báo này đề xuất một mô hình tổng thể cho việc nhận khuôn mặt người từ webcam. Trong đó tập trung chính vào công đoạn phát hiện và nhận dạng khuôn mặt. Hiệu quả của mô hình đã được đánh giá trên các tập dữ liệu chuẩn, dùng chung cho cộng đồng nghiên cứu về nhận dạng khuôn mặt người trên thế giới bao gồm cơ sở dữ liệu UOF, FEI, JAFFE và LZW. Quá trình đánh giá thực nghiệm được chia thành 2 bước, trong đó hiệu quả của phương pháp phát hiện khuôn mặt được đánh giá dựa trên 3 độ đo là độ chính xác (Precision), khả năng tìm hết (recall) và độ đo F-measure, hiệu quả của của mô hình nhận dạng khuôn mặt được đánh giá dựa trên độ chính xác nhận dạng. Các kết quả thực nghiệm cho thấy mô hình đề xuất đạt được độ chính xác cao và ổn định trong môi trường thực tế, có thể ứng dụng mô hình để giải quyết các bài toán ứng dụng điển hình như hệ thống camera giám sát cho phép phát hiện, nhận dạng và cảnh báo các đối tượng lạ mặt đột nhập tại các khu vực an ninh, nhà ga, sân bay, các cơ quan chính phủ, tòa nhà, chung cư, tra cứu thông tin tội phạm, chấm công tự động tại các khu công nghiệp, nhà máy, công trường, cải thiện chất lượng của các thuật toán giao tiếp người-máy, v.v.

Cám ơn đề tài "nghiên cứu phương pháp tra cứu ảnh dựa vào đa truy vấn" (mã số PTNTĐ17.04) đã hỗ trợ.

TÀI LIỆU THAM KHẢO

[1] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–9, 2015.

[2] Davis E King. Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research, 10:1755–1758, 2009.

[3] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. arXiv preprint arXiv:1411.7923, 2014.

[4] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. The Journal of Machine Learning Research,12:2825–2830, 2011.

[5] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 815–823, 2015.

[6] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[7] Hiyam Hatem, Zou Beiji,Raed Majeed, "A Survey of Feature Base Methods for Human Face Detection", International Journal of Control and Automation Vol.8, No.5 (2015), pp.61-78.

[8] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. IEEE International Conference on Image Processing (ICIP), 265(265):530, 2014.

[9] Hwai-Jung Hsu and Kuan-Ta Chen. Face recognition on drones: Issues and limitations. In Proceedings of the First Workshop on Micro Aerial Vehicle Networks, Systems, and Applications for Civilian Use, DroNet '15, pages 39–44, New York, NY, USA, 2015. ACM.

[10] Lawrence Sirovich and Michael Kirby. Low-dimensional procedure for the characterization of human faces. JOSA A, 4(3):519–524, 1987.

[11] N. Dalal, B. Triggs, Histograms of Oriented Gradients for Human Detection. IEEEcComputer Society Conference on Computer Vision and Pattern Recognition, 2005.

[12] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In Computer Vision, 2009 IEEE 12th International Conference on, pages 365–372. IEEE, 2009.

[13] Neeraj Singla, IISugandha Sharma, "Advanced Survey on Face Detection Techniques in Image Processing", International Journal of Advanced Research in Computer Science Technology (IJARCST 2014), vol. 2 Issue 1 Jan-March 2014.

[14] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. Proceedings of the British Machine Vision, 1(3):6, 2015.

[15] Rabia Jafri and Hamid R Arabnia. A survey of face recognition techniques. JIPS, 5(2):41–68, 2009.

[16] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. Neural Networks, IEEE Transactions on, 8(1):98–113, 1997.

[17] Turk, M. and Pentland , A. 1991. Eigenfaces for recognition. J. Cogn. Neurosci. 3, 72–86.

[18] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1867–1874, 2014.

[19] Viola, P. and Jones, M. 2001. Rapid object detection using a boosted cascade of simple features. In Proceedings, IEEE Conference on Computer Vision and Pattern Recognition.

[20] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. CoRR, abs/1412.1265, 2014. 1, 2, 5, 8.

[21] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In IEEE Conf. on CVPR, 2014. 1, 2, 5, 7, 8, 9.

[22] Z. Zhu, P. Luo, X. Wang, and X. Tang. Recover canonicalview faces in the wild with deep neural networks. CoRR, abs/1404.3543, 2014. 2

[23] http://opencv.org/

[24] http://www.numpy.org/

[25] http://scikit-learn.org/stable/

[26] http://www.face-rec.org/databases/