THIẾT KẾ MÔ HÌNH NHẬN DIỆN KHUÔN MẶT DỰA TRÊN MẠNG NƠ RON TÍCH CHẬP ĐA TẦNG VÀ ƯỚC TÍNH TƯ THẾ ĐẦU NGƯỜI

DESIGN OF A FACE RECOGNITION TECHNIQUE BASED MTCNN AND HEAD POSE ESTIMATION

Ngô Việt Đức[1], Đặng Thái Việt[1,*], Vũ Ngọc Hải[1], Nguyễn Như Trường[1]

DOI: https://doi.org/10.57001/huih5804.2024.023

TÓM TẮT

Trí tuệ nhân tạo và Internet kết nối vạn vật thu hút nhiều sự quan tâm của những học giả và nhà nghiên cứu, không chỉ bởi tính ứng dụng cao mà còn là những công nghệ tiêu biểu của Cách mạng công nghiệp lần thứ tư. Điểm nổi bật của trí tuệ nhân tạo là khả năng tự học, cho phép máy tính dự đoán và phân tích dữ liệu phức tạp như dấu vân tay, mống mắt và khuôn mặt. Nghiên cứu này đề xuất giải pháp hệ thống cơ điện tử tích hợp khả năng nhận dạng khuôn mặt của AI để tạo ra hệ thống điểm danh và đánh giá sự chuyên cần của học sinh. Độ chính xác của mô hình dao động từ 90% đến 95%. Kết quả của nghiên cứu được so sánh với nghiên cứu gần đây chứng minh khả năng sử dụng của hệ thống. Vì vậy, nhóm tác giả vận dụng kết quả của quá trình đào tạo để xây dựng hệ thống đánh giá chuyên cần và chuyên cần nhận diện khuôn mặt học viên.

Từ khóa: *Nhận dạng khuôn mặt, ước tính tư thế đầu người, thị giác máy tính, mạng tích chập xếp tầng đa tác vụ.*

ABSTRACT

Artificial Intelligence and IoT have always attracted a lot of attention from scholars and researchers, not only because of their high applicability but also typical technologies of the Fourth Industrial Revolution. The hallmark of AI is its self-learning ability, which enables computers to predict and analyze complex data such as fingerprints, irises, and faces. The study proposes a solution for a mechatronic system that integrates AI's face recognition capabilities to create an attendance system and assess student attendance. The model's accuracy ranges from 90% to 95%. The study's results are compared with recent research demonstrating the system's usability. Therefore, the authors apply the training process's outcomes to construct an attendance and diligence assessment system that recognizes students' faces.

Keywords: *Face recognition, head pose estimation, computer vision, MTCNN.*

[1]Đại học Bách khoa Hà Nội

[*]Email: viet.dangthai@hust.edu.vn

Ngày nhận bài: 10/6/2023

Ngày nhận bài sửa sau phản biện: 12/9/2023

Ngày chấp nhận đăng: 20/01/2024

DANH MỤC KÝ HIỆU

MTCNN: Mạng tích chập xếp tầng đa tác vụ

DCNN: Mạng tích chập học sâu

CNN: Mạng tích chập

FPS: Số khung hình/giây

AI: Trí tuệ nhân tạo

IoT: Internet kết nối vạn vật

1. GIỚI THIỆU

Trong kỷ nguyên Công nghiệp 4.0, với sự phát triển nhanh chóng của các thiết bị phần cứng cung cấp khả năng tính toán cho các mạng trí tuệ nhân tạo. Các thành tựu khoa học kỹ thuật đã tạo sự phát triển của các kỹ thuật cơ bản trong trí tuệ nhân tạo [1, 2]. Thị giác máy tính với ứng dụng rộng rãivào trong các thiết bị, hệ thống thông minh. Những công nghệ hiện nay đóng một vai trò ngày càng quan trọng trong việc hỗ trợ các hệ thống sử dụng các phương pháp dựa trên tầm nhìn, chẳng hạn như nhận dạng và phát hiện đối tượng, xử lý hình ảnh và trích xuất thông tin [3, 4]. Một trong các ứng dụng quan trọng của thị giác máy tính phát triển phần mềm chấm công và đánh giá hiệu suất trong sản xuất công nghiệp cũng như trong đời sống.

Mặc dù các hệ thống thẻ dựa trên RFID mang lại sự tiện lợi và chính xác, nhưng chúng dễ bị các hoạt động gian lận trong môi trường giáo dục, nơi học sinh có thể dựa vào bạn bè của mình để tham dự [4, 5]. Tương tự, các phương pháp nhận dạng mống mắt cung cấp tính bảo mật cao nhưng liên quan đến chi phí triển khai cao và gây lo ngại về quyền riêng tư [4, 5]. Do đó, chúng tôi đề xuất sử dụng nhận dạng khuôn mặt như một phương pháp mang lại độ chính xác và bảo mật cao. Để vận hành một hệ thống nhận dạng đáng tin cậy, cần phải đáp ứng một số tiêu chí nhất định, chẳng hạn như độ chính xác, thời gian phản hồi và độ bền. Để đáp ứng các yêu cầu này, việc chọn một mô hình nhận dạng khuôn mặt phù hợp là rất quan trọng [6, 7]. Hiện tại, có một số mô hình phổ biến có sẵn, chẳng hạn như ArcFace, DeepFace và FaceNet [6-9]. Mặc dù ArcFace và DeepFace cung cấp độ chính xác cao, nhưng chúng yêu cầu tài nguyên tính toán lớn hơn so với FaceNet, khiến chúng không phù hợp để triển

khai quy mô lớn. Do đó, chúng tôi chọn sử dụng mô hình FaceNet.

Để tạo điều kiện lấy mẫu hiệu quả, nhóm nghiên cứu kết hợp thuật toán xác định góc xoay khuôn mặt để chụp ảnh khuôn mặt từ nhiều góc nhìn. Sau đó, kết hợp quá trình ghi nhãn để công nhận các khuôn mặt trên máy tính Jetson Nano nhúng. Cuối cùng, kết quả đánh giá được gửi đến trang web đã phát triển để giảng viên và giám sát viên truy cập. Việc đưa vào thuật toán xoay khuôn mặt để lấy mẫu hình ảnh đã làm tăng đáng kể độ chính xác của phương pháp lên khoảng 90 - 95%.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Mô hình mạng tích chấp xếp tầng đa tác vụ (MTCNN)

MTCNN (Multi-tasked Cascaded Convolutional Networks) là một lựa chọn phổ biến để sử dụng trong Deep Convolutional Neural Networks (DCNN) để phát hiện khuôn mặt và nhận dạng đặc điểm khuôn mặt có liên quan vì những lý do sau:

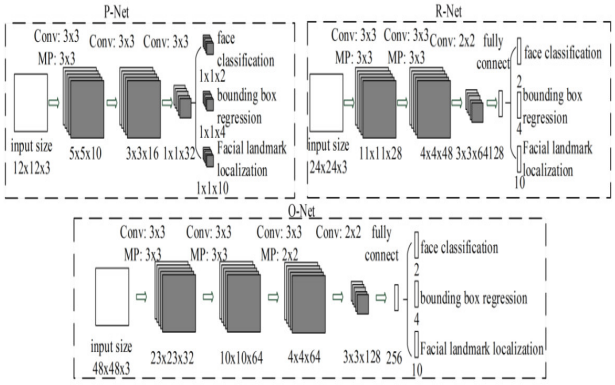
1) Đa tác vụ: MTCNN là một hệ thống đa tác vụ có khả năng thực hiện đồng thời nhiều tác vụ như nhận diện khuôn mặt, định vị các mốc trên khuôn mặt (mắt, mũi, miệng) và cung cấp các hộp giới hạn. Điều này cho phép MTCNN cung cấp thông tin chi tiết về khuôn mặt và các thành phần của nó, tạo điều kiện thuận lợi cho các tác vụ như nhận dạng khuôn mặt, phân loại và phân tích biểu thức.

2) Độ tin cậy cao: MTCNN sử dụng mạng nơ ron tích chập (CNN) để nhận diện khuôn mặt và bản địa hóa tính năng. CNN là một kiến trúc mạnh mẽ để xử lý dữ liệu hình ảnh, cho phép MTCNN đạt được độ tin cậy cao trong việc phát hiện và mạnh mẽ chống lại nhiễu, thay đổi tỷ lệ và thay đổi ánh sáng.

3) Hiệu quả cao: MTCNN được thiết kế để hoạt động nhanh chóng và hiệu quả. Với các giai đoạn phát hiện và bản địa hóa tính năng tuần tự, sử dụng các kỹ thuật như mạng tích chập và hội tụ nhận biết cục bộ, MTCNN có thể xử lý hình ảnh và video có độ phức tạp khác nhau một cách nhanh chóng.

4) Ứng dụng đa dạng: MTCNN có thể được ứng dụng trong nhiều lĩnh vực khác nhau, bao gồm nhận dạng khuôn mặt, phân loại tuổi và giới tính, phân tích nét mặt và hệ thống giám sát an ninh. Việc hợp nhất MTCNN vào DCNN mở ra cơ hội nghiên cứu và phát triển các ứng dụng trí tuệ nhân tạo liên quan đến xử lý ảnh và nhận dạng khuôn mặt.

Mô hình MTCNN bao gồm ba giai đoạn liên tiếp: phát hiện khuôn mặt, định vị mốc khuôn mặt và hồi quy hộp giới hạn chính xác. Các giai đoạn này được thiết kế để phát hiện và định vị các khuôn mặt trong hình ảnh với các mức độ phức tạp khác nhau, bao gồm các biến thể về tỷ lệ, tư thế và ánh sáng. MTCNN đã thể hiện khả năng mạnh mẽ trong việc phát hiện khuôn mặt và nhận dạng đặc điểm khuôn mặt tương ứng trong hình ảnh và video. Nó được sử dụng rộng rãi trong các ứng dụng như nhận dạng khuôn mặt, phân loại tuổi và giới tính, nhận dạng nét mặt và trong các hệ thống giám sát an ninh.



Hình 1. Cấu trúc mạng MTCNN bao gồm 3 lớp P-Net, R-Net và O-Net

2.2. Mô hình mạng Facenet

FaceNet là một mạng nơ ron học sâu được sử dụng để trích xuất các đặc điểm từ hình ảnh khuôn mặt của một người. Facenet [4-6] được công bố lần đầu vào năm 2015 bởi các nhà nghiên cứu của Google Schroff et al. FaceNet lấy đầu vào là hình ảnh khuôn mặt và xuất ra một vectơ nhúng 128 chiều chứa thông tin quan trọng về khuôn mặt. Vectơ này có thể được hiểu là một điểm trong hệ tọa độ 128 chiều, biểu diễn các đặc điểm của mặt đầu vào. Với điểm đặc trưng của vectơ được nhúng, chúng ta có thể gắn nhãn hình ảnh đầu vào khi chúng đủ gần với các phần nhúng của một người đã biết. Quá trình đào tạo của FaceNet tạo ra các vectơ có giá trị ngẫu nhiên cho tập hợp hình ảnh đầu vào.

Mô hình mạng FaceNet học theo các bước sau:

- Bước 1. Chọn ngẫu nhiên một ảnh làm ảnh đánh dấu (anchor image).

- Bước 2. Chọn ngẫu nhiên một hình ảnh của cùng một người làm hình ảnh đánh dấu làm hình ảnh tích cực (positive image).

- Bước 3. Chọn ngẫu nhiên một hình ảnh của một người khác với hình ảnh đánh dấu làm hình ảnh tiêu cực (negative image).

- Bước 4. Điều chỉnh các thông số của mạng Facenet sao cho hình ảnh tích cực gần với hình ảnh đánh dấu hơn hình ảnh tiêu cực.

$$\left\|f\left(x_i^a\right)-f\left(x_i^p\right)\right\|_2^2+\alpha<\left\|f\left(x_i^a\right)-f\left(x_i^n\right)\right\|_2^2, \quad \forall f\left(x_i^a\right), f\left(x_i^p\right), f\left(x_i^n\right)\in T \tag{1}$$

Trong đó tham số α được sử dụng để đảm bảo khoảng cách giữa các lần nhúng của các mẫu khác nhau là đủ xa. Điều này giúp tăng độ chính xác của hệ thống phân loại và giảm sự phụ thuộc vào các tính năng không cần thiết. Khi đó, hàm mất mát được biểu diễn như sau:

$$L=\sum_i^N\left[\left\|f\left(x_i^a\right)-f\left(x_i^p\right)\right\|_2^2-\left\|f\left(x_i^a\right)-f\left(x_i^n\right)\right\|_2^2\right]+\alpha \tag{2}$$

Các tác giả tiếp tục quá trình cho đến khi tất cả hình ảnh của cùng khuôn mặt người ở gần nhau và hình ảnh của những khuôn mặt người khác nhau cách xa nhau. Để đạt hiệu quả huấn luyện cao, ta cần chọn bộ ba trong đó ảnh

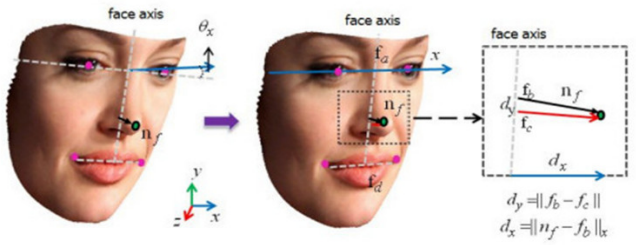
đánh dấu và ảnh tích cực càng xa nhau càng tốt, ảnh đánh dấu và ảnh tiêu cực càng gần nhau càng tốt (trong công thức 3).

$\left\|f\left(x_i^a\right)-f\left(x_i^p\right)\right\|_2^2$ là tối đa có thể $\left\|f\left(x_i^a\right)-f\left(x_i^n\right)\right\|_2^2$ và tối thiểu có thể (3)

Với các tập dữ liệu lớn, việc chọn các điểm tích cực và tiêu cực thích hợp từ toàn bộ tập dữ liệu có thể tốn kém về mặt tính toán, vì vậy chúng tôi có thể chia tập dữ liệu thành các lô nhỏ hơn và chọn các điểm tích cực và tiêu cực phù hợp từ mỗi lô nhỏ hơn.

2.3. Dự đoán tư thế đầu người

Tư thế đầu đề cập đến hướng hoặc vị trí đầu của một người so với hệ quy chiếu. Nó thường mô tả chuyển động quay của đầu dọc theo ba trục: cao độ, ngáp và lăn (xoay quanh x, y, z). Ba bậc tự do (DOF) có thể được tính toán bằng cách sử dụng hệ tọa độ cầu để ước tính các pháp tuyến bề mặt n_f. Dựa trên một tập hợp các tỷ lệ khoảng cách cố định từ nhiều cá nhân, góc nghiêng được tính dựa trên bất kỳ hướng xem nào. Trong phương pháp này, chiều dài được chiếu của các quy tắc bề mặt n_f được sử dụng trong cả hai trục tọa độ. Ngoài ra, các tác giả cố định hướng nhìn của mình và không sử dụng hệ tọa độ hình cầu.



Hình 2. Minh họa về việc khử xoay để tính toán bề mặt bình thường một cách tin cậy từ một hình ảnh khuôn mặt

Ước tính vị trí đầu bằng phương pháp 4 điểm (f_a, f_b, f_c, f_d) trong hình 2. Đầu tiên, f_a và f_d đại diện cho hai điểm chính giữa hai mắt và hai khóe miệng. Sau đó f_b là điểm nằm trên trục vuông góc với pháp tuyến mặt phẳng. Cuối cùng, f_c có thể được tính bằng các phương trình sau trong khi t được chọn trong phạm vi 0,45.

$$f_c = (1-t) * f_a + t * f_d \quad (4)$$

Giá trị n_f trong [7] tính dựa trên khoảng cách tương đối theo công thức sau:

$$n_f = \frac{\|f_b - f_d\|}{\|f_a - f_d\|} \quad (5)$$

Tỷ lệ chiều dài mũi tương đối chung có thể được tìm thấy theo phương trình (6):

$$l_n = \frac{L_n}{\lfloor f_a' - f_d' \rfloor} \quad (6)$$

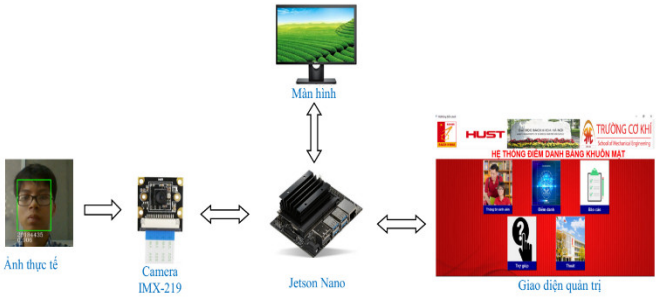
Các góc quay quanh trục z và y được tính bằng cách sử dụng khoảng cách tương đối quan sát được chia cho chiều cao mũi tối đa l_n:

$$\theta_y = \arcsin\left(\frac{d_x'}{l_n}\right), \text{ và } \theta_z = \arcsin\left(\frac{d_y'}{l_n}\right) \quad (7)$$

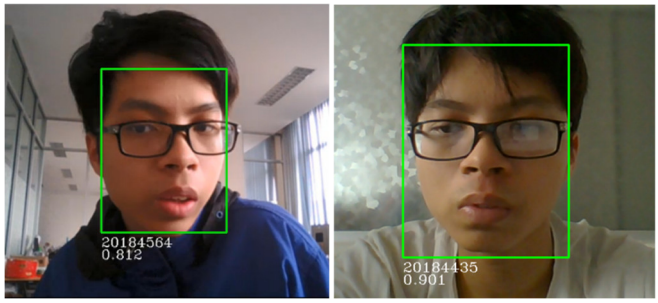
Dựa theo hình 2 và các phương trình từ (2) đến (7) chúng ta có thể biểu diễn các góc quay của khuôn mặt người theo các tư thế đầu đặc trưng. Điều này có thể hỗ trợ cho kỹ thuật nhận dạng khuôn mặt, thay vì độ chính xác cao ở góc nhìn trực diện, thì độ chính xác được tăng cao với các góc quay khuôn mặt tương ứng với các tư thế của đầu.

3. KẾT QUẢ VÀ THẢO LUẬN

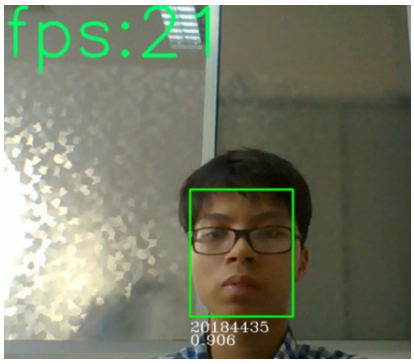
Trong hình 3, mô hình hệ thống điểm danh học sinh sử dụng nhận dạng khuôn mặt sẽ chạy trên máy tính nhúng Jetson Nano 8GB. Hệ thống giám sát có thể trực tuyến thông qua điện thoại hoặc website. Kết quả điểm danh sẽ được lưu dưới dạng file Excel để tạo báo cáo cho giáo viên. Khối xử lý trung tâm gồm: Jetson Nano 4Gb. Dữ liệu đầu vào ảnh từ camera, tín hiệu từ mạch điều khiển sẽ truyền đến Jetson Nano để xử lí và sau đó có thể kết nối đến các thiết bị chấp hành.



Hình 3. Sơ đồ cấu trúc hệ thống hệ điểm danh IoT



(a) (b)



(c)

Hình 4. Chất lượng của nhận dạng khuôn mặt với phương pháp lấy mẫu khuôn mặt: (a) chưa có kết hợp Head pose, (b) và (c) có kết hợp Head pose

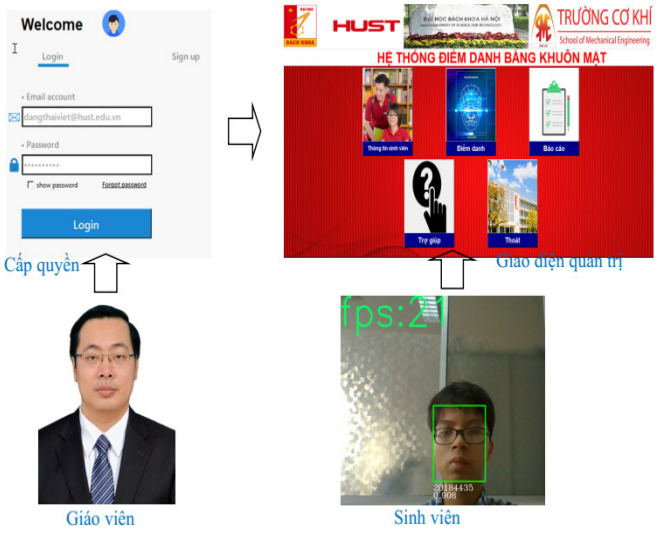
So với kết quả thu được khi sử dụng kỹ thuật Haar Cascades để nhận diện khuôn mặt và chụp ảnh ngay lập tức, bộ ảnh thu được sẽ chỉ bao gồm một đến hai góc khác nhau của khuôn mặt. Tuy nhiên, độ chính xác của việc lấy mẫu với hình ảnh từ nhiều góc độ mang lại độ chính xác cao hơn. Điều này là do phương pháp mới giúp quá trình đào tạo tạo ra các vectơ có nhiều đặc điểm khuôn mặt hơn. Do đó, trong các tình huống nhận dạng khuôn mặt trong thế giới thực, nếu hình ảnh đầu vào khác với hình ảnh mẫu (do góc, ánh sáng,...), khả năng tìm thấy các vectơ có nhiều tính năng tương đương hơn sẽ tăng lên. Trong hình 4, nhóm tác giả đã kết hợp đầu ra của mạng MTCNN kết hợp với kỹ thuật dự đoán tư thế đầu người (Head pose estimation) để đảm bảo độ chính xác và chất lượng trong nhận diện khuôn mặt.

Phương pháp lấy mẫu mới tăng độ chính xác trong khi vẫn đủ hiệu quả về tài nguyên để Jetson Nano xử lý. Trong hình 3(c), khi chạy với bộ dự liệu 40 khuôn mặt thì FPS duy trì ổn định trong khoảng 21 đến 25. Các kết quả so sánh được thể hiện trong bảng 1 đã chỉ ra độ chính xác được cải thiện và tốc độ xử lý đảm bảo yêu cầu trong bài toán điểm danh dựa vào nhận dạng khuôn mặt.

Bảng 1. Bảng so sánh giữa các phương pháp nhận dạng theo độ chính xác và số khung hình/giây (FPS)

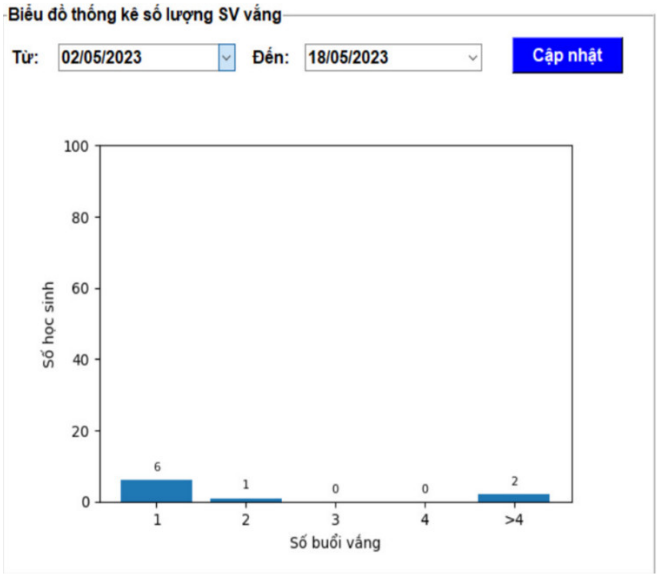
Mô hình	Độ chính xác	Số khung hình/giây (FPS)
LBP [4-6]	84 - 88	19 - 21
DLIB [4-6]	52 - 59	9 - 10
MTCNN+Facenet + Head Pose	90 - 95	21 - 25

Dựa trên kết quả thu được của hệ thống nhận dạng khuôn mặt, một giao diện quản lý độ chuyên cần sinh viên được xây dựng, đảm bảo kiểm tra tần suất số lần vắng/nghỉ và có khả năng tra cứu trực tuyến và xuất file lưu trữ dạng Excelt (hình 5).

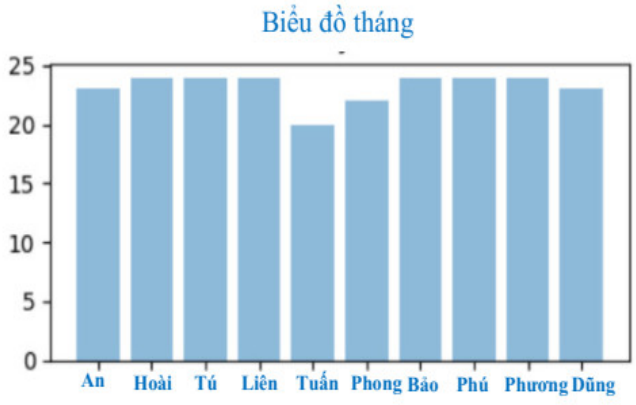


Hình 5. Giao diện quản trị hệ điểm danh IoT sinh viên

Biểu đồ thể hiện tần suất điểm danh sinh viên được thể hiện như hình 6, kết quả cũng có thể được xuất file lưu trữ dạng Excel.



(a)



(b)

Hình 6. Kết quả điểm danh sinh viên: (a) Trạng thái thực tế tại lần điểm danh và (b) Tổng kết cuối kỳ

4. KẾT LUẬN

Kỹ thuật nhận diện khuôn mặt bằng kỹ thuật Head pose kết hợp với mô hình Facenet làm tăng hiệu quả nhận diện cho thuật toán và giúp ngăn chặn một số phương thức giả danh hiện nay. Độ chính xác đạt 90 - 95% với tốc độ xử lý hình ảnh 21-25 FPS là kết quả rất khả quan so với các phương pháp nhận diện trực quan trong các công trình công bố trước đây. Kỹ thuật đã được tích hợp vào hệ thống IoT nhằm nâng cao tính bảo mật đảm bảo quá trình điểm danh sinh viên Đại học Bách khoa Hà Nội. Mở rộng, kết quả phần nhận diện khuôn mặt có thể kết hợp điều khiển, giám sát thiết bị điện, giám sát và phân tích các dữ liệu của thời tiết môi trường và đưa ra những tín hiệu khẩn cấp nhằm đối phó với những tình huống bất ngờ xảy ra. Hệ thống hoàn thiện góp phần tạo một sản phẩm công nghệ đảm bảo cho quá trình bảo mật trong hệ thống thông minh IoT

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Đại học Bách khoa Hà Nội trong đề tài mã số T2023-PC-008.

TÀI LIỆU THAM KHẢO

[1]. Dang T. V., Bui N. T., "Multi-scale Fully Convolutional Network based Semantic Segmentation for Mobile Robot Navigation," *Electronics*, 12(3), 533, 2022.

[2]. Dang T. V., Bui N. T., "Obstacle Avoidance Strategy for Mobile Robot based on Monocular Camera," *Electronics*, 12(8), 1932, 2023.

[3]. Dang T. V., Bui N. T., "Design the abnormal object detection system using template matching and subtract background algorithm," *MMMS 2022, LNME*, 2023.

[4]. Dang T. V., Phan V. T, Nguyen H. T., Hoang G. M., Bui N. T., "Design of a Face Recognition Technique based MTCNN and ArcFace." *MMMS 2022, LNME*, 2023.

[5]. Dang T. V., "Smart Attendance System based on Improved Facial Recognition," *Journal of Robotics and Control (JRC)*, 4(1), 46-53, 2023.

[6]. Dang T. V., "Smart home Management System with Face Recognition based on ArcFace model in Deep Convolutional Neural Network," *Journal of Robotics and Control (JRC)*, 3 (6), 754-761, 2022.

[7]. Gee A., Cipolla R., "Determining the gaze of faces in im ages," *Image and Vision Computing*, 12(10):639-647, 1994

[8]. Nguyen N. Q., Su S. F., Tran Q. V., Nguyen V. T., Jeng J. T., "Real time human tracking using improved CAM-shift," In *2017 Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS),* pp. 1-5, 2017.

[9]. Nguyen V. T., Nguyen A. T., Nguyen V. T., Bui H. A., "A real-time human tracking system using convolutional neural network and particle filter," In *Intelligent Systems and Networks*: Selected Articles from ICISN 2021, Vietnam, pp. 411-417, 2021.

AUTHORS INFORMATION

Ngo Viet Duc, Dang Thai Viet, Vu Ngoc Hai, Nguyen Nhu Truong

Hanoi University of Science and Technology, Vietnam