



XÂY DỰNG HỆ THỐNG ĐIỂM DANH HỌC VIÊN VỚI THÔNG TIN 3D CỦA KHUÔN MẶT TẠI TRUNG TÂM TIN HỌC, TRƯỜNG ĐẠI HỌC AN GIANG

Huỳnh Việt Trung¹, Nguyễn Văn Hoà¹

¹Trường Đại học An Giang, ĐHQG-HCM

Thông tin chung:

Ngày nhận bài: 17/02/2024

Ngày nhận kết quả bình duyệt:
07/06/2024

Ngày chấp nhận đăng:
01/2025

Title:

Building a system of recognizing learners with 3D information at CICT of An Giang University

Keywords:

Face Recognition, 3D, ResNet-50, RGB (Red Green Blue), RGB-D (Red Green Blue-Depth)

Từ khóa:

Nhận dạng khuôn mặt, 3D, ResNet-50, RGB (Red Green Blue), RGB-D (Red Green Blue-Depth)

ABSTRACT

We present the system of face recognition and identification based on real time using camera 3D and small computer Jetson Nano, We perform training for convolutional neural network model in order to extract features from face photos, with transfer learning technique, we train the model on 2 datasets, face photo having depth information and one with no depth information. Training data used from the set of data supplied by the organization of Ai BenchCouncil. In order to evaluate the model, we use a set of data collected by camera 3D, from 302 learners of CICT, An Giang university. The result identifies position of archives reaching exactness to 99,87% basing on the MTCNN (Multi-task Cascaded Convolutional Networks). The model of face recognition reaches to 99,77% with data of photos containing depth information of faces when using the machine learning SVM (Support Vector Machine).

TÓM TẮT

Trong bài viết này, chúng tôi trình bày một hệ thống nhận dạng và điểm danh khuôn mặt thời gian thực sử dụng camera 3D và máy tính nhỏ Jetson Nano, chúng tôi huấn luyện mô hình mạng nơ-ron tích chập để trích chọn đặc trưng của hình ảnh khuôn mặt, với kỹ thuật học chuyển giao trên 2 loại dữ liệu, hình ảnh khuôn mặt có thông tin chiều sâu và hình ảnh khuôn mặt không có thông tin chiều sâu. Dữ liệu huấn luyện mô hình trích đặc trưng khuôn mặt được sử dụng từ bộ dữ liệu được cung cấp bởi tổ chức Ai BenchCouncil. Để đánh giá các mô hình, chúng tôi sử dụng tập dữ liệu thu thập bằng camera 3D, từ 302 đối tượng là học viên Trung tâm Tin học, Trường Đại học An Giang. Kết quả xác định vị trí khuôn mặt trong bức ảnh đạt độ chính xác 99,87% dựa trên mạng nơ-ron xếp chồng MTCNN (Multi-task Cascaded Convolutional Networks). Mô hình định danh khuôn mặt đạt độ chính xác 99,77% với dữ liệu hình ảnh có thông tin chiều sâu của khuôn mặt khi sử dụng thuật toán máy học SVM (Support Vector Machine).

1. GIỚI THIỆU

Các hệ thống nhận dạng mặt người đã được nghiên cứu vào những năm 1970, tiêu biểu như công trình nhận dạng mặt người (Kanade, 1974)

đạt độ chính xác 75%. Tuy nhiên do các hạn chế của công nghệ tại thời điểm đó nên hệ thống xử lý còn phức tạp và kết quả nhận dạng

phụ thuộc nhiều vào tính chuẩn xác của dữ liệu ảnh đầu vào.

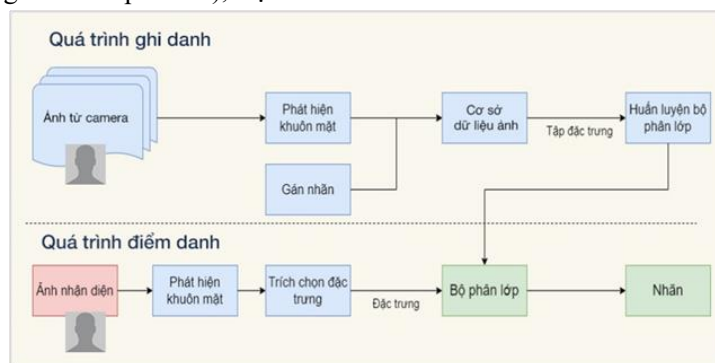
Quá trình nhận dạng mặt người bị chi phối bởi các yếu tố khác nhau như thay đổi về tỉ lệ ảnh, tịnh tiến, quay, che khuất, thay đổi độ sáng, nhiễu... từ đó thu hút nhiều nghiên cứu nhằm đưa ra các hướng tiếp cận khác nhau để có thể nhận dạng được mặt người tốt nhất. Một phương pháp được sử dụng và tỏ ra hiệu quả đó là sử dụng đặc trưng SIFT (Lowe, 2004) (Scale-Invariant Feature Transform) không bị thay đổi trước những biến đổi tỉ lệ ảnh, tịnh tiến, phép quay, không bị thay đổi một phần đối với phép biến đổi hình học affine (thay đổi góc nhìn) và mạnh với những thay đổi về độ sáng, nhiễu và sự che khuất. Các nghiên cứu tiêu biểu sử dụng đặc trưng SIFT như sau: đề xuất so khớp các đặc trưng SIFT để định danh khuôn mặt (Aly, 2006). Huấn luyện mô hình máy học SVM trên mô hình biểu diễn túi từ Bag-of-Words được tạo từ các đặc trưng SIFT cho phân lớp ảnh (Do, 2011). Bên cạnh đó, một số phương pháp khác được các nhà nghiên cứu đề xuất như sử dụng đặc trưng GIST (Oliva & Torralba, 2001), bài báo “Điểm danh bằng mặt người với đặc trưng GIST và máy học véc-tơ hỗ trợ” (Phạm Nguyên Khang, Trần Nguyễn Minh Thư & Đỗ Thanh Nghị, 2017) chỉ ra rằng các đặc trưng GIST có thể rút trích tập hợp các đặc trưng quan trọng như: tính tự nhiên (degree of naturalness), tính cởi mở (degree of openness), tính mở rộng (degree of expansion), độ nhám

(degree of roughness) và độ chắc chắn (degree of ruggedness), cho phép trình bày cấu trúc không gian của một cảnh. Bên cạnh đó đặc trưng GIST được trích chọn theo các hướng của các tần số khác nhau từ đó làm nổi bật thêm cấu trúc của ảnh.

Các thuật toán nhận dạng mặt người đã đạt được một bước tiến mới khi sử dụng mô hình học sâu (LeCun, Bottou, Bengio & Haffner, 1998), và được đánh giá trên các tập dữ liệu lớn như ImageNet (Fei-Fei, Deng & Li, 2009) với hơn 14 triệu ảnh các loại giúp nâng cao độ chính xác của kỹ thuật nhận dạng, một số công trình khác cũng mang lại độ chính xác cao tiêu biểu như OpenFace (Amos, Ludwiczuk & Satyanarayanan, 2016). DeepFace (Taigman, Yang, Ranzato & Wolf, 2014).

Hệ thống nhận dạng và điểm danh khuôn mặt thường gồm 2 quá trình: quá trình ghi danh và quá trình điểm danh, quá trình ghi danh thực hiện các bước: thu thập hình ảnh của các đối tượng trong hệ thống từ camera, thực hiện phát hiện khuôn mặt tự động, lưu ảnh khuôn mặt của các đối tượng vào cơ sở dữ liệu, rút trích đặc trưng khuôn mặt các đối tượng và huấn luyện mô hình phân lớp các đối tượng.

Quá trình điểm danh: thu ảnh đối tượng từ camera, định vị khuôn mặt, rút trích đặc trưng từ ảnh khuôn mặt và sử dụng mô hình phân lớp (được huấn luyện trong quá trình ghi danh) để định danh đối tượng.



Hình 1. Sơ đồ hệ thống nhận dạng và điểm danh

Gần đây một xu hướng mới được đề xuất, chỉ ra rằng cải thiện được độ chính xác, là nhận dạng khuôn mặt 3D (Xiong, Wen & Huang, 2019). Kỹ thuật này bổ sung bản đồ chiều sâu của khuôn mặt giúp phân biệt rõ ràng các đường phân cách của hốc mắt, mũi và cằm.

Nghiên cứu sử dụng thông tin 3D được tăng cường bởi sự phát triển của các bộ cảm biến tinh vi giúp nắm bắt hình ảnh chụp khuôn mặt 3D được tốt hơn. Các cảm biến hoạt động bằng cách chiếu ánh sáng có cấu trúc lên gương mặt (Integrated Electronics [Intel], 2019).

Bên cạnh các ưu điểm, hướng tiếp cận 3D cũng có những khó khăn về tập dữ liệu chiều sâu (RGB-D), như có số lượng nhỏ hơn tập dữ liệu 2D và không đủ để đào tạo mô hình CNN (Convolutional Neural Network) hiệu quả như các mô hình dựa trên ảnh RGB. Để giải quyết những khó khăn này, một số phương pháp được sử dụng như học chuyển đổi từ mô hình RGB được huấn luyện trước và lấy hình ảnh RGB-D làm đầu vào (Xiong, Wen & Huang, 2019).

Trung tâm Tin học - Trường Đại học An Giang là đơn vị hoạt động chuyên về lĩnh vực đào tạo, hiện trung tâm có 8 phòng máy đang hoạt động với số lượng học viên trung bình 2.000 học viên/năm. Công tác quản lý lớp học đã được tự động hóa mang lại một số hiệu quả đáng kể như công tác đăng ký ghi danh, tổ chức lớp,... Tuy

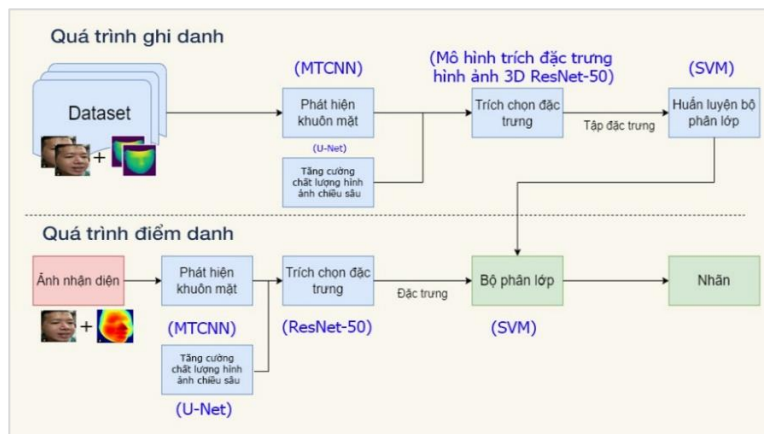
nhiên, công tác điểm danh học viên trong các buổi học vẫn còn thực hiện thủ công, chưa mang lại hiệu quả cao.

Hiện nay, công tác điểm danh tại Trung tâm Tin học, Trường Đại học An Giang được thực hiện như sau: trong các buổi học, nhân viên quản lý lớp đến từng phòng liên hệ giáo viên đứng lớp để cập nhật tình hình học viên vào sổ theo dõi. Sổ theo dõi được dùng làm căn cứ để theo dõi tình hình lớp học như tỉ lệ vắng, tỉ lệ tham dự và được dùng để giải quyết các vấn đề liên quan đến học vụ. Việc thực hiện thủ công này chưa mang lại hiệu quả cao do một số hạn chế như: quá tải công việc cho nhân viên quản lý lớp, việc thống kê còn nhiều sai sót, quá trình xác nhận học viên đã từng học tại Trung tâm còn mất nhiều thời gian. Xuất phát từ yêu cầu này chúng tôi xây dựng hệ thống điểm danh tự động các học viên bằng nhận dạng khuôn mặt 3D.

Nội dung tiếp theo của bài viết sẽ giới thiệu phương pháp nghiên cứu. Các kết quả thực nghiệm và thảo luận được trình bày ở phần 3, Kết luận và hướng phát triển sẽ được nêu ở phần 4 của bài báo.

2. PHƯƠNG PHÁP NGHIÊN CỨU

Chúng tôi sử dụng phương pháp thực nghiệm để xây dựng hệ thống nhận dạng và điểm danh khuôn mặt với 2 bước chính: quá trình ghi danh và quá trình điểm danh như Hình 1.



Hình 2. Mô hình tổng quan

Quá trình ghi danh: thu thập hình ảnh của các đối tượng trong hệ thống từ camera 3D, thực hiện phát hiện khuôn mặt tự động và tăng cường chất lượng hình ảnh chiều sâu, rút trích đặc trưng khuôn mặt và huấn luyện mô hình phân lớp các đối tượng.

Quá trình điểm danh: thu ảnh đối tượng từ camera, thực hiện phát hiện khuôn mặt tự động, tăng cường chất lượng hình ảnh chiều sâu, rút trích đặc trưng từ ảnh khuôn mặt và sử dụng mô hình phân lớp (được huấn luyện trong quá trình ghi danh) để định danh đối tượng.

Các nội dung chính của 2 quá trình trên bao gồm: huấn luyện mô hình rút trích đặc trưng hình ảnh khuôn mặt 3D, phát hiện khuôn mặt thu được từ camera 3D, tăng cường chất lượng hình ảnh chiều sâu khuôn mặt, và huấn luyện mô hình phân lớp đối tượng.

Các phương pháp thực hiện như sau:

- Huấn luyện mô hình rút trích đặc trưng hình ảnh khuôn mặt kết hợp thông tin chiều sâu, sử dụng mạng nơ-ron tích chập, kiến trúc mạng ResNet-50 (He, Zhang, Ren & Sun, 2016), dựa trên tập dữ liệu 3D Intellifusion được cung cấp bởi tổ chức Ai BenchCouncil bao gồm hình ảnh của 1.200 người nổi tiếng với các sắc tộc khác nhau.
- Phát hiện khuôn mặt với mạng nơ-ron xếp chồng sử dụng thư viện MTCNN (Zhang, K.,

Zhang, Z., Li & Qiao, 2016).

- Thu thập thông tin chiều sâu khuôn mặt sử dụng camera 3D Intel RealSense D435 (Integrated Electronics [Intel], 2019).

- Tăng cường chất lượng chiều sâu của khuôn mặt (Schlett, Rathgeb & Busch, 2021).

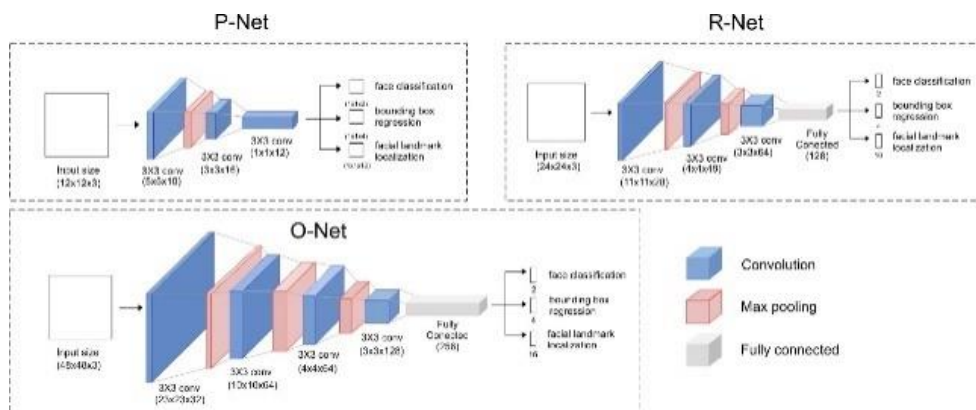
- Huấn luyện mô hình phân lớp bằng thuật toán máy học SVM (Vapnik, 1995) để định danh đối tượng từ ảnh khuôn mặt, dựa trên tập dữ liệu thu thập từ camera 3D bao gồm hình ảnh của 302 đối tượng là học viên của Trung tâm Tin học, Trường Đại học An Giang.

Thời gian nghiên cứu xây dựng hệ thống trong vòng 6 tháng, sau đây là nội dung thảo luận các bước nghiên cứu xây dựng hệ thống.

2.1 Định vị khuôn mặt sử dụng thư viện MTCNN

Hình ảnh thu được từ camera 3D có thể chứa phần nền và các đối tượng xung quanh, để hệ thống nhận dạng và định danh khuôn mặt hoạt động hiệu quả cần có một phương pháp xác định vùng bao quanh khuôn mặt, giai đoạn này gọi là phát hiện khuôn mặt, chúng tôi sử dụng thư viện MTCNN (Zhang, K., và cs., 2016).

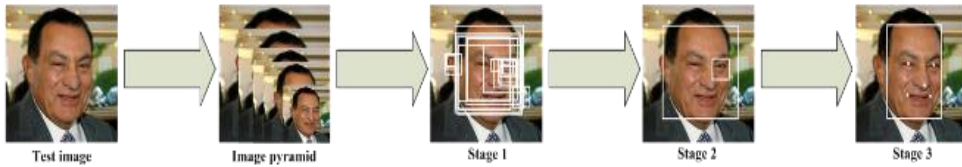
MTCNN gồm 3 mạng nơ-ron xếp chồng hoạt động cùng lúc để phát hiện khuôn mặt có trong hình ảnh, mỗi mạng nơ-ron có chức năng khác nhau lần lượt (P-Net, O-Net, R-Net) như Hình 3.



Hình 3. Kiến trúc mạng P-Net, R-Net, O-Net.

Phương pháp này hoạt động theo 3 bước chính: đầu tiên hình ảnh được đưa vào mạng P-Net (Proposal Network) nhằm dự đoán các vùng trong ảnh có chứa khuôn mặt, tiếp theo các hình ảnh được đưa vào R-Net (Refine Network) sử dụng kết quả của P-Net để loại bỏ các vùng không chứa khuôn mặt, cuối cùng hình ảnh

được đưa vào mạng O-Net (Output Network) xác định kết quả cuối cùng là tọa độ góc trên cùng bên trái và góc dưới bên phải của khuôn mặt và 5 đặc điểm trên khuôn mặt bao gồm, 2 điểm mắt, 1 điểm mũi và 2 điểm khóe miệng như Hình 4 Stage 3.



Hình 4. Các bước phát hiện khuôn mặt (Zhang, K., và cs., 2016).

2.2 Thu thập thông tin chiều sâu của khuôn mặt

Sau khi định vị được khuôn mặt, bước tiếp theo thu thập thông tin chiều sâu của khuôn mặt. Để thu thập thông tin chiều sâu khuôn mặt, chúng tôi sử dụng camera 3D Intel RealSense D435 (Intel, 2019). Camera này sử dụng 2 camera kép để thu nhận hình ảnh bên trái, và bên phải của khuôn mặt, đồng thời cảm biến hồng ngoại được quét trên bề mặt của khuôn mặt để tăng cường độ chính xác của chiều sâu, kết thúc quá trình này các thông tin thu thập được từ camera 3D được tính toán tạo thành bản đồ chiều sâu của khuôn mặt.

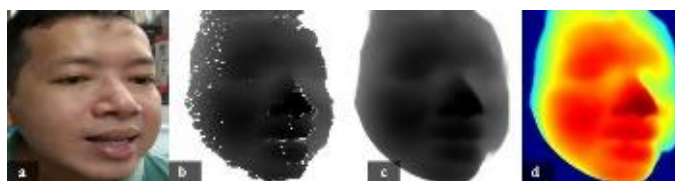
Ưu điểm phương pháp này: tạo ra bản đồ chiều sâu của khuôn mặt tương đối chính xác cho từng đối tượng. Tuy nhiên trong quá trình thu thập thông tin chiều sâu hình ảnh của đối tượng có thể bị phản xạ bởi ánh sáng của môi trường, do bề mặt của đối tượng bị ướt bởi mồ hôi, hay khoảng cách quét của cảm biến hồng ngoại đến đối tượng quá xa làm mất thông tin chiều sâu ở một số vùng trên khuôn mặt, vì vậy cần có một

biện pháp thực hiện tăng cường để phục hồi các thông tin chiều sâu này.

Chúng tôi sử dụng một phương pháp tăng cường hình ảnh chiều sâu khuôn mặt (Schlett, Rathgeb & Busch, 2021).

Phương pháp này giúp tăng cường hình ảnh thông tin chiều sâu của khuôn mặt bằng cách sử dụng một mô hình mạng nơ-ron tích chập dựa trên kiến trúc mạng U-Net (Ronneberger, Fischer & Brox, 2015) (Convolutional Networks for Biomedical Image Segmentation) đã được đào tạo trước, nhiệm vụ của mô hình này là phát hiện các lỗ hổng trên bản đồ chiều sâu của khuôn mặt và tiến hành tăng cường tại các vị trí hổng này.

Kết quả của phương pháp này thu được mô hình, dùng tăng cường hình ảnh độ sâu của khuôn mặt khi sử dụng các thiết bị như camera 3D như: camera 3D Intel RealSense, hình ảnh chiều sâu của khuôn mặt được tăng cường tại các vị trí bị mất thông tin chiều sâu như Hình 5b, kết quả thu được như Hình 5c.

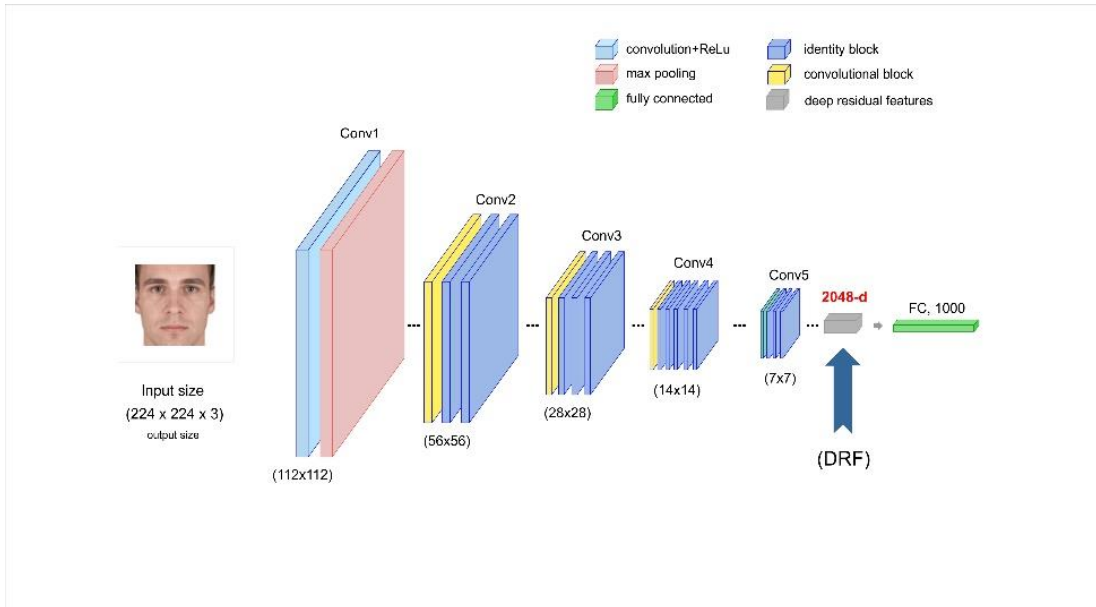


Hình 5. Thông tin chiều sâu khuôn mặt

2.3 Xây dựng mô hình trích đặc trưng

Tiếp theo chúng tôi xây dựng mô hình rút trích các đặc trưng quan trọng từ ảnh khuôn mặt, chúng tôi sử dụng kiến trúc mạng ResNet-50 (He và cs., 2015). Đây là kiến trúc mạng được

chứng minh là hiệu quả trong các bài toán xử lý ảnh với số lớp lớn (He và cs., 2015), sau khi hình ảnh được đưa qua kiến trúc mạng này sẽ được mã hoá thành 1 vec-tơ 2.048 chiều biểu diễn các đặc trưng của từng khuôn mặt.



Hình 6. Kiến trúc mạng ResNet-50

Kiến trúc mạng ResNet-50 (He và cs., 2015) bao gồm 55 lớp tích chập, 2 lớp pooling và 1 lớp kết nối đầy đủ. Số lớp có các tham số có thể điều chỉnh là 55 bao gồm 54 lớp tích chập và 1 lớp kết nối đầy đủ. Số lượng bộ lọc trong khối đầu tiên là 64, và được tăng dần trong các khối tiếp theo đó cho đến khi đạt 2.048, và lớp tích chập cuối cùng DRF (Deep Residual features) được sử dụng để trích chọn các đặc trưng của hình ảnh thành vector 2.048 chiều. Cuối cùng là một lớp kết nối đầy đủ với 1.000 nơ-ron.

Chúng tôi sử dụng kiến trúc mạng này kết hợp kỹ thuật học chuyển giao, với bộ trọng số đã được huấn luyện trên tập dữ liệu lớn ImageNet (Fei-Fei và cs., 2006) trong quá trình xây dựng mô hình trích đặc trưng hình ảnh khuôn mặt, để đảm bảo tính chính xác cho mô hình, chúng tôi

tiến hành thí nghiệm để so sánh ảnh hưởng của hình ảnh chiều sâu và kỹ thuật học chuyển giao trên dữ liệu ảnh thông thường và dữ liệu ảnh có thông tin chiều sâu.

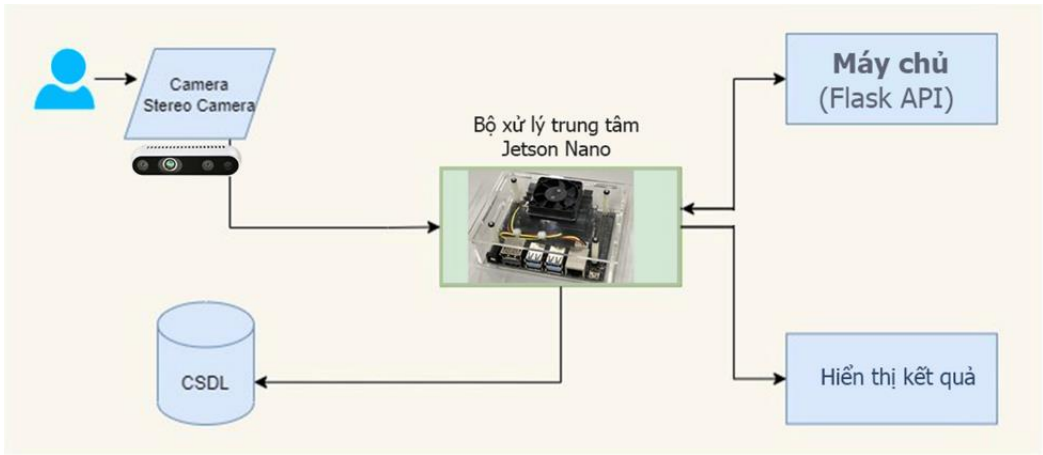
2.4 Xây dựng mô hình phân lớp

Sau khi qua bước huấn luyện trên, chúng tôi thu được mô hình dùng để trích đặc trưng hình ảnh khuôn mặt, tiếp theo chúng tôi sử dụng mô hình này duyệt qua tập hình ảnh khuôn mặt cần rút trích đặc trưng, để tạo dữ liệu đầu vào cho quá trình huấn luyện mô hình phân lớp, và để có tính đối sánh, chúng tôi sử dụng thuật toán máy học SVM (Vapnik, 1995) và kNN (Guo, Wang, Bell, Bi & Greer, 2003) (K-Nearest Neighbors), để xây dựng mô hình phân lớp định danh đối tượng.

3. KẾT QUẢ THỰC NGHIỆM, THẢO LUẬN

Mô hình thực tế của hệ thống vận hành theo hình thức máy chủ và máy khách, hình ảnh khuôn mặt và thông tin chiều sâu của khuôn mặt được thu thập và xử lý thông qua camera

3D và máy tính nhỏ Jetson Nano, sau đó được gửi về máy chủ sử dụng framework Flask để định danh và trả kết quả cho máy tính nhỏ Jetson Nano thông qua giao thức HTTP (Hypertext Transfer Protocol).



Hình 7. Mô hình chi tiết

Để đánh giá hiệu suất có đối sánh nên nghiên cứu tiến hành trên 2 thiết bị là máy tính và máy tính nhỏ Jetson Nano, chức năng và nhiệm vụ của 2 thiết bị này đóng vai trò thu thập và gửi

thông tin về cho máy chủ thực hiện định danh đối tượng và hiển thị kết quả định danh. Thông tin chi tiết về cấu hình hai thiết bị này được trình bày trong Bảng 1 và Bảng 2.

Bảng 1. Bảng mô tả thông số kỹ thuật máy tính

Thành phần	Thông số kỹ thuật
Hệ điều hành	Window 10, 64 bit
CPU	Core i7
Ram	16 GB
HDD	750 GB
GPU	x

Bảng 2. Bảng mô tả thông số kỹ thuật máy tính nhỏ Jetson Nano

Thành phần	Thông số kỹ thuật
Hệ điều hành	Ubuntu 20.4
CPU	4 Core ARM A57
Ram	4 GB
HDD	256 GB

GPU	128 Core MaxWell
-----	------------------

3.1 *Đánh giá giải thuật phát hiện khuôn mặt*

Chúng tôi thực hiện đánh giá giải thuật phát hiện khuôn mặt với 2 phương pháp HOG (Dalal & Triggs, 2005) (Histogram of Oriented

Gradients) và mạng nơ-ron xếp chồng thư viện MTCNN (Zhang, K., và cs., 2016). Kết quả của bước phát hiện khuôn mặt được trình bày trong Bảng 3.

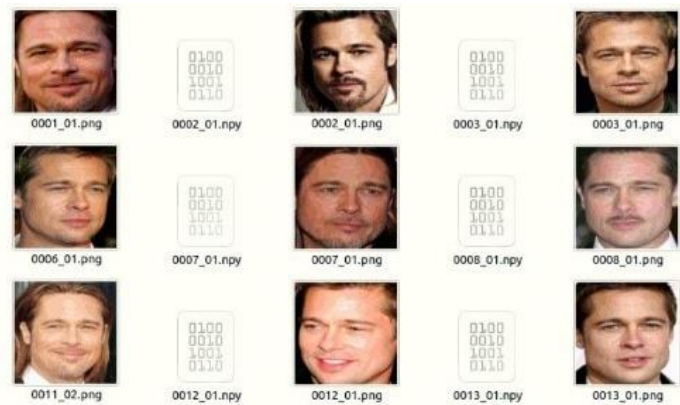
Bảng 3. So sánh giải thuật phát hiện khuôn mặt HOG và MTCNN

Phương pháp	Tổng số khuôn mặt	Số khuôn mặt nhận dạng đúng	Độ chính xác (%)	Thời gian (Giây/Hình)
HOG	17.776	13.751	77,35	0,03s
MTCNN	17.776	17.754	99,87	0,14s

Nhận xét: qua kết quả thực nghiệm chúng tôi thấy rằng giải thuật phát hiện khuôn mặt khi sử dụng đặc trưng HOG có độ chính xác thấp hơn giải thuật phát hiện khuôn mặt sử dụng mạng nơ-ron xếp chồng, cụ thể giải thuật phát hiện khuôn mặt HOG không phát hiện được các khuôn mặt ở góc nghiêng, với tiêu chí ưu tiên độ chính xác, do vậy chúng tôi đề xuất sử dụng giải thuật phát hiện khuôn mặt sử dụng mạng nơ-ron xếp chồng MTCNN.

3.2 *Xây dựng mô hình trích đặc trưng*

Dữ liệu huấn luyện mô hình trích đặc trưng khuôn mặt, chúng tôi sử dụng tập dữ liệu Intellifusion có thông tin chiều sâu của khuôn mặt RGB-D (Xiong và cs., 2019). Tập dữ liệu bao gồm hình ảnh của 1.200 người nổi tiếng với các sắc tộc khác nhau, tổng số hình ảnh 402.068. Mỗi thư mục với tên thư mục là tên nhân của đối tượng, số lượng ảnh trong mỗi lớp từ 100 - 600 ảnh, mỗi ảnh bao gồm 1 ảnh màu và một ảnh chiều sâu của khuôn mặt, các hình ảnh có độ phân giải khác nhau.



Hình 8. Tập dữ liệu 3D Intellifusion

Quy trình huấn luyện dữ liệu tạo mô hình trích đặc trưng khuôn mặt.

Bước 1: tiền xử lý dữ liệu.

Chúng tôi sử dụng thư viện MTCNN (Zhang, K., và cs., 2016) để phát hiện khuôn mặt. Tiếp

theo hình ảnh được thay đổi kích thước thành 224x224 pixel như Hình 8, trước khi đưa vào mạng ResNet-50 (He và cs., 2015).

Bước 2: tập dữ liệu huấn luyện được chia ngẫu nhiên theo tỉ lệ 8:1:1 như Bảng 4.

Bảng 4. Bảng phân chia dữ liệu huấn luyện mô hình trích đặc trưng theo tỉ lệ 8:1:1

Tập dữ liệu	Số lớp	Tổng số ảnh	Số ảnh huấn luyện	Số ảnh đánh giá	Số ảnh kiểm tra
Intellifusion	1.200	402.608	321.608	40.191	40.809

Bước 3: huấn luyện mô hình trích đặc trưng. Chúng tôi sử dụng kỹ thuật học chuyển giao, với bộ trọng số đã được huấn luyện trên tập dữ liệu lớn ImageNet để đảm bảo tính chính xác cho mô hình trích đặc trưng, chúng tôi huấn luyện mô hình bằng kiến trúc mạng ResNet-50 (He và cs., 2015) và sử dụng thư viện khoa học máy tính Pytorch. Cấu hình huấn luyện: máy tính hệ điều hành Ubuntu, Ram 8GB, GPU Nvidia GTX 1070 8GB.

Thông số huấn luyện: số lớp: 1.200, kích thước ảnh: 224x224 pixel, Batch size: 16, Epoch: 50, Loss function: Entropy, tối ưu lỗi: SGD (Stochastic Gradient Descent), Learning rate: 0,001.

Chúng tôi tiến hành 3 thí nghiệm để so sánh ảnh hưởng của hình ảnh chiều sâu và phương pháp học chuyển giao, kết quả chúng tôi thu được 3 mô hình như trong Bảng 5.

Bảng 5. Kết quả huấn luyện mô hình trích đặc trưng

Mô hình	Thông tin chiều sâu	Transfer learning (ImageNet)	Thời gian huấn luyện (phút)	Độ chính xác (%)
RGB	Không	Có	11.515	95,87
RGB-D	Có	Có	14.996	95,53
RGB-D*	Có	Không	17.501	91,06

Ghi chú: tên mô hình RGB-D là mô hình có sử dụng thông tin chiều sâu trong quá trình huấn luyện mô hình trích đặc trưng khuôn mặt, nhưng không sử dụng kỹ thuật học chuyển giao.*

Thời gian huấn luyện đối với mô hình có thông tin chiều sâu RGB-D và RGB-D* tương đương 10 ngày, và không có thông tin chiều sâu RGB tương đương 8 ngày.

Bước 4: thực hiện điều chỉnh loại bỏ lớp kết nối đầy đủ của kiến trúc mạng ResNet-50 (He và cs., 2015) không sử dụng hàm softmax để phân lớp đối tượng, với mục đích chỉ sử dụng mô hình thu được để trích đặc trưng khuôn mặt.

Nhận xét: qua kết quả thực nghiệm chúng tôi thấy rằng mô hình trích đặc trưng hình ảnh đạt kết quả cao nhất khi sử dụng kỹ thuật học chuyển giao đối với dữ liệu RGB và dữ liệu RGB-D.

3.3 Xây dựng mô hình phân lớp

Chúng tôi thu thập bộ dữ liệu của các đối tượng là học viên Trung tâm Tin học, Trường Đại học An Giang, với các độ tuổi khác nhau, hình ảnh được thu thập bằng camera 3D Intel RealSense D435 (Intel, 2019) với các góc nhìn khác nhau, chính diện, nhìn lên, nhìn xuống, bên trái và phải... mỗi thư mục là một lớp với tên thư mục là tên nhân của đối tượng.

Bước tiếp theo chúng tôi sử dụng thư viện MTCNN (Zhang, K., và cs., 2016) để phát hiện khuôn mặt và tăng cường thông tin chiều sâu của khuôn mặt (Schlett, Rathgeb & Busch, 2021), các hình ảnh được thay đổi kích thước về 224x224 pixel, như Hình 9.



Hình 9. Tập dữ liệu 3D thu thập

Để đảm bảo các hình ảnh thu thập không bị trùng hay quá giống nhau trong cùng một lớp, chúng tôi tiến hành phân loại theo ngưỡng khác nhau (*là sự khác biệt của một hình ảnh với tất cả các hình ảnh còn lại trong cùng 1 lớp*), chúng tôi chọn ngưỡng khác nhau là 25% (*tất cả các hình ảnh trong cùng một lớp phải khác nhau 25% cho các góc độ chụp khuôn mặt khi thu thập*).

Chúng tôi sử dụng thư viện Pillow (thư viện hỗ trợ xử lý hình ảnh được viết bằng ngôn ngữ

Python) để thực hiện phân loại theo ngưỡng tìm sự khác nhau, chúng tôi đề xuất chọn lựa 15 ảnh ngẫu nhiên trong một lớp với mục tiêu số lượng ảnh thu thập vừa đủ nhưng vẫn đảm bảo độ chính xác của mô hình phân lớp. Kết quả sau khi phân loại theo ngưỡng, chúng tôi thu được 302 lớp, số lượng ảnh là 4.530 ảnh.

Chúng tôi tiến hành phân chia dữ liệu huấn luyện mô hình phân lớp theo tỉ lệ 8:2 như Bảng 6.

Bảng 6. Bảng phân chia dữ liệu huấn luyện mô hình phân lớp theo tỉ lệ 8:2

Tập dữ liệu	Số lớp	Tổng số ảnh	Số ảnh huấn luyện	Số ảnh kiểm tra
Học viên	302	4.530	3.624	906

Để đánh giá mô hình đề xuất có đối sánh, chúng tôi sử dụng 2 thuật toán máy học: SVM (Vapnik, 1995) với tham số $C = 1$ (C là hằng số dương dùng để điều chỉnh độ rộng của lề phân hoạch và khoảng cách lề), $\gamma = 0,001$ (γ là tham số của hàm nhân được lựa chọn), $\text{kernel} = \text{linear}$ (tham số thể hiện hàm

nhân tuyến tính được lựa chọn), các tham số được chọn thông qua thí nghiệm kiểm tra chéo trên tập dữ liệu huấn luyện) và kNN (G. Guo và cs., 2003) với $k = 1$ sử dụng khoảng cách Euclid được lựa chọn thông qua quá trình thực nghiệm nhiều lần). Kết quả đánh giá mô hình phân lớp được trình bày trong Bảng 7.

Bảng 7. Đánh giá độ chính xác mô hình phân lớp

Mô hình	Tổng số ảnh	Số ảnh huấn luyện	Số ảnh đánh giá	SVM (%)	kNN (k = 1) (%)
RGB	4.530	3.624	906	99,77	99,55
RGB-D	4.530	3.624	906	99,77	99,22

Nhận xét: chúng tôi thấy rằng phương án sử dụng 15 ảnh trong 1 lớp, đều cho kết quả phân lớp đạt 99,77% đối với hai loại dữ liệu RGB và RGB-D khi sử dụng thuật toán SVM (Vapnik, 1995) tuyến tính, tiếp theo chúng tôi sử dụng mô hình phân lớp SVM (Vapnik, 1995) vừa đạt được, tiến hành thực nghiệm trên thiết bị máy tính và máy tính nhỏ Jetson Nano.

3.4 Đánh giá kết quả định danh khuôn mặt trong điều kiện ánh sáng bình thường

Bảng 8. Kết quả đánh giá độ chính xác trong điều kiện ánh sáng bình thường				
Mô hình	Tổng số ảnh	Số ảnh nhận đúng	Số ảnh nhận dạng sai	Độ chính xác (%)
RGB	255	237	18	92,94
RGB-D	255	241	14	94,50

Nhận xét: chúng tôi thấy rằng trong điều kiện ánh sáng bình thường, bộ phân lớp RGB-D định danh đối tượng cho kết quả tốt hơn bộ phân lớp RGB.

3.5 Đánh giá kết quả định danh khuôn mặt trong điều kiện ánh sáng kém

Chúng tôi tiến hành thu thập hình ảnh của 2 đối tượng với 150 hình ảnh có thông tin chiều sâu

Bảng 9. Kết quả đánh giá độ chính xác trong điều kiện ánh sáng kém.				
Mô hình	Tổng số ảnh	Số ảnh nhận đúng	Số ảnh nhận dạng sai	Độ chính xác (%)
RGB	150	65	85	43,30
RGB-D	150	77	73	51,00

Nhận xét: chúng tôi thấy rằng trong điều kiện ánh sáng kém, bộ phân lớp RGB-D định danh đối tượng cho kết quả tốt hơn bộ phân lớp RGB.

Dựa trên các kết quả thực nghiệm có thể chỉ ra sự đóng góp trong nghiên cứu như sau: với mô hình trích đặc trưng hình ảnh khuôn mặt được bổ sung thông tin chiều sâu (RGB-D) có thể

Chúng tôi tiến hành thu thập hình ảnh của 17 đối tượng với 255 ảnh có thông tin chiều sâu của khuôn mặt trong điều kiện ánh sáng bình thường và tiến hành so sánh thực nghiệm trên 2 mô hình phân lớp RGB và RGB-D.

Độ chính xác được tính theo công thức (Số lượng khuôn mặt định danh đúng/Tổng số khuôn mặt). Kết quả định danh khuôn mặt được trình bày trong Bảng 8.

của khuôn mặt trong điều kiện ánh sáng kém và tiến hành so sánh thực nghiệm trên 2 mô hình phân lớp RGB và RGB-D.

Độ chính xác được tính theo công thức (Số lượng khuôn mặt định danh đúng/Tổng số khuôn mặt). Kết quả định danh khuôn mặt được trình bày trong Bảng 9.

trích xuất được nhiều đặc trưng hơn các đặc điểm của khuôn mặt vì được bổ sung thông tin chiều sâu làm nổi bật các vùng phân cách của mắt mũi miệng..., dẫn đến độ chính xác của mô hình phân lớp có thông tin chiều sâu đạt độ chính xác tốt hơn.

3.6 Đánh giá thời gian định danh khuôn mặt trên máy tính và máy tính nhỏ Jetson Nano sử dụng bộ phân lớp RGB-D

Chúng tôi thực hiện đánh giá thời gian định

danh khuôn mặt với 5 lớp, bao gồm 625 ảnh có dữ liệu thông tin chiều sâu khuôn mặt, để so sánh thời gian thực hiện trên 2 thiết bị. Kết quả được trình bày trong Bảng 10.

Bảng 10. Thời gian định danh đối tượng trên máy tính và Jetson Nano

Thiết bị	Số lượng ảnh	Độ chính xác (%)	Tổng thời gian (Giây)	Một đơn vị thời gian (Giây/Hình)
Máy tính	625	87.84	109s	0,17s
Jetson Nano	625	85.76	254s	0,40s

Nhận xét: chúng tôi thấy rằng mô hình phân lớp RGB-D khi triển khai trên máy tính nhỏ Jetson Nano có độ chính xác thấp hơn và thời gian thực hiện định danh đối tượng lâu hơn trên máy tính lý giải cho việc mô hình có độ chính xác thấp hơn là vì trên máy tính nhỏ Jetson Nano sử dụng kiểu Int8 cho các phép toán thay vì trên

máy tính là kiểu Int32, Int64, dẫn đến ảnh hưởng độ chính xác khác nhau trên 2 thiết bị.

3.7 Đánh giá hiệu suất nhận dạng và định danh thời gian thực trên máy tính và máy tính nhỏ Jetson Nano sử dụng bộ phân lớp RGB-D

Bảng 11. Đánh giá hiệu suất nhận dạng và định danh thời gian thực

Thiết bị	Dlib	MTCNN	Độ trễ
Máy tính	4 FPS	1 FPS	0s
Jetson Nano	3 - 4 FPS	7 - 10 FPS	5 - 6s

Ghi chú: FPS (Frames per second là tốc độ khung hình hoặc tỷ suất khung hình, được tính bằng số khung hình trên một giây).

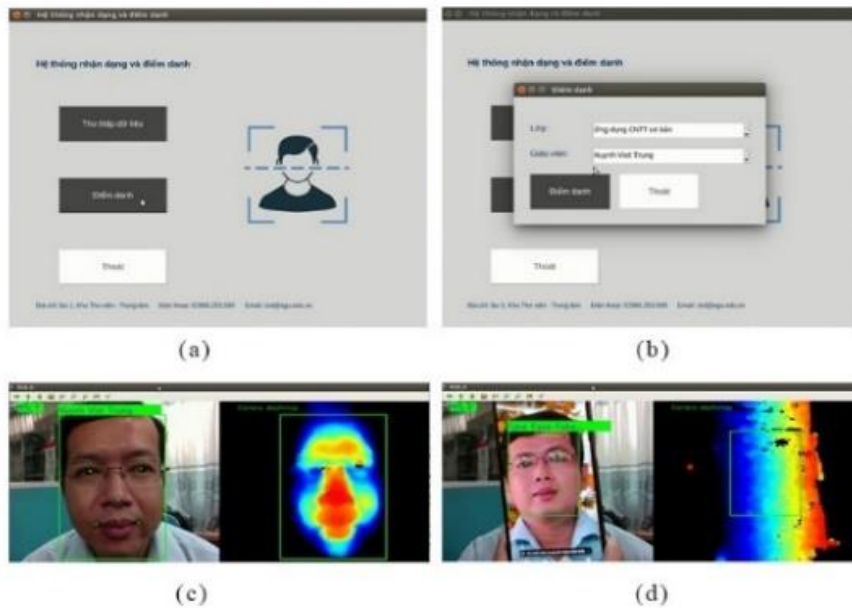
Độ trễ trên Jetson Nano: mô hình vận hành theo hình thức máy chủ và máy khách thông qua giao thức HTTP để truyền nhận dữ liệu nên phát sinh độ trễ, các bước thực hiện như sau:

Bước 1: nhận dạng khuôn mặt thời gian thực sử dụng thư viện MTCNN (Zhang, K., và cs., 2016).

Bước 2: hệ thống tiến hành gửi thông tin định danh của đối tượng bao gồm hình ảnh khuôn mặt và bản đồ chiều sâu của khuôn mặt đến máy chủ để tiến hành định danh và trả kết quả sau 50 khung hình.

Giá trị 50 khung hình được lựa chọn thông qua kết quả quan sát thực tế, hệ thống chạy ổn định không xảy ra hiện tượng giật khung hình. Kết quả định danh đối tượng được hiển thị theo quy ước sau: kết quả định danh hiển thị tại khung ảnh m2 = thời gian xử lý + kết quả yêu cầu dự đoán tại khung hình m1.

Nhận xét: mô hình đạt FPS cao nhất khi sử dụng bộ phát hiện khuôn mặt MTCNN (Zhang, K., và cs., 2016) với sự hỗ trợ của GPU trên máy tính nhỏ Jetson Nano.



Hình 10. Một số giao diện của hệ thống

Giao diện chính của hệ thống với 2 chức năng chính thu thập dữ liệu và điểm danh Hình 10(a), giao diện chức năng chọn lớp học điểm danh Hình 10(b), hệ thống hiển thị kết quả định danh là “tên của đối tượng” bên trên phần khung bao của khuôn mặt Hình 10(c), hệ thống hiển thị thông báo trong trường hợp người dùng điểm danh không hợp lệ bằng ảnh trên điện thoại hay hình ảnh không phải đối tượng thật Hình 10(d).

Chương trình được viết bằng ngôn ngữ Python, và vận hành theo hình thức máy khách và máy chủ sử dụng framework Flask (thư viện giúp xây dựng ứng dụng Website viết bằng ngôn ngữ Python), hệ thống nhận dạng bao gồm camera 3D và máy tính nhỏ Jetson Nano sẽ được bố trí tại lớp học.

4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi vừa trình bày một hệ thống nhận dạng và điểm danh khuôn mặt thời gian thực sử dụng camera 3D với máy học SVM (Vapnik, 1995) sử dụng đặc trưng được rút trích từ hình ảnh chiều sâu khuôn mặt bằng mạng nơ-ron tích chập. Hệ thống thực hiện định vị khuôn mặt thu được từ camera 3D bằng mạng nơ-ron xếp

chồng, đáp ứng tiêu chí về độ chính xác trong định vị, kết quả thực nghiệm trên tập dữ liệu gồm 4.530 của 302 đối tượng là học viên của Trung tâm Tin học, Trường Đại học An Giang đạt 99,77%, bên cạnh đó chúng tôi thực hiện thí nghiệm trên tập dữ liệu được thu thập trong môi trường ánh sáng kém gồm 150 hình ảnh, kết quả thực nghiệm cho thấy mô hình phân lớp có chứa thông tin chiều sâu khuôn mặt RGB-D cho kết quả cao hơn mô hình không có chiều sâu khuôn mặt RGB. Chúng tôi triển khai hệ thống nhận dạng và điểm danh khuôn mặt trên 2 thiết bị máy tính và máy tính nhỏ Jetson Nano trong thời gian thực, kết quả đạt được mức FPS ổn định 7 - 10 FPS trên máy tính nhỏ Jetson Nano nhờ sự hỗ trợ của GPU.

Bên cạnh đó, hệ thống vận hành với độ trễ nhất định trong thời gian thực trên thiết bị máy tính nhỏ Jetson Nano vì phụ thuộc tốc độ đường truyền mạng. Hệ thống chỉ điểm danh khuôn mặt với cự ly gần do ảnh hưởng của thiết bị camera 3D đặt ra, và chỉ điểm danh được 01 học viên tại cùng một thời điểm. Hệ thống này đang trong quá trình thử nghiệm tại Trung tâm

Tin học, Trường Đại học An Giang với phương án, camera 3D và máy tính nhỏ Jetson Nano được bố trí tại các lớp học.

Trong thời gian tới, chúng tôi tiếp tục hoàn thiện giao diện và hoàn thiện sản phẩm để có thể tạo thành mô đun tích hợp vào các phần mềm quản lý hệ thống khác như quản lý học viên, quản lý nhân viên.

TÀI LIỆU THAM KHẢO

- Aly, M. (2006). Face Recognition using SIFT features. *Technical Report*.
- Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6(2), 20.
- Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 886-893)*. Ieee.
- Do, T. N. (2011). Detection of pornographic images using Bag-of-Visual-Words and ARCX4 of random multinomial Naive Bayes, *Journal of Science and Technology* 49.
- Fei-Fei, L., Deng, J., & Li, K. (2009). ImageNet: Constructing a large-scale image database. *Journal of vision*, 9(8), 1037-1037.
- Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 986-996). Springer, Berlin, Heidelberg.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
- Huỳnh Phước Hải, Nguyễn Văn Hòa & Đỗ Thanh Nghị. (2017). So sánh mô hình học sâu với các phương pháp học tự động khác trong phân lớp dữ liệu biểu hiện Gen MicroArray. *Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR)*, Đà Nẵng.
- Intel. (2019). Intel RealSense D435. Intel-RealSense-D400-Series-Datasheet.pdf.
- Kanade, T. (1974). Picture Processing System by Computer Complex and Recognition of Human Faces. *Kyoto University (京都大学)*, May 23, 1974. Accessed: Oct. 22, 2021. [Online]. Available: <https://doi.org/10.14989/doctor.k1486>.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, vol. 86, no. 11, p. 47.
- Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110.
- Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3), 145-175.
- Phạm Nguyên Khang, Trần Nguyễn Minh Thư & Đỗ Thanh Nghị. (2017). Điểm danh bằng mặt người với đặc trưng GIST và máy học vectơ hỗ trợ. *Hội thảo khoa học về Nghiên cứu cơ bản và Ứng dụng Công nghệ Thông tin*.

- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241).
- Schlett, T., Rathgeb, C., & Busch, C. (2021). Deep learning-based single image face depth data enhancement. *Computer Vision and Image Understanding*, 210, 103247.
- Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1701-1708).
- Vapnik. (1995). The Nature of Statistical Learning Theory, *Springer-Verlag*.
- Xiong, X., Wen, X., & Huang, C. (2019). Improving RGB-D face recognition via transfer learning from a pretrained 2D network. In *International Symposium on Benchmarking, Measuring and Optimization* (pp. 141-148). *Springer, Cham*.
- Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10), 1499-1503.