DOI: https://doi.org/10.62985/j.huit_ojs.vol26.no1E.355

NGHIÊN CỨU VÀ PHÁT TRIỂN MÔ HÌNH NHẬN DIỆN KHUÔN MẶT ỨNG DỤNG ĐIỂM DANH CHO SINH VIÊN

Bùi Công Danh, Phạm Nguyễn Huy Phương*

Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: *phuongpnh@huit.edu.vn*

Ngày nhận bài: 03/12/2025; Ngày nhận bài sửa: 23/12/2025; Ngày chấp nhận đăng: 20/3/2026

TÓM TẮT

Nhận diện khuôn mặt là một trong những hướng nghiên cứu quan trọng của thị giác máy tính, nhằm tự động xác định hoặc xác minh danh tính cá nhân dựa trên các đặc trưng sinh trắc học từ hình ảnh hoặc video khuôn mặt. Bài báo trình bày một hệ thống nhận diện khuôn mặt ứng dụng cho việc điểm danh sinh viên trong lớp học. Trong bài báo này chúng tôi đề xuất một hệ thống MFSN với ba giai đoạn chính bao gồm: phát hiện MTCNN, trích xuất đặc trưng bằng FaceNet và xác minh danh tính theo cơ chế học một lần (one-shot learning) nhờ Siamese Network, kết quả thực nghiệm cho thấy mô hình đề xuất của chúng tôi đạt được độ chính xác 98,3% trên VGGFace2 và đạt hiệu suất cao trên tập dữ liệu thực tế về sinh viên HUIT với chỉ một hình ảnh đã đăng ký cho mỗi sinh viên. Hệ thống được triển khai hoàn toàn trên máy tính cá nhân cấu hình phổ thông, hoạt động độc lập trong môi trường mạng nội bộ (LAN) mà không yêu cầu kết nối Internet hay dịch vụ đám mây. Hệ thống điểm danh theo thời gian thực cho một lớp học 50 sinh viên trong thời gian trung bình 8 giây.

Từ khóa: Nhận dạng khuôn mặt, One-shot learning, Siamese network, điểm danh sinh viên.

1. GIỚI THIỆU

Nhận diện sinh trắc học là một lĩnh vực nghiên cứu chính và rất quan trọng cho hệ thống bảo mật. Các ứng dụng thực tiễn của công nghệ sinh trắc học khuôn mặt bao gồm: kiểm soát ra vào, xác thực giao dịch tài chính, hệ thống chấm công tự động, giám sát an ninh đô thị và hỗ trợ nhận dạng đối tượng trong an ninh quốc gia. Hệ thống sinh trắc học giúp người sử dụng công nghệ bảo mật an toàn và tiết kiệm thời gian hơn vì sinh trắc học bảo vệ an toàn hơn mật khẩu và thẻ định danh. Công nghệ nhận diện khuôn mặt sinh trắc học cũng rất tiện lợi, hoàn toàn không xâm lấn và dễ dàng thiết lập trên các thiết bị công nghệ hiện có. Ngoài việc cải thiện độ chính xác của việc xác thực, nghiên cứu này cũng tập trung vào hiệu quả của các hệ thống quy mô nhỏ trong việc đảm bảo triển khai trong các ứng dụng thực tế như kiểm soát truy cập cá nhân, an ninh thông tin và các môi trường có rủi ro cao trong việc xác minh danh tính. Bài báo này nhằm nghiên cứu và phát triển một hệ thống điểm danh sinh viên tự động sử dụng công nghệ nhận diện khuôn mặt thời gian thực trong một môi trường LAN hạn chế. Việc triển khai hệ thống dựa trên mô hình Client-Server, trong đó phía Server đảm nhận xử lý thời gian thực, xử lý hình ảnh, nhận diện khuôn mặt và ghi lại sự có mặt, còn phía Client thực hiện chức năng ứng dụng di động chụp hình khuôn mặt của sinh viên và truyền tải chúng đến máy chủ qua LAN để nhận diện và ghi lại sự có mặt. Mô hình nhận dạng khuôn mặt sử dụng MTCNN (phát hiện mặt), FaceNet (vector trích xuất đặc trưng), và Mạng Siamese (xác thực so sánh) gộp lại làm một. Mạng Siamese với cơ chế học tập so sánh độ tương đồng giữa các cặp ảnh giúp cho xác minh danh tính học sinh dễ chính xác hơn. FaceNet là bộ chuyển đổi ảnh khuôn mặt thành các vectơ nhúng trong không gian đặc trưng được tối ưu hóa cho tính khoảng cách giữa các mẫu. Bài báo của chúng tôi không chỉ dừng lại ở lí thuyết mà còn thực nghiệm trên dữ liệu học sinh thực tế, xây dựng web và phát triển ứng dụng di động cho việc điểm danh trong các lớp học thực tế.

Điểm mới chính của bài báo là chúng tôi đề xuất một hệ thống điểm danh nhận diện khuôn mặt thực tiễn được tối ưu hóa cho việc điểm danh một lần và triển khai ngoại tuyến trên mạng LAN, thu hẹp khoảng cách giữa các mô hình học sâu và các ứng dụng lớp học thực tế trong điều kiện hạn chế.

Đóng góp cụ thể trong bài báo của chúng tôi như sau:

- Nghiên cứu các kỹ thuật nhận diện khuôn mặt bao gồm: mô hình MTCNN, FaceNet, và so sánh khả

năng định danh khuôn mặt: SVM và các mô hình học sâu tiên tiến như mạng Siamese.

- Thu thập và tiền xử lý tập dữ liệu hình ảnh của sinh viên từ Trường Đại học Công Thương Thành phố Hồ Chí Minh để kiểm tra triển khai thực nghiệm mô hình.

- Đề xuất hệ thống nhận diện khuôn mặt thời gian thực tích hợp bao gồm: (1) MTCNN, (2) FaceNet, và (3) Siamese.

- Triển khai mô hình trên máy chủ cục bộ, cho phép điểm danh với sinh viên di động qua mạng LAN.

- Phát triển ứng dụng di động để gửi hình ảnh đến máy chủ nhằm tự động điểm danh sinh viên.

- Đặc biệt chúng tôi xây dựng được một bộ công cụ thể hiện tính ưu việt trên phương pháp đề xuất: (i) Độ chính xác định danh 98,3% chỉ cần 1 ảnh đăng ký, tốt hơn SVM tới 52,6%. (ii) Hệ thống LAN chạy trọn vẹn trong 480 mili-giây, đã điểm danh thật cho 487 sinh viên. (iii) Ứng dụng đa nền tảng gồm trang web và app giúp giảng viên thực hiện công tác điểm danh sinh viên nhanh chóng, tiết kiệm thời gian một cách đáng kể so với các phương pháp khác.

2. CÁC NGHIÊN CỨU LIÊN QUAN

Nhận dạng khuôn mặt là một lĩnh vực nghiên cứu quan trọng trong lĩnh vực thị giác máy tính, đã được các nhà khoa học nghiên cứu và phát triển trong rất nhiều năm. Nghiên cứu trong lĩnh vực này có thể được chia thành ba giai đoạn chính: Giai đoạn đầu, dùng các phương pháp truyền thống dựa trên các đặc điểm thủ công, giai đoạn thứ hai dùng các kỹ thuật sử dụng mô hình thống kê và giai đoạn thứ ba các phương pháp hiện đại sử dụng học sâu.

2.1. Phương pháp truyền thống

Phương pháp Eigenfaces do Turk và cộng sự đề xuất [1], đây là một trong những công trình tiên phong trong lĩnh vực nhận diện khuôn mặt. Phương pháp này tập trung vào sự kết hợp giữa các bản đồ cạnh của khuôn mặt và các thuật toán để tính toán khoảng cách, chẳng hạn như Khoảng cách Hausdorff đã sửa đổi (MHD) và Khoảng cách Hausdorff của các đoạn thẳng (LHD), để nâng cao độ chính xác của việc nhận diện khuôn mặt gọi là "Eigenfaces". Việc nhận diện khuôn mặt diễn ra bằng cách phân tích các khoảng cách trong không gian như vậy. Phương pháp này đơn giản; tuy nhiên, nó nhạy cảm với những thay đổi về ánh sáng và góc nhìn.

Bảng 1. Bảng kết quả nhận dạng với điều kiện lý tưởng

Thuật toán	Bern	AR
Eigenface – 20 eigenvectors	100%	53%
MHD	100%	69%
LHD	100%	93%

Đối với cơ sở dữ liệu Bern, tất cả các thuật toán đều đạt độ chính xác nhận dạng 100%, do sự khác biệt giữa ảnh trong tập cơ sở dữ liệu và ảnh dùng để nhận dạng là rất nhỏ. Tuy nhiên, đối với cơ sở dữ liệu AR, sự khác biệt giữa các cặp ảnh lớn hơn do các ảnh được chụp ở hai phiên khác nhau cách nhau khoảng hai tuần. Vì vậy, độ chính xác nhận dạng của các thuật toán trên bộ dữ liệu này có sự khác biệt đáng kể. Kết quả thực nghiệm cho thấy các phương pháp MHD và LHD đạt hiệu năng tốt hơn so với phương pháp Eigenface, vốn là một phương pháp kinh điển thường được sử dụng trong các hệ thống nhận dạng khuôn mặt [1]. Ngoài ra, phương pháp LHD còn thể hiện tính ổn định trước các biến đổi của ảnh khuôn mặt như thay đổi điều kiện chiếu sáng, góc chụp và biểu cảm khuôn mặt. Bộ dữ liệu AR được thiết kế với nhiều biến đổi về ánh sáng, biểu cảm và che khuất, và các ảnh được thu thập trong hai phiên cách nhau khoảng hai tuần [2]. Với những ưu điểm đó, phương pháp nhận dạng khuôn mặt dựa trên cạnh, đặc biệt là phương pháp LHD, là một phương pháp rất tốt để ứng dụng trong nhận dạng khuôn mặt [1] SIFT và LBP (những năm 2000): Các đặc trưng SIFT (Scale-Invariant Feature Transform) [3] và LBP (Local Binary Patterns) [4] được sử dụng để trích xuất các điểm đặc trưng cục bộ từ khuôn mặt. Các phương pháp này giảm thiểu ảnh hưởng của các yếu tố như ánh sáng và biểu cảm khuôn mặt.

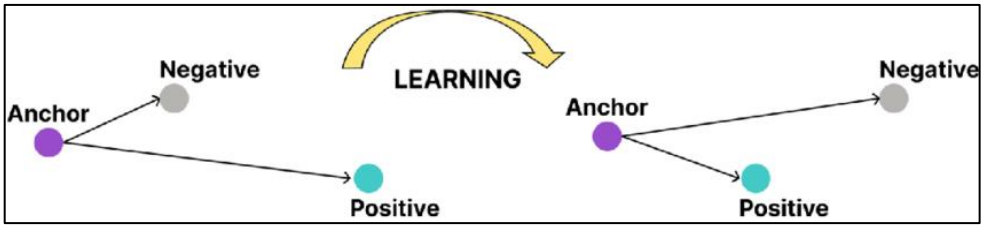
2.2. Giai đoạn sử dụng mô hình thống kê

Trong nghiên cứu này, trọng tâm là việc triển khai và so sánh các kỹ thuật phân loại hình ảnh sử dụng Máy Vector Hỗ trợ và Mạng Nơ-ron Tích chập. Đầu tiên, đội ngũ đã làm việc với một tập dữ liệu nhỏ gồm 350 bức ảnh và triển khai SVM, đạt được độ chính xác 93%. Tuy nhiên, kết quả này có thể là một dị thường do kích thước nhỏ của tập dữ liệu. Bước tiếp theo, đội đã thực hiện tăng cường dữ liệu và tăng kích thước tập dữ liệu lên hơn ba lần. Khi SVM được triển khai trên tập dữ liệu mở rộng, độ chính xác giảm xuống còn 82%. Để cải thiện kết quả này, đội đã chuyển sang sử dụng Mạng Nơ-ron Tích chập, một kỹ thuật học sâu, trong đó đội đã thực hiện các bước tiền xử lý hình ảnh và triển khai mô hình cần thiết [5].

2.3. Giai đoạn sử dụng các phương pháp hiện đại

DeepFace [6]: Bài báo này được thiết kế để giải quyết vấn đề nhận diện khuôn mặt trong các điều kiện không bị hạn chế và thu hẹp khoảng cách với khả năng nhận diện của con người. Hệ thống sử dụng một phương pháp đột phá kết hợp căn chỉnh khuôn mặt 3D chính xác với mạng nơ-ron sâu. Căn chỉnh 3D cho phép chuẩn hóa các hình ảnh khuôn mặt đầu vào và khắc phục các vấn đề về xoay, ánh sáng và biểu cảm khuôn mặt. Mạng nơ-ron sâu với 9 lớp (hơn 120 triệu tham số) được đào tạo trên tập dữ liệu khuôn mặt lớn nhất chứa 4 triệu hình ảnh của hơn 4,000 người dùng. Đặc biệt hệ thống sử dụng các tầng được kết nối cục bộ thay vì chỉ các tầng tích chập thông thường, tận dụng đặc điểm khu vực khuôn mặt các vùng mặt để tăng hiệu quả biểu diễn dữ liệu. Nghiên cứu cho thấy DeepFace đạt độ chính xác 97,35% trên tập dữ liệu Labeled Faces in the Wild (LFW), giảm hơn 27% sai số so với kỹ thuật tốt nhất trước đó, và 91,4% trên tập YouTube Faces (YTF), cũng cải thiện rõ rệt so với các phương pháp hiện tại. Điều này chứng minh hệ thống nhận diện chính xác khuôn mặt trong các điều kiện thực tế với hình ảnh chất lượng kém, góc độ khuôn mặt và biểu cảm khác nhau. Hệ thống đã phát triển khả năng tạo ra các ký hiệu mặt với ký hiệu hóa tích cực. Điều này là nhờ vào sự kết hợp giữa Conde 3D và deep learning. Nghiên cứu cũng phát hiện DNN mạng đạt hiệu suất tốt nhất trên các tập dữ liệu lớn [6].

Năm 2015 Google đã giới thiệu FaceNet, một hệ thống học sâu tạo ra các không gian nhúng để xác định và phân nhóm các khuôn mặt. Điều đặc biệt về hệ thống này là nó học nhúng các khuôn mặt sao cho khoảng cách giữa các điểm trong không gian nhúng đại diện cho mức độ tương đồng của các khuôn mặt [7]. Cách tiếp cận này loại bỏ các bước trung gian và nâng cao hiệu quả với kích thước nhúng chỉ có 128 byte cho mỗi khuôn mặt. FaceNet áp dụng mạng nơ-ron tích chập sâu cùng với hàm mất mát triplet loss để cải thiện không gian nhúng. Triplet loss giữ cho các khuôn mặt có cùng danh tính gần nhau trong không gian nhúng, trong khi các khuôn mặt khác danh tính được phân tách bằng một khoảng cách tối thiểu.



Hình 1. Hàm mất mát ba phần tử (Triplet Loss) [7]

Các nhà khoa học đã tìm ra những cách tốt hơn để dạy máy tính nhận diện khuôn mặt. Họ đã sử dụng các phương pháp thông minh khác nhau và các chương trình đặc biệt mô phỏng não bộ mang tên Zeiler, Fergus, và Inception, để giúp máy tính học nhanh hơn và hoạt động tốt hơn. Máy tính có thể nhận diện khuôn mặt rất chính xác gần như hoàn hảo trong một số bài kiểm tra. Nó cũng làm rất tốt việc nhận diện khuôn mặt trong các bộ sưu tập ảnh riêng tư, ngay cả khi ảnh được chụp trong điều kiện ánh sáng khác nhau, từ các góc độ khác nhau, hoặc nếu người trong ảnh trông hơi khác so với tuổi tác. FaceNet là một hệ thống mới dễ sử dụng hơn và hoạt động tốt hơn so với các phương pháp cũ, vốn cần thêm các bước hoặc công cụ phức tạp hơn [7].

Tiếp theo, Deng và cộng sự phát triển CosFace, phương pháp này bổ sung thêm một biên độ đơn giản để mọi thứ ổn định hơn. Phương pháp mới nhất và tốt nhất được gọi là ArcFace. Phương pháp này bổ sung thêm một chút góc (gọi là biên độ) vào cách máy tính đo khuôn mặt, làm cho sự khác biệt trở nên rõ ràng hơn và giúp máy tính học nhanh hơn và chính xác hơn. Điều này giúp máy tính nhận dạng khuôn mặt tốt hơn nhiều, ngay cả trong những tình huống khó khăn. Các nhà khoa học đang cố gắng dạy máy tính nhận dạng khuôn mặt, giống như cách chúng ta làm. Để làm được điều này, họ dạy máy tính tìm các đặc điểm đặc biệt

trên khuôn mặt (như hình dạng của mắt hoặc mũi) giúp phân biệt khuôn mặt của người này với người khác. Mục tiêu là đảm bảo rằng các đặc điểm từ cùng một người rất giống nhau (gần nhau) và các đặc điểm từ những người khác nhau rất khác nhau (xa nhau). Trước đây, họ đã sử dụng một phương pháp phổ biến gọi là Softmax, nhưng nó không hoàn hảo vì nó không làm cho các đặc điểm khác biệt như cần thiết. Vì vậy, các nhà khoa học đã đưa ra những ý tưởng mới để giúp máy tính hoạt động tốt hơn. Một ý tưởng được gọi là Center Loss, giúp các đặc điểm của cùng một người luôn gần nhau. Sau đó, họ thử SphereFace, sử dụng phương pháp đặc biệt để đo góc, nhưng rất khó để huấn luyện máy tính bằng phương pháp đó [8].

2.4. Giai đoạn sử dụng học sâu

Trong lĩnh vực nhận diện khuôn mặt, học sâu đã chuyển từ phương pháp truyền thống (như Eigenfaces[1]) sang CNN với trọng tâm tối ưu hóa loss function để giảm khoảng cách nội lớp và tăng ngoại lớp [Wang et al., 2021]. Loss truyền thống như Softmax thiếu hiệu quả với đặc thù khuôn mặt (intra-class lớn do ánh sáng/tư thế). Các hướng SOTA chính: (1) Thêm penalty như Center Loss [9] giảm nội lớp, Orthogonality Loss [10] tăng trực giao; (2) Margin-based như ArcFace đạt 99,83% trên LFW và 98,36% trên MegaFace [11]; Triplet Loss và kết hợp HST/ACT Loss cải thiện hội tụ [8] [12]. Những cải tiến này nâng cao tổng quát hóa, đặc biệt cho ứng dụng thực tế như điểm danh sinh viên với FaceNet (dựa trên Triplet Loss).

Các yếu tố cần cân nhắc khi lựa chọn mô hình:

Mặc dù các mô hình nhẹ gần đây như MobileNet và EdgeFace đã cho thấy kết quả đầy hứa hẹn trong nhận dạng khuôn mặt trên thiết bị và trên biên, mục tiêu thiết kế của chúng khác với phạm vi của nghiên cứu này. MobileNet chủ yếu được tối ưu hóa cho suy luận phía thiết bị di động dưới các ràng buộc tính toán nghiêm ngặt [13]. Ngược lại, hệ thống được đề xuất áp dụng kiến trúc máy khách-máy chủ, trong đó mobile chỉ phát hiện khuôn mặt sau đó quá trình nhận dạng khuôn mặt được thực hiện trên máy chủ cục bộ. Do đó, việc sử dụng MobileNet không mang lại lợi ích thực tiễn bổ sung trong kịch bản triển khai được xem xét.

EdgeFace được thiết kế đặc biệt cho môi trường AI nhúng và biên, nhấn mạnh tính nhỏ gọn của mô hình và hiệu quả năng lượng [14]. Tuy nhiên, nghiên cứu này tập trung vào độ chính xác xác minh khuôn mặt một lần và độ ổn định của hệ thống trong thiết lập điểm danh lớp học dựa trên mạng LAN. Hơn nữa, EdgeFace thường yêu cầu một quy trình huấn luyện chuyên dụng và tinh chỉnh quy mô lớn để khai thác tối đa các ưu điểm của nó, điều này nằm ngoài phạm vi của nghiên cứu hướng ứng dụng này.

Vì những lý do này, học metric dựa trên FaceNet/ArcFace được chọn là giải pháp phù hợp hơn để cân bằng độ chính xác nhận dạng, tính ổn định của hệ thống và tính khả thi triển khai. Việc đánh giá so sánh với các mô hình tối ưu hóa cho thiết bị biên như EdgeFace sẽ được xem xét trong tương lai khi mở rộng hệ thống sang các kịch bản triển khai hoàn toàn trên thiết bị hoặc nhúng.

2.5. Lý do lựa chọn FaceNet cho bài toán

Mặc dù các mô hình margin-based như ArcFace đạt hiệu năng SOTA trên các benchmark lớn (LFW, MegaFace), chúng chủ yếu tối ưu hóa bài toán phân lớp đóng (closed-set classification). Trong khi đó, bài toán trong nghiên cứu này thuộc dạng nhận diện mở, nơi các danh tính mới có thể xuất hiện mà không cần huấn luyện lại mô hình.

FaceNet học trực tiếp ánh xạ khuôn mặt vào không gian embedding thông qua Triplet Loss, đảm bảo rằng khoảng cách giữa các mẫu cùng danh tính luôn nhỏ hơn khoảng cách với các mẫu khác danh tính. Cách tiếp cận này giúp embedding có tính hình học rõ ràng, thuận lợi cho các tác vụ so khớp, clustering và mở rộng hệ thống.

Kết quả thực nghiệm với ArcFace tại bảng 5 cho thấy mặc dù đạt độ chính xác cao, phân bố khoảng cách giữa các cặp cùng và khác danh tính vẫn tồn tại chồng lấn. Ngược lại, FaceNet cho phép kiểm soát trực tiếp mối quan hệ khoảng cách, phù hợp hơn với mục tiêu tối ưu hóa độ tin cậy trong ứng dụng thực tế như điểm danh sinh viên.

Trái ngược với các hệ thống điểm danh dựa trên nhận diện khuôn mặt hiện có, chủ yếu dựa vào các phương pháp phân loại và triển khai trên nền tảng đám mây, hệ thống của chúng tôi được thiết kế như một giải pháp dựa trên xác thực, hoạt động hoàn toàn trong môi trường mạng LAN cục bộ. Thiết kế này cho phép điểm danh một lần duy nhất, khả năng mở rộng tăng dần và triển khai thực tế trong môi trường lớp học thực tế.

3. TỔNG QUAN LÝ THUYẾT

3.1. Tổng quan về hệ thống nhận dạng khuôn mặt sinh trắc học

3.1.1. Hệ thống sinh trắc học

Sinh trắc học là một phương pháp sử dụng các bộ phận trên cơ thể hoặc hành vi của chúng ta để chứng minh bản thân. Ví dụ, những thứ như dấu vân tay, diện mạo khuôn mặt, kiểu mắt, giọng nói hoặc cách đi lại có thể giúp nhận dạng bạn. Ngày nay, chúng ta sử dụng sinh trắc học để giữ an toàn cho các địa điểm, kiểm soát người được phép ra vào, thanh toán trực tuyến, và thậm chí trong trường học để tự động điểm danh, giúp chúng ta biết ai đang ở đó. Một thông tin sinh trắc học lý tưởng cần đáp ứng các tiêu chí sau:

- Tính phổ biến: Mỗi cá nhân đều sở hữu thông tin sinh trắc học này.
- Tính duy nhất: Đặc điểm sinh trắc học phải phân biệt tối đa giữa các cá nhân khác nhau.
- Tính vĩnh viễn: Đặc điểm sinh trắc học không thay đổi đáng kể trong suốt cuộc đời.
- Tính thu thập được: Có thể đo lường và thu thập dễ dàng.
- Tính chấp nhận được: Phương pháp thu thập và sử dụng đặc điểm sinh trắc học phải được người dùng chấp nhận.

Các quy tắc trên giúp chúng ta quyết định xem việc sử dụng công nghệ đặc biệt để nhận diện khuôn mặt có phải là một ý tưởng tốt hay không. Trong số các loại công nghệ này, nhận diện khuôn mặt được ưa chuộng vì dễ sử dụng, không gây phiền hà cho người khác, có thể được thiết lập với camera đã lắp đặt sẵn và hoạt động hiệu quả trong việc theo dõi thời gian mọi người đến hoặc rời khỏi trường học và nơi làm việc.

3.1.2. Những thách thức trong nhận dạng khuôn mặt

Bài toán nhận dạng khuôn mặt người đã được nghiên cứu từ những năm 1970 và đến nay vẫn là một chủ đề quan trọng trong lĩnh vực thị giác máy tính và trí tuệ nhân tạo. Tuy nhiên, đây là một bài toán phức tạp do sự biến đổi lớn của khuôn mặt trong các điều kiện khác nhau, vì vậy nhiều thách thức vẫn chưa được giải quyết hoàn toàn. Do đó, bài toán này vẫn tiếp tục thu hút sự quan tâm của nhiều nhóm nghiên cứu trên thế giới [9] [15]. Một số khó khăn chính trong bài toán nhận dạng khuôn mặt có thể kể đến như sau: (i) **Tư thế và góc chụp:** Hình ảnh khuôn mặt có thể thay đổi đáng kể do sự khác nhau về góc chụp giữa máy ảnh và đối tượng. Ví dụ, khuôn mặt có thể được chụp trực diện, nghiêng sang trái hoặc phải, hoặc từ các góc trên xuống và dưới lên. Với các góc chụp khác nhau, một số đặc điểm như mắt, mũi hoặc miệng có thể bị che khuất một phần hoặc hoàn toàn. (ii) **Sự thay đổi của các đặc điểm ngoại hình:** Một số yếu tố như râu, ria mép, kính hoặc phụ kiện khác có thể xuất hiện hoặc biến mất theo thời gian, làm tăng độ khó của quá trình nhận dạng. (iii) **Biểu cảm khuôn mặt:** Các biểu cảm như cười, buồn hoặc sợ hãi có thể làm thay đổi đáng kể hình dạng và đặc trưng của khuôn mặt. (iv) **Che khuất (Occlusion):** Khuôn mặt có thể bị che khuất bởi các vật thể khác hoặc bởi các khuôn mặt khác trong ảnh. (v) **Biến đổi tư thế (Pose variation):** Sự thay đổi về hướng của đầu hoặc trục máy ảnh có thể làm cho khuôn mặt bị nghiêng hoặc biến dạng trong ảnh. (vi) **Điều kiện chụp ảnh:** Các yếu tố như điều kiện ánh sáng, loại thiết bị chụp hoặc chất lượng ảnh có thể ảnh hưởng đáng kể đến hiệu quả của hệ thống nhận dạng. (vii) **Sự lão hóa:** Khuôn mặt con người thay đổi theo thời gian, khiến việc nhận dạng qua nhiều năm trở nên khó khăn hơn. (viii) **Quy mô cơ sở dữ liệu lớn:** Trong các hệ thống thực tế, cơ sở dữ liệu khuôn mặt có thể chứa hàng triệu hoặc thậm chí hàng tỷ ảnh, tạo ra thách thức lớn về khả năng tính toán và lưu trữ [16].

3.2. Tổng quan về kiến trúc của hệ thống nhận dạng khuôn mặt con người

Hệ thống đề xuất của chúng tôi cho nhận dạng điểm danh khuôn mặt sinh viên bao gồm bốn bước xử lý như sau: (i) Phát hiện khuôn mặt. (ii) Căn chỉnh hoặc phân đoạn khuôn mặt. (iii) Trích xuất đặc trưng. (iv) Nhận dạng phân loại khuôn mặt. Trong phần sau đây chúng tôi giới thiệu các nội dung kiến thức liên quan trong hệ thống này.

3.2.1. Tổng quan về Multi-Task Cascaded Convolutional Networks

MTCNN (Multi-Task Cascaded Convolutional Networks) là một mạng nơ-ron tích chập sâu phát hiện khuôn mặt, có nghĩa là nó phát hiện khuôn mặt trong hình ảnh tĩnh hoặc video. Việc xác định khuôn mặt và các đặc điểm quan trọng khác như mắt, mũi và miệng là một vấn đề chính trong xử lý hình ảnh. MTCNN đạt

được điều này bằng cách sử dụng ba lớp mạng tích chập chính, tương ứng với ba giai đoạn (hoặc quy trình phụ) của việc xử lý khuôn mặt. Các giai đoạn này được gọi là P-Net (Mạng Đề xuất), R-Net (Mạng Tinh chỉnh) và O-Net (Mạng Đầu ra) hoặc ngắn gọn là PRO.

Ưu điểm: Tích hợp nhiều tác vụ, MTCNN kết hợp phát hiện khuôn mặt và phát hiện điểm mốc trong khuôn mặt giúp giảm độ phức tạp thay vì phải dùng các mô hình riêng lẻ. Hiệu suất cao, MTCNN được thiết kế theo mô hình thác đổ với khả năng nhanh chóng lọc bỏ các vùng không liên quan từ rất sớm trong quy trình giúp tối ưu hóa thời gian và tài nguyên tính toán. Độ chính xác cao, do được đào tạo trên các bộ dữ liệu lớn và các phương pháp tinh chỉnh như hồi quy hộp giới hạn, độ chính xác trong phát hiện khuôn mặt và điểm mốc của MTCNN tăng đáng kể. Khả năng xử lý các khuôn mặt có nhiều kích thước nhờ dùng mô hình kim tự tháp hình ảnh. MTCNN có thể phát hiện khuôn mặt ở nhiều kích thước và góc độ khác nhau. Nhược điểm: MTCNN có thể tốn nhiều thời gian hơn và yêu cầu rất nhiều tài nguyên do tính toán hiệu quả của ba mạng nơ-ron và việc xử lý nhiều tỷ lệ hình ảnh. Khi ảnh chất lượng thấp, MTCNN bị suy yếu hiệu suất với ảnh mờ, nhiễu, khuôn mặt bị che hay phân giải thấp. Khó cân bằng thời gian thực, với thiết bị tính toán yếu như điện thoại hay IoT, MTCNN bị chậm và suy giảm hiệu suất [17].

3.2.2. Tổng quan về FaceNet

FaceNet là một mô hình học sâu tiên tiến do Google phát triển, nhằm phục vụ các ứng dụng trong nhận diện và xác thực khuôn mặt. Ngoài việc nhận diện và xác thực khuôn mặt, FaceNet còn thực hiện việc chuyển đổi các khuôn mặt thành các vectơ đặc trưng (embedding). Điều này cho phép thực hiện các tác vụ như: nhận diện khuôn mặt và xác định một người trong tập ảnh khuôn mặt đã biết. Xác minh khuôn mặt, Kiểm tra hai khuôn mặt có thuộc về một người hay không. Phân cụm khuôn mặt, Tương tự các mặt được nhóm lại thành cụm dựa trên các đặc trưng [18].

3.2.3. Tổng quan Siamese Network

Kiến trúc cơ bản của Mạng Siamese: Mạng Siamese (Mạng Nơ-ron Siamese - SNN) là một kiến trúc học sâu bao gồm hai hoặc nhiều mạng con giống hệt nhau, với cùng tham số và trọng số. Mỗi mạng con nhận một ảnh đầu vào, trích xuất các vectơ đặc trưng, sau đó so sánh các vectơ để xác định độ tương đồng. Mạng Siamese đặc biệt hữu ích trong các tác vụ như xác minh khuôn mặt, chữ ký hoặc nhận dạng với dữ liệu hạn chế. Trong bài báo hệ thống đề xuất của chúng tôi thực hiện quá trình xác minh khuôn mặt sinh viên diễn ra theo các bước sau: (1) MTCNN phát hiện và cắt khuôn mặt từ ảnh đầu vào. (2) FaceNet trích xuất vectơ đặc trưng (nhúng) của khuôn mặt đã cắt. (3) Vector nhúng này được so sánh với các vectơ được lưu trữ trong cơ sở dữ liệu sinh viên bằng Mạng Siamese. (4) Nếu độ tương đồng vượt quá một ngưỡng nhất định, hệ thống sẽ xác minh danh tính của sinh viên. Quy trình xử lý của SNN có thể được tóm tắt như sau: (Bước 1) Chuẩn bị các cặp dữ liệu đầu vào chọn một cặp dữ liệu (trong trường hợp này là hình ảnh) từ tập dữ liệu. Một cặp có thể bao gồm hai hình ảnh thuộc cùng một lớp (cặp dương) hoặc khác lớp (cặp âm). Việc tổ chức này giúp mạng học được sự khác biệt hoặc tương đồng giữa các dữ liệu. (Bước 2) Xử lý thông qua các mạng con mỗi hình ảnh trong cặp được truyền qua một mạng con (sub-network) giống hệt nhau về kiến trúc và trọng số. Các mạng con này chịu trách nhiệm chuyển đổi dữ liệu đầu vào thành một biểu diễn vector gọi là vector nhúng. Các vector nhúng này chứa thông tin đặc trưng của dữ liệu đầu vào, giúp biểu diễn dữ liệu dưới dạng số để dễ dàng tính toán. (Bước 3) Tính toán sự khác biệt cuối cùng, chúng tôi lấy hai vectơ nhúng từ các mạng con và sử dụng các phương pháp đo độ tương đồng khác nhau để tìm kiếm các sự khác biệt của chúng. Có một số phép đo độ tương đồng hữu ích như là ED, sử dụng để đếm đường thẳng trên mặt Phẳng độ dài giữa 2 vectơ. Các kỹ thuật khác, như Cosine và L1/L2 Norm cũng có thể được áp dụng. (Bước 4) Áp dụng hàm sigmoid khoảng cách giữa hai vectơ, được chuyển qua hàm Sigmoid, chuẩn hóa kết quả thành một giá trị ở khoảng [0,1]. Điều này được gọi là điểm, thể hiện cho độ giống nhau giữa hai vector. Nếu điểm càng gần 1, hai vectơ càng giống nhau, nghĩa là chúng thuộc cùng một lớp của hình. Ngược lại, nếu tích vô hướng gần bằng 0, nghĩa là hai ảnh thuộc các lớp khác nhau. (Bước 5) Đưa ra quyết định cuối cùng bằng cách áp dụng điểm giá trị, SNN dự đoán liệu hai hình ảnh chứa cùng một lớp hay không. Đây không chỉ đơn giản hóa quá trình phân loại mà còn cho phép thêm lớp mà không cần đào tạo mạng từ đầu [19].

Ưu điểm của mạng Siamese trong nhận dạng khuôn mặt: Mạng nơ-ron Siamese (SNN) giúp xử lý các vấn đề với dữ liệu ít hoặc dữ liệu phức tạp hơn. Dưới đây là những ưu điểm mà SNN nổi bật hơn so với mạng nơ-ron truyền thống trong vấn đề này: Lượng dữ liệu huấn luyện cần thiết nhỏ SNN là mạng nơ-ron hiếm hoi có khả năng hoạt động tốt trong điều kiện dữ liệu huấn luyện rất ít. Không như mạng nơ-ron truyền thống cần tới hàng ngàn mẫu để huấn luyện, SNN chỉ cần từ 1 tới 5 mẫu. Điều này được giải thích là do SNN áp dụng các kĩ thuật One-Shot Learning hoặc Few-Shot Learning. Điều này giúp tiết kiệm chi phí và thời gian thu thập dữ liệu, đồng thời có ý nghĩa quan trọng đối với các trường hợp dữ liệu khó hoặc không thể thu thập

đầy đủ, chẳng hạn như chữ ký hoặc các dữ liệu chuyên biệt khác. Ngoài ra, việc phần dữ liệu sử dụng càng ít sẽ càng giúp giải quyết vấn đề mất cân bằng dữ liệu, do không buộc phải cân bằng các lớp dữ liệu cho đủ. Kết hợp linh hoạt với các bộ phân loại khác vì phương pháp học của SNN dựa trên sự tương đồng của các cặp dữ liệu thay vì phân loại trực tiếp nên nó có thể dễ dàng kết hợp với các bộ phân loại khác như Support Vector Machine hoặc Random Forest để tăng hiệu suất. Khi kết hợp SNN, nó hoạt động như một bộ kết xuất đặc trưng, đo vẽ cho mô hình phân loại và cửa phân loại khác. Kỹ thuật này cung cấp các kết quả vượt trội so với việc sử dụng một mô hình, đặc biệt là trong các bài toán phân loại phức tạp. Học từ sự tương đồng ngữ nghĩa bởi vì SNN học các sự tương đồng và mối quan hệ giữa các cặp dữ liệu thay vì trực tiếp phân loại hai tập dữ liệu, nó có thể cải thiện hiệu suất của các bộ phân loại truyền thống như Support Vector Machines (SVM) hoặc Random Forests. Sự tích hợp SNN đóng vai trò như một bộ trích xuất đặc trưng với một biểu diễn vector mạnh mẽ để sử dụng trong các bộ phân loại khác. Điều này là do xu hướng của nó cung cấp hiệu suất vượt trội hơn so với việc sử dụng một mô hình đơn lẻ, đặc biệt là trong các nhiệm vụ phân loại khó khăn hơn. Bên cạnh những ưu điểm, trình xác minh Siamese cũng gặp ba vấn đề về khả năng mở rộng: Chi phí huấn luyện: Lấy mẫu cặp dẫn đến số lượng bộ ba tăng cao; Không có xác suất trên lớp: Đầu ra là một điểm tương đồng duy nhất và không thể báo cáo độ tin cậy theo kiểu softmax. Độ trễ trong việc đăng ký: Thêm một sinh viên vẫn kích hoạt 487 phép so sánh cặp (42 ms trên CPU). Các vấn đề này được giải quyết một phần khi khai thác hard-negative trên-web và sử dụng FAISS[19].

4. PHƯƠNG PHÁP GIẢI QUYẾT VẤN ĐỀ

4.1. Định nghĩa

MTCNN - Phát hiện khuôn mặt: MTCNN là một mạng nơ-ron tích chập (CNN) đa tác vụ, bao gồm ba mạng con giúp phát hiện khuôn mặt, xác định các điểm đặc trưng trên khuôn mặt (như mắt, mũi, miệng), và ước lượng kích thước khuôn mặt. Phương pháp này có thể phát hiện khuôn mặt trong các điều kiện ánh sáng khác nhau và các góc nhìn khác nhau. Chúng tôi định nghĩa như sau:

$$\mathcal{L}_{\text{MTCNN}} = \mathcal{L}^{\text{det}} + 0.5\mathcal{L}^{\text{box}} + 0.5\mathcal{L}^{\text{lmk}} \quad (1)$$

$$\begin{aligned} \mathcal{L}^{\text{det}} &= -[y \log p + (1-y) lo\, g(1-p)] \\ \mathcal{L}^{\text{box}} &= \|\hat{b} - b\|2^2 \\ \mathcal{L}^{\text{lmk}} &= \|\hat{l} - l\|2^2 \end{aligned} \quad (2)$$

Trong đó: $y \in \{0,1\}$ nhãn thật (1= khuôn mặt), $p \in [0,1]$ xác suất dự đoán bởi P/R/O-Net, $b = (x, y, w, h)$ giới hạn thật vùng mặt, $\hat{b}$ vùng dự đoán, $l \in \mathbb{R}^{10}$ 5 điểm mốc (mắt trái/phải, mũi, miệng trái/phải), (mắt trái: (x_1, y_1), mắt phải: (x_2, y_2), mũi: (x_3, y_3), miệng trái: (x_4, y_4), miệng phải: (x_5, y_5)), $\hat{l}$ 5 điểm mốc dự đoán.

Facenet - Trích xuất đặc trưng khuôn mặt: Facenet sử dụng một mô hình CNN để trích xuất đặc trưng của khuôn mặt dưới dạng vector.

FaceNet ánh xạ ảnh → vector 128-D:

$$f(I) = \frac{\text{CNN}(I)}{\|\text{CNN}(I)\|_2} \in \mathbb{R}^{128} \quad (3)$$

Huấn luyện bằng Triplet Loss:

$$\mathcal{L}_{\text{triplet}} = [\underbrace{\|f^a - f^p\|2^2}_{\text{khoảng cách cùng người}} - \underbrace{\|f^a - f^n\|2^2}_{\text{khoảng cách khác người}} + 0.2\,]_+ \quad (4)$$

Trong đó: f^a anchor - ảnh gốc, f^p positive - cùng người với f^a, f^n negative - khác người với f^a, $[]_+ = \max(0, \cdot)$

Siamese Network – Xác minh khuôn mặt: Sau khi trích xuất đặc trưng khuôn mặt bằng mô hình FaceNet, hệ thống sử dụng Siamese Network để thực hiện xác minh danh tính thông qua việc **so** sánh độ tương đồng giữa hai vector đặc trưng. Nhờ cơ chế học dựa trên tương đồng, Siamese Network đặc biệt hiệu quả trong các bài toán xác minh danh tính, và có thể hoạt động tốt ngay cả với những danh tính chưa xuất hiện trong tập huấn luyện. Mô hình kiến trúc sử dụng trong bài toán. Bài toán chứng thực sinh trắc học sử dụng nhận diện khuôn mặt có thể được phân thành ba bước chính: phát hiện khuôn mặt, trích xuất đặc trưng

khuôn mặt, và phân loại khuôn mặt. Mỗi bước đều sử dụng các mô hình khác nhau để xử lý và đạt được kết quả tối ưu. Siamese so sánh 2 vector bằng Cosine Similarity:

$$s(q,g) = \frac{q \cdot g}{\|q\|_2 \|g\|_2} \overset{\tau=\theta}{\rightarrow} \begin{cases} \text{"Có mặt"} & s \geq \theta \\ \text{"Unknown"} & \text{ngược lại} \end{cases} \quad (5)$$

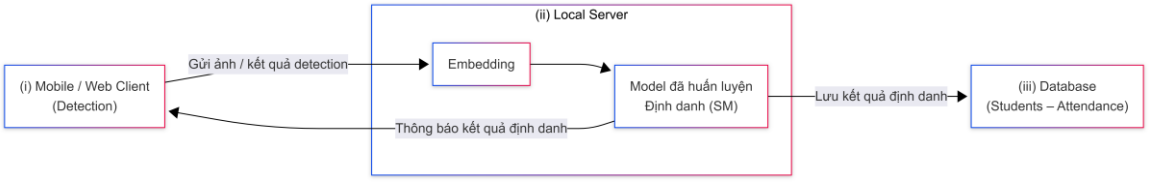
Trong đó: $q \in \mathbb{R}^{128}$ vector query - ảnh hiện tại, $g \in \mathbb{R}^{128}$ vector gallery - ảnh đăng ký, $g_{\min}$ vector gần nhất tìm. θ là ngưỡng tương đồng Cosine, thường chọn dựa trên ROC curve từ validation set.

4.2. Kiến trúc phần mềm

Hệ thống điểm danh nhận diện khuôn mặt được đề xuất sử dụng kiến trúc máy khách-máy chủ được thiết kế để triển khai thực tế trong lớp học. Hệ thống của chúng tôi bao gồm ba thành phần chính: (i) thu thập dữ liệu phía máy khách, (ii) máy chủ cục bộ để xử lý nhận diện khuôn mặt, và (iii) cơ sở dữ liệu và mô-đun quản lý.

Ở phía máy khách, cả ứng dụng web và ứng dụng di động được sử dụng để chụp ảnh và phát hiện (detection) khuôn mặt trong các buổi học. Những hình ảnh này được truyền đến máy chủ cục bộ qua mạng LAN để đảm bảo độ trễ thấp và bảo mật dữ liệu. Máy chủ thực hiện trích xuất đặc trưng bằng cách sử dụng các mô hình học sâu dựa trên số liệu (FaceNet), và so khớp độ tương đồng để xác minh danh tính.

Kết quả nhận diện sau đó được lưu trữ trong cơ sở dữ liệu tập trung, nơi lưu giữ hồ sơ học sinh, hình ảnh đăng ký và hồ sơ điểm danh. Kiến trúc này tách biệt việc thu thập dữ liệu khỏi quá trình xử lý nhận diện, cho phép hệ thống đạt được hiệu suất mạnh mẽ trong khi vẫn có khả năng mở rộng và dễ dàng triển khai trong môi trường lớp học thực tế.



Hình 2. Kiến trúc phần mềm của hệ thống chấm công nhận diện khuôn mặt được đề xuất.

4.3. Đề xuất thuật toán

Công thức tổng quát toàn bộ quá trình:

$$\underbrace{I_t}_{\text{selfie}} \overset{\text{MTCNN}}{\rightarrow} \underbrace{(B, L)}_{\text{khuôn mặt+ lmk}} \overset{\text{Affine}}{\rightarrow} \underbrace{f}_{160\times160} \overset{\text{FaceNet}}{\rightarrow} \underbrace{q}_{\mathbb{R}^{128}} \overset{\text{Siamese}}{\rightarrow} \underbrace{\text{"SV201234"}}_{\text{ID}} \quad (6)$$

Algorithm 1: ĐIỂM DANH KHUÔN MẶT

Input: Selfie $I_t \in \mathbb{R}^{720\times1280\times3}$
Output: ID sinh viên + timestamp

$\mathcal{P} \leftarrow$ Pyramid(I_t, scales={1.0,0.71,0.5,0.35}) // MTCNN
$C \leftarrow$ P-Net($\mathcal{P}$), NMS(iou=0.7) $B \leftarrow$ R-Net(C), NMS(iou=0.5) $(B, L) \leftarrow$ O-Net(B) //
if $B = \emptyset$ **then**
  return "Không thấy mặt"
else
  $M \leftarrow$ Affine($L_{\text{ref}} \rightarrow L$) //
  $f \leftarrow$ warp(I_t, M, 160×160) $q \leftarrow$ FaceNet(f) → L2-norm(q)
  $g_{\min} \leftarrow$ FAISS-kNN(q, $\mathcal{G}$) $s \leftarrow$ Siamese(q, $g_{\min}$) **if** $s \geq \theta$ **then**
    Database: INSERT attendance(id,time) VALUES($g_{\min}$.id, NOW())
    return "SV$g_{\min}$.id – Có mặt"
  else
    return "Unknown"

Hình 3. Thuật toán điểm danh nhận dạng sinh viên.

Trong nghiên cứu này, chúng tôi đề xuất **Thuật toán 1** tại hình 2 để giải quyết toàn bộ quá trình điểm danh sinh viên tự động dựa trên nhận diện khuôn mặt. Thuật toán được thiết kế một cách hệ thống, kế thừa

các mô hình hiện đại như MTCNN, FaceNet và Siamese Network, tạo thành một pipeline có tên MFSN khép kín từ khi thu nhận ảnh selfie đến khi ghi nhận kết quả điểm danh vào cơ sở dữ liệu. Toàn bộ quy trình có thể được tóm tắt và mô tả chi tiết theo các giai đoạn chính sau:

Giai đoạn phát hiện và căn chỉnh khuôn mặt (Dòng 1-6)

Mục tiêu của giai đoạn này là xác định chính xác vị trí khuôn mặt trong ảnh đầu vào và chuẩn hóa nó để sẵn sàng cho việc trích xuất đặc trưng.

– Đầu vào: Một khung hình selfie I_t được thu thập từ thiết bị camera

– Xây dựng Image Pyramid (Dòng 1): Ảnh đầu vào được biến đổi thành một Image Pyramid đa tỷ lệ (scales $= \{1{,}0; 0{,}71; 0{,}5; 0{,}35\}$) để đảm bảo thuật toán có thể phát hiện khuôn mặt ở nhiều kích thước và khoảng cách khác nhau so với camera

– Pipeline MTCNN (Dòng 2-4): Ảnh được đưa tuần tự qua ba mạng con:

- P-Net (Proposal Network): Thực hiện phát hiện sơ bộ, đưa ra các ứng viên là khuôn mặt tiềm năng. Kết quả sau đó được làm sạch bằng kỹ thuật Non-Maximum Suppression (NMS) với ngưỡng IoU=0,7 để loại bỏ các hộp bao trùng lặp
- R-Net (Refine Network): Tiếp nhận các ứng viên từ P -Net để tinh chỉnh và loại bỏ thêm các hộp bao không khớp, với NMS(iou=0,5)
- O-Net (Output Network): Mạng cuối cùng cho đầu ra chính xác nhất, bao gồm hộp giới hạn khuôn mặt B và vị trí của 5 điểm mốc quan trọng L (mắt trái/phải, mũi, mép trái/phải của miệng)

– Kiểm tra & Căn chỉnh (Dòng 5-6): Nếu không phát hiện thấy khuôn mặt nào ($B = \emptyset$), hệ thống ngay lập tức trả về thông báo "Không thấy mặt". Ngược lại, một phép biến đổi Affine được áp dụng dựa trên các điểm mốc L để căn chỉnh và "uốn" khuôn mặt về góc nhìn chính diện chuẩn, sau đó cắt và resize thành ảnh khuôn mặt chuẩn hóa có kích thước 160×160 pixel.

Giai đoạn trích xuất đặc trưng (Dòng 7-8)

– Giai đoạn này chuyển đổi hình ảnh khuôn mặt đã được chuẩn hóa thành một biểu diễn toán học (embedding vector) trong không gian đặc trưng.

– FaceNet (Dòng 7): Ảnh khuôn mặt 160×160 được đưa vào mô hình FaceNet (sử dụng kiến trúc nền tảng Inception-ResNet-v1). Mô hình này ánh xạ hình ảnh thành một vector đặc trưng 128 chiều $f(I) = \mathrm{CNN}(I)$

– Chuẩn hóa L2 (Dòng 8): Vector đặc trưng f được chuẩn hóa bằng chuẩn L2 để thành vector đơn vị q, tức $q = f/\|f\|_2$. Bước này là cực kỳ quan trọng, nó đảm bảo rằng phép đo khoảng cách hoặc độ tương đồng ở bước tiếp theo được thực hiện một cách chính xác và hiệu quả. Vector q chính là "dấu vân tay số" của sinh viên trong hệ thống.

Giai đoạn truy vấn và xác minh định danh (Dòng 9-13)

– Đây là giai đoạn quyết định, nơi hệ thống so sánh "dấu vân tay số" vừa thu được với cơ sở dữ liệu để tìm ra danh tính phù hợp.

– Tìm kiếm nhanh với FAISS (Dòng 9): Thay vì so sánh tuần tự với tất cả các vector trong cơ sở dữ liệu $\mathcal{G}$ (gallery), hệ thống sử dụng thư viện FAISS để tìm kiếm nhanh vector $g_{\min}$ trong $\mathcal{G}$ có khoảng cách gần nhất với vector truy vấn q. Điều này đảm bảo tốc độ xử lý thời gian thực ngay cả khi số lượng sinh viên trong cơ sở dữ liệu là rất lớn

– Tính toán độ tương đồng với Siamese Network (Dòng 10): Độ tương đồng Cosine s giữa q và $g_{\min}$ được tính toán: $s = (q \cdot g_{\min})/(\|q\|_2 \|g_{\min}\|_2)$. Do cả hai vector đều đã được chuẩn hóa, công thức này tương đương với tích vô hướng $s = q \cdot g_{\min}$, cho kết quả trong khoảng $[-1,1]$, với giá trị càng gần 1 thì hai khuôn mặt càng giống nhau.

– Ra quyết định và ghi nhận điểm danh (Dòng 11-13): Hệ thống so sánh độ tương đồng s với một ngưỡng xác định trước θ (theta).

– Nếu $s \geq \theta$: Hệ thống kết luận đã nhận diện thành công sinh viên. Hệ thống sẽ thực hiện hai thao tác:

1. Ghi lại lịch sử điểm danh vào cơ sở dữ liệu với lệnh INSERT: lưu ID của sinh viên ($g_{\min}.\mathrm{id}$) và thời điểm điểm danh (NOW())

2. Trả về thông báo kết quả cho người dùng, ví dụ: "SV201234 - Có mặt"

– Ngược lại ($s < \theta$) : Hệ thống không đủ tự tin để khớp với bất kỳ sinh viên nào đã đăng ký và trả về kết quả "Unknown".

5. THỰC NGHIỆM & KẾT QUẢ

5.1. Bộ dữ liệu

5.1.1. VGGFace2

VGGFace2 được xem là một trong những bộ dữ liệu đầu tiên có sẵn công khai để sử dụng trong nghiên cứu nhận dạng và xác minh khuôn mặt. Nó được xây dựng bởi các nhóm nghiên cứu tại Đại học Oxford [20]. So với các phiên bản trước của nó là VGGFace, VGGFace2 toàn diện hơn, bao gồm hình ảnh khuôn mặt chính diện và không chính diện với các độ phân giải, biểu cảm và ánh sáng khác nhau. Bộ dữ liệu này được tạo ra nhằm giúp các mô hình học sâu học được các đặc trưng nhận dạng khuôn mặt hiệu quả hơn. VGGFace2 chứa hơn 3,3 triệu hình ảnh cho 9.131 danh tính khác nhau, với mỗi danh tính có từ 87 đến hơn 800 hình ảnh. Các hình ảnh được thu thập từ các nguồn công khai để nắm bắt một loạt các biến thể trong danh tính, và cân bằng về giới tính và dân tộc. Tập dữ liệu VGGFace2 được chia làm hai phần theo chức năng để phục vụ cho huấn luyện và kiểm thử các mô hình nhận diện khuôn mặt. Phần **tập huấn luyện (train)** với hơn 3 triệu hình ảnh đã được thu thập và 8.631 ảnh danh tính được mô hình hóa và nội dung khá đa dạng. Các ảnh trong phần tập huấn luyện thể hiện các biểu cảm khác nhau ở các khuôn mặt và từ nhiều góc khác nhau trong những điều kiện môi trường khác nhau. Đối với phần **tập kiểm thử (test)** được sử dụng để đánh giá mô hình, đã thu thập được hơn 400.000 ảnh cho 500 danh tính. Số hình ảnh này tách biệt và không dùng trong huấn luyện, với mục đích chính là đo đạc khả năng tổng quát hóa của mô hình.

VGGFace2 được thiết kế để có những đặc điểm phong phú nhằm tái tạo điều kiện nhận diện khuôn mặt một cách thực tế khi được sử dụng trong các tình huống đời thực. Sự biến đổi chủ yếu đến từ các góc nhìn khác nhau. Các hình ảnh được chụp từ các góc nhìn khác nhau: chính diện, bên trái, bên phải, phía trên và phía dưới. Điều này là cần thiết nếu một mô hình muốn nhận diện một khuôn mặt khi kiểu dáng chuẩn không có. Thứ hai, phạm vi biểu cảm là quan trọng, và một khuôn mặt có thể được chụp trong một trong các trạng thái sau: mỉm cười, nhắm mắt, nhăn mặt, và trung tính, điều này cho phép mô hình nhận diện các trạng thái khác nhau của một người trong bối cảnh tương tác.

Bước tiền xử lý dữ liệu VGGFace2 ban đầu, dữ liệu VGGFace2 được tiền xử lý qua bảy bước chung, với mục tiêu chính là tối ưu hóa dữ liệu cho việc đào tạo mô hình học sâu.

Bước 1: Tiêu chuẩn hóa hình ảnh tất cả hình ảnh được thay đổi kích thước về kích thước tiêu chuẩn 160×160 pixel, chuyển đổi sang định dạng jpg và đảm bảo không gian màu là RGB. **Bước 2:** Cắt và định hình khuôn mặt: MTCNN được sử dụng để phát hiện khuôn mặt, khuôn mặt được cắt theo đúng và định hình dựa trên các đặc điểm khuôn mặt: mắt, mũi và miệng. **Bước 3**: Tiêu chuẩn hóa dữ liệu: các giá trị pixel được tiêu chuẩn hóa về các khoảng [0,1] hoặc [-1,1] và Z-score được sử dụng để loại bỏ ảnh hưởng của ánh sáng và tiếng ồn một cách thống kê. **Bước 4**: Dọn dẹp và lọc dữ liệu: các hình ảnh có tiếng ồn và mờ hoặc với khuôn mặt không hoàn chỉnh được loại bỏ và các danh tính có quá ít hình ảnh bị xóa. **Bước 5**: Chia theo danh tính với 8.631 danh tính cho tập huấn luyện và 500 danh tính cho tập kiểm thử. **Bước 6:** Tăng cường dữ liệu áp dụng các kỹ thuật như lật ngang, điều chỉnh độ sáng, dịch chuyển, xoay và phóng to/thu nhỏ để tăng tính đa dạng. **Bước 7:** Tổ chức và lưu trữ sắp xếp ảnh theo thư mục danh tính với mã ID cụ thể, đảm bảo cấu trúc nhất quán cho quá trình huấn luyện. Quy trình này đảm bảo dữ liệu đầu vào chất lượng cao, giúp nâng cao hiệu suất của mô hình nhận diện khuôn mặt.

5.1.2. Bộ dữ liệu HUIT

Ngoài bộ dữ liệu công khai VGGFace2, một bộ dữ liệu nội bộ nhỏ đã được thu thập tại Trường Đại học Công Thương Thành phố Hồ Chí Minh (HUIT) để phản ánh kịch bản điểm danh lớp học thực tế. Bộ dữ liệu này bao gồm 30 sinh viên từ một lớp học, với 5 hình ảnh khuôn mặt mỗi sinh viên được chụp trong điều kiện không hạn chế, dẫn đến thiết lập ít mẫu.

Các mẫu nội bộ của HUIT được kết hợp với VGGFace2 và được sử dụng để huấn luyện và kiểm tra

theo cùng một giao thức thử nghiệm được mô tả trong bài báo này. Mục đích của việc kết hợp bộ dữ liệu nội bộ này không phải để thiết lập một chuẩn mực mới, mà để mô phỏng quá trình thực tế điểm danh sinh viên mới với số lượng mẫu hạn chế cho mỗi danh tính trong một hệ thống điểm danh thực tế.

Ngoài ra, để xác thực ở cấp độ hệ thống, ứng dụng web và di động được phát triển đã được triển khai trong ba lớp học thực tế, mỗi lớp gồm khoảng 40 sinh viên với 5 hình ảnh khuôn mặt mỗi sinh viên. Dữ liệu này chỉ được sử dụng để đánh giá chức năng đầu cuối và tính khả thi thực tiễn của hệ thống được đề xuất, và không được sử dụng trong quá trình huấn luyện mô hình hoặc so sánh hiệu suất định lượng.

5.2. Chỉ số đánh giá

Chúng tôi đánh giá model định danh tập trung vào các chỉ số hiệu năng chính bao gồm:

Độ chính xác (Accuracy): Tỷ lệ dự đoán đúng trên tổng số dự đoán.

Độ lớn dữ liệu huấn luyện: Quy mô và chất lượng (số lượng, độ đa dạng) của dữ liệu dùng để train model.

Thời gian xử lý: Thời gian cần để xử lý và so khớp một khuôn mặt. Tốc độ, tính khả thi cho ứng dụng thời gian thực.

Khả năng mở rộng: Khả năng duy trì hiệu suất khi số lượng người trong cơ sở dữ liệu tăng lên rất lớn. Tính ổn định và hiệu quả khi triển khai ở quy mô lớn.

Hiệu suất (dữ liệu ít): Khả năng nhận diện chính xác một người mới chỉ với 1 hoặc vài ảnh mẫu. Tính linh hoạt và khả năng ứng phó với tình huống thực tế, ít dữ liệu.

5.3. Kết quả trên VGGFace2

Chúng tôi tiến hành thực nghiệm theo thuật toán tổng quát đề xuất: Phương pháp đề xuất **Bước 1**: MTCNN - Phát hiện khuôn mặt, **Bước 2**: Facenet - Trích xuất đặc trưng khuôn mặt, **Bước 3**: Giai đoạn truy vấn và xác minh định danh dùng **SVM** và **Siamese Network**.

5.3.1. So sánh với các nghiên cứu cùng nhóm

Bảng 2. So sánh các nghiên cứu tương tự trên VGGFace2.

STT	Công trình nghiên cứu	Accuracy
1	Vggface2: A dataset for recognising faces across pose and age [20]	98,2%
2	Deep Learning Based Real Time Face Recognition For University Attendance System [21]	95,3%
3	Phương pháp chúng tôi đề xuất MFSN	98,3%

5.3.2. Phân tích định lượng hành vi của mô hình

Trong nghiên cứu này, SVM không được xem là mô hình tiên tiến cạnh tranh với các phương pháp học sâu. Thay vào đó, nó được đưa vào như một mô hình cơ sở truyền thống thường được sử dụng trong các nhiệm vụ xác thực khuôn mặt nhị phân, trong đó mục tiêu là xác định xem hai hình ảnh khuôn mặt có thuộc cùng một danh tính hay không.

Việc so sánh giữa phân loại dựa trên SVM và các mô hình học Siamese/metric nhằm mục đích làm nổi bật khoảng cách hiệu suất giữa các quy trình hai giai đoạn thông thường (các đặc trưng được tạo thủ công hoặc trích xuất trước, sau đó là bộ phân loại) và các phương pháp học metric sâu từ đầu đến cuối. Sự so sánh như vậy cung cấp những hiểu biết thực tiễn về hiệu quả của học sâu khi thay thế các hệ thống xác thực truyền thống trong các triển khai thực tế.

Bên cạnh đó, chúng tôi có mở rộng thực nghiệm một mô hình nhận dạng khuôn mặt dựa trên học sâu hiện đại là ArcFace. Điều này cho phép so sánh trực tiếp giữa các phương pháp học metric sâu hiện đại, trong khi SVM vẫn là mô hình cơ sở tham chiếu chứ không phải là đối thủ cạnh tranh chính.

Bảng 3. So sánh định lượng giữa SVM và mạng Siamese

Chỉ số	SVM	Siamese Network
Problem formulation	Multi-class classification	Pairwise verification
Accuracy	0,47	0,983
Precision (macro avg)	0,47	0,983
Recall (macro avg)	0,47	0,983
F1-score (macro avg)	0,47	0,983
Precision (positive class)	0,47	0,93
Recall (positive class)	0,50	0,96
False Acceptance Rate (FAR)	High	Low
False Rejection Rate (FRR)	High	Low
Macro avg ≈ Weighted avg	Yes	Yes
Class imbalance sensitivity	Low	Low
Suitability for one-shot learning	Low	High
Scalability to new identities	Limited (retraining required)	High

Ngoài độ chính xác tổng thể, các chỉ số phân loại chi tiết hơn nữa giải thích khoảng cách hiệu suất quan sát được. Đối với phương pháp dựa trên SVM, độ chính xác, độ thu hồi và điểm F1 luôn ở mức khoảng 0,45 - 0,47, và các giá trị trung bình vĩ mô và có trọng số gần như giống hệt nhau. Điều này cho thấy hiệu suất thấp không phải do sự mất cân bằng lớp hoặc dự đoán thiên lệch, mà là do khả năng phân biệt hạn chế của SVM trong thiết lập một lần, với số lượng lớp lớn. Ngược lại, mạng Siamese đạt được độ chính xác và độ thu hồi cân bằng và cao (≈0,98), với tỷ lệ chấp nhận sai thấp hơn đáng kể, điều này rất quan trọng đối với các hệ thống điểm danh thực tế. Ma trận nhầm lẫn cũng xác nhận rằng mô hình Siamese duy trì sự phân tách rõ ràng giữa các cặp khớp và không khớp, trong khi SVM thể hiện sự chồng chéo đáng kể giữa các lớp. Những kết quả này chứng minh rằng sự khác biệt về hiệu suất phản ánh tính phù hợp vốn có của phương pháp học dựa trên xác minh đối với nhận dạng khuôn mặt với dữ liệu đăng ký tối thiểu, chứ không phải do cấu hình thử nghiệm không công bằng.

5.3.3. Thảo luận về hiệu suất

Bảng 4. Kết quả chỉ số so sánh 2 mô hình SVM và Siamese Network

Chỉ số	SVM (Support Vector Machine)	Siamese Network
Độ chính xác (Accuracy)	45,7% - 47%	98,3%
Dữ liệu huấn luyện	Cần nhiều ảnh mẫu	Chỉ cần 1 ảnh/người
Thời gian xử lý	1,2s/ảnh	0,5s/ảnh
Khả năng mở rộng	Phải huấn luyện lại khi thêm người	Không cần huấn luyện lại
Hiệu suất (dữ liệu ít)	Kém hơn	Ổn định, vượt trội

Để đánh giá hiệu quả của các phương pháp nhận diện khuôn mặt, chúng tôi tiến hành so sánh giữa mô hình sử dụng SVM và mô hình học sâu Siamese Network. Cả hai mô hình đều sử dụng MTCNN cho bước phát hiện khuôn mặt và FaceNet để trích xuất vector đặc trưng, do đó đảm bảo tính công bằng trong giai đoạn tiền xử lý và biểu diễn dữ liệu. Sự khác biệt chính giữa hai phương pháp nằm ở cách thức xây dựng bài toán và cơ chế ra quyết định ở bước cuối. Cụ thể, phương pháp dựa trên SVM tiếp cận bài toán theo hướng **phân loại đa lớp**, trong đó mỗi sinh viên được xem như một lớp riêng biệt và danh tính được xác định trực tiếp từ vector đặc trưng. Ngược lại, Siamese Network sử dụng kiến trúc mạng đôi để học **khoảng cách giữa các cặp khuôn mặt**, từ đó thực hiện nhiệm vụ **xác minh danh tính** thông qua các hàm mất mát như Contrastive Loss hoặc Triplet Loss.

Kết quả thực nghiệm trên tập dữ liệu VGGFace2 cho thấy Siamese Network đạt độ chính xác lên đến **98,3%**, vượt trội so với mức **45,7%–47%** của phương pháp sử dụng SVM. Khoảng cách hiệu năng lớn này chủ yếu xuất phát từ sự khác biệt trong cách xây dựng bài toán và mức độ phù hợp của mô hình đối với kịch

bản dữ liệu thực tế. Trong thiết lập đăng ký một lần (one-shot enrollment), mỗi danh tính chỉ có một hoặc rất ít mẫu huấn luyện, khiến SVM gặp khó khăn trong việc học các ranh giới quyết định đáng tin cậy trong không gian embedding có số chiều cao. Điều này dẫn đến hiện tượng chồng chéo giữa các lớp và làm suy giảm đáng kể hiệu suất phân loại. Ngược lại, Siamese Network xây dựng bài toán nhận diện khuôn mặt như một nhiệm vụ xác minh, vốn được thiết kế đặc biệt cho các kịch bản học một lần hoặc rất ít mẫu. Hơn nữa, embedding của FaceNet được tối ưu hóa bằng hàm mất mát bộ ba nhằm phục vụ việc so sánh độ tương đồng giữa các khuôn mặt, thay vì phân tách đa lớp. Do đó, mô hình Siamese có thể khai thác hiệu quả hơn cấu trúc hình học của không gian embedding và đạt được độ chính xác cao hơn đáng kể. Về mặt hiệu năng tính toán, Siamese Network cũng cho thấy ưu thế khi chỉ cần tính toán khoảng cách giữa các vector đặc trưng mà không yêu cầu huấn luyện lại bộ phân loại khi bổ sung danh tính mới. Điều này giúp giảm độ phức tạp triển khai, tăng khả năng mở rộng và phù hợp với các ứng dụng xác minh danh tính theo thời gian thực. Ngược lại, phương pháp sử dụng SVM yêu cầu tái huấn luyện mô hình khi số lượng người dùng thay đổi, đồng thời tiêu tốn nhiều tài nguyên hơn nhưng không mang lại hiệu quả phân loại tương ứng.

Bảng 5. So sánh ArcFace (Siamese) và FaceNet (Siamese)

Chỉ số	ArcFace (Siamese / Triplet Loss)	FaceNet (Siamese / Triplet Loss)
Problem formulation	Multi-class classification	Pairwise / metric-based verification
Embedding separability	Trung bình – còn chồng lấn	Cao – tách biệt rõ
Triplet Accuracy	≈ 0,85	≈ 0,98
Accuracy (verification)	≈ 0,85	≈ 0,98
Precision (macro avg)	≈ 0,84 – 0,86	≈ 0,98
Recall (macro avg)	≈ 0,83 – 0,86	≈ 0,98
F1-score (macro avg)	≈ 0,84 – 0,86	≈ 0,98
Precision (positive class)	≈ 0,85	≈ 0,93
Recall (positive class)	≈ 0,86	≈ 0,96
False Acceptance Rate (FAR)	Trung bình	Thấp
False Rejection Rate (FRR)	Trung bình	Thấp
Macro avg ≈ Weighted avg	Có	Có
Class imbalance sensitivity	Trung bình	Thấp
Suitability for one-shot learning	Thấp	Cao
Scalability to new identities	Hạn chế (cần retrain)	Cao
Deployment in open-set systems	Không tối ưu	Phù hợp

Bảng 5 trình bày so sánh giữa ArcFace kết hợp Softmax và FaceNet trong bối cảnh bài toán nhận diện mở. Cần lưu ý rằng ArcFace được thiết kế chủ yếu cho bài toán phân lớp đa lớp (closed-set classification), do đó các chỉ số hiệu năng của ArcFace trong bảng được suy diễn gián tiếp từ phân bố khoảng cách embedding và độ chính xác triplet, thay vì từ đánh giá phân lớp trực tiếp.

Kết quả cho thấy mặc dù ArcFace đạt hiệu năng cao trên các benchmark chuẩn, không gian embedding sinh ra trong điều kiện dữ liệu thực tế vẫn tồn tại sự chồng lấn giữa các cặp cùng danh tính và khác danh tính, dẫn đến tỷ lệ chấp nhận sai (FAR) và từ chối sai (FRR) ở mức trung bình. Ngược lại, FaceNet được tối ưu trực tiếp thông qua Triplet Loss cho bài toán xác thực cặp, tạo ra không gian embedding có tính phân tách rõ ràng hơn, giúp cải thiện độ ổn định của các chỉ số precision, recall và F1-score trong các kịch bản one/few-shot và open-set.

Do đó, bảng so sánh không nhằm khẳng định ưu thế tuyệt đối của một mô hình so với mô hình khác, mà nhằm làm rõ sự phù hợp của từng phương pháp đối với mục tiêu bài toán, trong đó FaceNet thể hiện lợi thế rõ ràng hơn cho các ứng dụng nhận diện mở và mở rộng danh tính mà không cần huấn luyện lại mô hình.

Từ các kết quả so sánh trên, có thể khẳng định rằng **Siamese Network là lựa chọn phù hợp hơn cho**

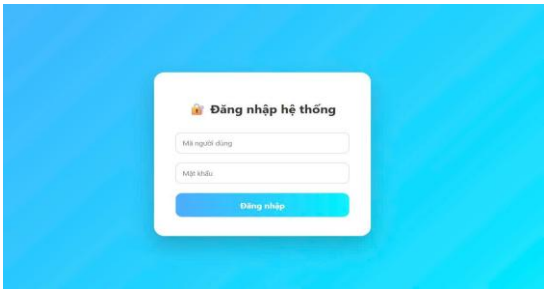
các hệ thống nhận diện khuôn mặt trong thực tế, đặc biệt đối với các bài toán điểm danh sinh viên với các ràng buộc như dữ liệu huấn luyện hạn chế, yêu cầu mở rộng linh hoạt và độ chính xác cao. Vì những lý do này, chúng tôi lựa chọn mô hình Siamese Network cho quá trình triển khai và đánh giá hệ thống điểm danh trong các thí nghiệm tiếp theo.

5.4. Triển khai thực tế mở rộng mô hình

Sau khi thực nghiệm mô hình trên dataset VGGFace2 chúng tôi chọn triển khai thực tế mở rộng mô hình trên ứng dụng thế tế điểm danh sinh viên. Chi tiết theo mô tả sau.

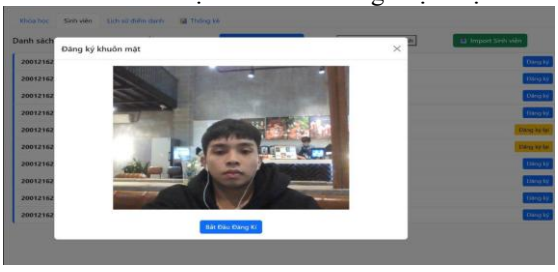
5.4.1. Dashboard Web

Giao diện đăng nhập hệ thống nhận diện khuôn mặt cho phép người dùng nhập thông tin tài khoản để truy cập hệ thống.



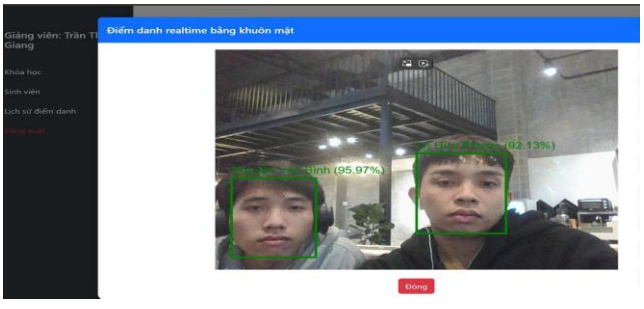
Hình 2. Đăng nhập.

Giao diện đăng ký khuôn mặt cho phép sinh viên chụp và lưu hình ảnh khuôn mặt để sử dụng trong hệ thống điểm danh tự động. Khi nhấn nút "Bắt đầu đăng ký", hệ thống sẽ ghi nhận khuôn mặt và liên kết với mã sinh viên, giúp xác thực danh tính và hỗ trợ điểm danh bằng nhận diện khuôn mặt trong các buổi học.



Hình 3. Đăng ký khuôn mặt.

Hệ thống điểm danh tự động tiên tiến, sử dụng công nghệ nhận diện khuôn mặt theo thời gian thực để điểm danh sinh viên.



Hình 4. Điểm danh real-time.

5.4.2. Mobile app

Hình 5. Hiển thị kết quả điểm danh.

6. KẾT LUẬN VÀ KHUYẾN NGHỊ

Trong nghiên cứu này, chúng tôi đã triển khai và đánh giá một hệ thống chứng thực sinh trắc học khuôn mặt sử dụng các kỹ thuật hiện đại, bao gồm MTCNN cho phát hiện khuôn mặt, FaceNet cho trích xuất đặc trưng và SVM cho phân loại. Tuy nhiên, sau khi huấn luyện mô hình trên tập dữ liệu VGGFace2, kết quả cho thấy độ chính xác của hệ thống sử dụng SVM chỉ đạt 45,7% trên tập kiểm thử với 500 danh tính và 169.396 hình ảnh. Điều này phản ánh hạn chế của SVM trong việc phân loại với số lượng lớn danh tính và dữ liệu có tính biến thiên cao. Ngược lại, khi áp dụng kiến trúc Siamese FaceNet – sử dụng mạng đôi để học khoảng cách giữa các cặp khuôn mặt – mô hình đạt độ chính xác lên đến 98,3%, cùng với F1-score ấn tượng, cho thấy hiệu quả vượt trội trong bài toán xác minh khuôn mặt. Nhờ khả năng học độ tương đồng thay vì phân lớp trực tiếp, Siamese Network chứng minh được độ tổng quát cao hơn, đặc biệt phù hợp với các môi trường thực tế phức tạp. Những kết quả này cho thấy rằng mô hình Siamese FaceNet là lựa chọn phù hợp hơn trong các hệ thống nhận diện khuôn mặt hiện đại, đặc biệt là khi cần xử lý tập dữ liệu lớn và đa dạng. SVM có thể vẫn hữu ích trong các bài toán phân lớp đơn giản, nhưng không đáp ứng tốt yêu cầu về độ chính xác và khả năng mở rộng như các mô hình học sâu hiện nay.

Mặc dù hệ thống đạt được nhiều kết quả khả quan, vẫn còn tồn tại một số hạn chế và cần được cải tiến trong tương lai: Hạn chế trong điều kiện môi trường khắc nghiệt: Hệ thống vẫn gặp khó khăn trong việc nhận diện khuôn mặt khi có sự thay đổi mạnh về ánh sáng, hoặc khi khuôn mặt bị che khuất (ví dụ, do khẩu trang, tóc hoặc các vật thể che khuất một phần khuôn mặt). Mặc dù MTCNN khá mạnh mẽ trong việc phát hiện khuôn mặt trong điều kiện ánh sáng thay đổi, nhưng khả năng nhận diện chính xác vẫn bị ảnh hưởng khi khuôn mặt bị che khuất hoặc trong môi trường tối. Tốc độ xử lý: Mặc dù hệ thống có thể hoạt động hiệu quả trong môi trường thời gian thực, nhưng với các mô hình có độ phức tạp cao, tốc độ xử lý có thể bị ảnh hưởng. Để có thể sử dụng trong các ứng dụng thực tế, cần tối ưu hóa các bước xử lý như trích xuất đặc trưng và phân loại. Tính chính xác trong các tình huống đông đúc: Khi có nhiều người trong cùng một khung hình hoặc trong các môi trường đông đúc, hệ thống có thể gặp khó khăn trong việc phân biệt và nhận diện chính xác từng khuôn mặt. Các phương pháp như phát hiện khuôn mặt theo vùng hay phân tích khuôn mặt đa lớp có thể cải thiện hiệu quả trong các trường hợp này.

Trong tương lai: Cần phát triển các thuật toán mạnh mẽ hơn để nhận diện khuôn mặt trong các điều kiện môi trường khắc nghiệt, bao gồm các phương pháp học sâu có khả năng nhận diện khuôn mặt trong các tình huống che khuất, bóng tối hoặc ánh sáng yếu. Tối ưu hóa tốc độ và hiệu suất cần tối ưu hóa tốc độ xử lý và giảm thiểu thời gian phản hồi của hệ thống để có thể sử dụng cho các ứng dụng đòi hỏi thời gian thực, như giám sát an ninh hoặc thanh toán qua nhận diện khuôn mặt. Ứng dụng các phương pháp học sâu mới như các kiến trúc mạng nơ-ron tích chập (CNN) cải tiến, học không giám sát hoặc học chuyển giao có thể được áp dụng để cải thiện độ chính xác và khả năng nhận diện khuôn mặt trong môi trường khó khăn hơn.

Lời cảm ơn: Nghiên cứu này do Trường Đại học Công Thương Thành phố Hồ Chí Minh bảo trợ và cấp kinh phí theo Hợp đồng số 30/HĐ-DCT ngày 17 tháng 01 năm 2025.

TÀI LIỆU THAM KHẢO

[1] M. Turk and A. Pentland, "Eigenfaces for recognition," J. Cogn. Neurosci., vol. 3, no. 1, pp. 71–86, Jan. 1991, doi: 10.1162/jocn.1991.3.1.71.

[2] R. Min, N. Kose, and J.-L. Dugelay, "KinectFaceDB: A Kinect database for face recognition," IEEE Trans. Syst. Man Cybern. Syst., vol. 44, no. 11, pp. 1534–1548, Oct. 2014, doi: 10.1109/TSMC.2014.2331215.

[3] E. N. Mortensen, H. Deng, and L. Shapiro, "A SIFT descriptor with global context," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Jun. 2005, pp. 184–190 vol. 1. doi: 10.1109/CVPR.2005.45.

[4] G. Zhang, X. Huang, S. Z. Li, Y. Wang, and X. Wu, "Boosting Local Binary Pattern (LBP)-based face recognition," in Advances in Biometric Person Authentication, vol. 3338, S. Z. Li, J. Lai, T. Tan, G. Feng, and Y. Wang, Eds, in Lecture Notes in Computer Science, vol. 3338. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 179–186. doi: 10.1007/978-3-540-30548-4_21.

[5] A. Tzotsos and D. Argialas, "Support vector machine classification for object-based image analysis," 2008. doi: 10.1007/978-3-540-77058-9_36.

[6] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "CVPR 2014 Open Access Repository," 2014. doi: https://openaccess.thecvf.com/content_cvpr_2014/html/Taigman_DeepFace_Closing_the_2014_CVPR_paper.html.

[7] F. Schroff, D. Kalenichenko, and J. Philbin, "CVPR 2015 Open Access Repository," 2015. doi: https://www.cvfoundation.org/openaccess/content_cvpr_2015/html/Schroff_FaceNet_A_Unified_2015_CVPR_paper.html.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "CVPR 2019 Open Access Repository," 2019. doi: https://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html.

[9] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," Sep. 2016. doi: 10.1007/978-3-319-46478-7_31.

[10] "Orthogonality loss: Learning discriminative representations for face recognition," IEEE Access, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9184823.

[11] "ArcFace: Additive angular margin loss for deep face recognition," ResearchGate, 2019. [Online]. Available: https://www.researchgate.net/publication/338506499_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition.

[12] S. Wang and Y. Chen, "A joint loss function for deep face recognition," Multidimens. Syst. Signal Process., vol. 30, Jul. 2019, doi: 10.1007/s11045-018-0614-0.

[13] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017. [Online]. Available: https://www.researchgate.net/publication/316184205_MobileNets_Efficient_Convolutional_Neural_Networks_for_Mobile_Vision_Applications.

[14] A. George, C. Ecabert, H. Otroshi, K. Kotwal, and S. Marcel, "EdgeFace: Efficient face recognition model for edge devices," IEEE Trans. Biom. Behav. Identity Sci., vol. PP, pp. 1–1, Apr. 2024, doi: 10.1109/TBIOM.2024.3352164.

[15] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," ACM Comput. Surv., vol. 35, no. 4, pp. 399–458, Dec. 2003. [Online]. Available: https://dl.acm.org/doi/10.1145/954339.954342.

[16] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," Oxford University Research Archive, 2015. [Online]. Available: https://ora.ox.ac.uk/objects/uuid:a5f2e93f-2768-45bb-8508-74747f85cad1.

[17] N. Zhang, J. Luo, and W. Gao, "Research on face detection technology based on MTCNN," in 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA), Sep. 2020, pp. 154–158. doi: 10.1109/ICCNEA50255.2020.00040.

[18] E. Jose, G. M., M. T. P. Haridas, and M. H. Supriya, "Face recognition based surveillance system using FaceNet and MTCNN on Jetson TX2," in 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Mar. 2019, pp. 608–613. doi: 10.1109/ICACCS.2019.8728466.

[19] E. Solomon, A. Woubie, and E. S. Emiru, "Self-supervised deep learning based end-to-end face verification method using Siamese network," in 2023 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), IEEE, 2023, pp. 1–6.

[20] Q. Cao, L. Shen, W. Xie, O. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," 2018, pp. 67–74. doi: 10.1109/FG.2018.00020.

[21] M. Singhal and G. Ahmad, "Deep learning based real time face recognition for university attendance system," in 2023 International Symposium on Devices, Circuits and Systems (ISDCS), May 2023, pp. 01–04. doi: 10.1109/ISDCS58735.2023.10153549.

ABSTRACT

RESEARCH AND DEVELOPMENT OF A FACE RECOGNITION MODEL FOR STUDENT ROLL CALL APPLICATION

Bui Cong Danh, Pham Nguyen Huy Phuong*

Ho Chi Minh City University of Industry and Trade

*Email: *phuongpnh@huit.edu.vn*

Face recognition is one of the important research directions of computer vision, aiming to automatically identify or verify personal identity based on biometric features from facial images or videos. This paper presents a face recognition system applied to student attendance in the classroom. In this paper, we propose a MFSN system with three main stages including MTCNN detection, feature extraction using FaceNet and identity verification using one-shot learning mechanism using Siamese Network. Experimental results show that our proposed model achieves 98.3% accuracy on VGGFace2 and achieves high performance on a real dataset of HUIT students with only one registered image for each student. The system is fully deployed on a low-end personal computer, operating independently in a local area network (LAN) environment without requiring an Internet connection or cloud services. The system can take real-time attendance for a class of 50 students in an average time of 8 seconds

Keywords: Face recognition, One-shot learning, Siamese network, Student attendance.