



(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ

(19) Cộng hòa xã hội chủ nghĩa Việt Nam (VN)
CỤC SỞ HỮU TRÍ TUỆ

(11)



1-0052802

(51)^{2020.01} G06N 3/02

(13) B

(21) 1-2022-00531

(22) 25/01/2022

(45) 27/10/2025 451

(43) 25/08/2023 425A

(73) Trường Đại học Công nghiệp Hà Nội (VN)

Số 298 đường Cầu Diễn, phường Minh Khai, quận Bắc Từ Liêm, thành phố Hà Nội

(72) Trần Đức Quý (VN); Hoàng Anh (VN); Hoàng Minh Tuấn (VN); Lê Trường Giang (VN); Phạm Bá Tuấn Chung (VN); Vũ Quý Trung (VN); Nguyễn Đắc Nam (VN); Mai Văn Thanh (VN); Trương Văn Khải (VN); Nguyễn Văn Lương (VN).

(54) PHƯƠNG PHÁP SỬ DỤNG MÁY TÍNH ĐỂ HUẤN LUYỆN MÔ HÌNH NHẬN DẠNG NGƯỜI DỰA TRÊN KHUÔN MẶT VÀ KHUNG XƯƠNG

(57) Sáng chế đề cập đến phương pháp sử dụng máy tính để huấn luyện mô hình nhận dạng người dựa trên khuôn mặt và khung xương từ ảnh thu được từ các camera và ảnh khuôn mặt tự chụp sử dụng thị giác máy tính với độ chính xác cao và thời gian tối ưu. Phương pháp được xây dựng dựa trên cơ sở một số mô hình học sâu nổi bật được sử dụng trong các kỹ thuật nhận diện khuôn mặt để tìm kiếm và trích xuất đặc trưng khuôn mặt, kết hợp cùng với các thuật toán tiền xử lý ảnh đầu vào, các kỹ thuật làm dày dữ liệu, và bộ phân loại đặc trưng dựa trên mạng nơ-ron sâu DNN. Để thực hiện được phương pháp theo như sáng chế, cần trải qua các bước: tìm kiếm khuôn mặt trong ảnh; chuyển đổi ảnh chụp khuôn mặt từ các góc khác nhau về ảnh chụp chính diện khuôn mặt; trích xuất các đặc trưng khuôn mặt; chuẩn hóa vectơ đặc trưng và tính toán khoảng cách giữa các vectơ đặc trưng; phân loại đặc trưng và dự đoán hai khuôn mặt là “cùng người” hay “khác người”.



Lĩnh vực kỹ thuật được đề cập

Sáng chế trình bày phương pháp định danh sinh viên tự động bằng việc so sánh ảnh khuôn mặt chụp từ các camera trong các phòng học, ảnh khuôn mặt và khung xương (skeleton) của sinh viên sử dụng công nghệ thị giác máy tính. Cụ thể, phương pháp được trình bày sử dụng trong lĩnh vực nhận dạng và định danh khuôn mặt.

Tình trạng kỹ thuật của sáng chế

Phương pháp so sánh ảnh khuôn mặt giữa ảnh chụp từ các camera (máy quay, máy chụp ảnh, ...) và ảnh chụp của sinh viên khi nhập học được áp dụng trong các sản phẩm liên quan tới định danh sinh viên tự động của các trường đại học. Thông thường, công việc định danh sẽ được thực hiện trực tiếp bởi giáo viên, người hướng dẫn bằng mắt thường. Phương pháp so sánh ảnh khuôn mặt đóng góp trong việc tự động hóa để tiết kiệm thời gian, nhân lực cho các hoạt động liên quan đến xác minh danh tính của sinh viên và cũng góp phần làm tăng thời gian thời gian hoạt động giảng dạy trong lớp học. Việc so sánh ảnh từ các camera và ảnh khuôn mặt tự chụp có một số khó khăn nhất định bao gồm chất lượng ảnh chụp của mỗi sinh viên là không đồng nhất, ảnh thu được từ các camera thường không thu được toàn bộ khuôn mặt, hình dáng của sinh viên, thay đổi khuôn mặt do góc độ, khoảng cách ghi hình, trang điểm, tập dữ liệu huấn luyện hạn chế do các vấn đề liên quan đến bảo mật cá nhân. Những năm trở lại đây, cùng với sự phát triển của công nghệ trí tuệ nhân tạo đặc biệt có thể nói đến công nghệ thị giác máy tính, đã có một số phương pháp so sánh ảnh khuôn mặt ra đời sử dụng các kỹ thuật học máy (machine learning) và học sâu (deep learning). Trong sáng chế này, chúng tôi đề xuất một phương pháp mới sử dụng công nghệ thị giác máy tính cho vấn đề so sánh ảnh khuôn mặt của sinh viên khắc phục được các khó khăn hiện có để đạt được độ chính xác tốt hơn

trong thời gian xử lý tối ưu hơn, đặc biệt là áp dụng hiệu quả trên tập dữ liệu khuôn mặt của người Việt Nam.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế là đề xuất phương pháp so sánh khuôn mặt và khung xương giữa ảnh khuôn mặt và hình dáng của sinh viên thu được từ các camera và ảnh chụp của sinh viên phục vụ cho các hệ thống xác minh danh tính với độ chính xác cao và thời gian tối ưu. Để đạt được mục đích trên, phương pháp này gồm các bước:

Bước 1: tìm kiếm khuôn mặt và hình dáng của sinh viên trong các ảnh từ các camera; bước này được thực hiện dựa trên một số mô hình học sâu như IresNet, MobileFaceNet, MobileNet, DenseNet, MxNet để xác định vị trí khuôn mặt trong ảnh và xác định vùng chứa toàn bộ hình dáng cơ thể của sinh viên; kết hợp cùng một số kỹ thuật tiền xử lý ảnh đầu vào để tăng cường chất lượng ảnh và làm đa dạng dữ liệu khi ảnh được chụp từ nhiều thiết bị đa dạng, góc chụp và cự li chụp khác nhau và các tập dữ liệu liên quan còn hạn chế.

Bước 2: chuyển đổi các ảnh khuôn mặt từ bước 1 về góc chụp chính diện khuôn mặt và xác định khung xương của sinh viên; Đối với việc chuyển góc chụp, bước này được thực hiện dựa trên mô hình AnchorFace để biến những ảnh khuôn mặt thu được từ bước 1 nhưng được chụp từ các góc nhìn xiên, nhìn chéo,... thành góc nhìn trực diện. Đối với việc xác định khung xương bước này thực hiện sử dụng mô hình DualHeadNet để xác định khung xương của sinh viên thông qua dữ liệu liên quan đến hình dáng của sinh viên đã được xác định từ bước 1 nhằm làm tăng cơ sở để xác định hai sinh viên là “cùng người” hay “khác người”.

Bước 3: trích xuất các đặc trưng từ khuôn mặt và khung xương; bước này được thực hiện ứng dụng một số mô hình học sâu như Insightface, DeepFace để biến đổi ảnh khuôn mặt và khung xương ban đầu vào thành các vectơ đặc trưng dùng để đối sánh.

Bước 4: chuẩn hóa các vectơ đặc trưng và tính toán sự khác biệt giữa các vectơ đặc trưng; bước này được thực hiện dựa trên phép chuẩn hóa Entropy, sau đó lựa chọn cặp ảnh khuôn mặt và khung xương ngẫu nhiên để huấn luyện mô hình phân loại ở bước 5 dựa trên hàm mất mát côsin (cosine similarity) và tính độ lớn khoảng cách của các vectơ đặc trưng sau khi chuẩn hóa và ghép cặp.

Bước 5: phân loại đặc trưng và dự đoán hai khuôn mặt là “cùng người” hay “khác người”; bước này được thực hiện dựa trên một số mạng nơron sâu (Deep neuron network - DNN) để học các đặc trưng dùng cho việc so sánh; khi đưa hai ảnh khuôn mặt và ảnh khung xương đầu vào ở bước 1, đầu ra ở bước 5 trả về xác suất trùng khớp (%) của hai cặp ảnh đó và đưa ra kết quả dự đoán.

Mô tả vắn tắt các hình vẽ

Hình 1: hình ảnh mô tả các bước trong phương pháp huấn luyện mô hình nhận dạng người dựa trên khuôn mặt và khung xương.

Hình 2: hình ảnh mô tả bước phân loại đặc trưng và dự đoán hai khuôn mặt có phải cùng một người hay không.

Hình 3: sơ đồ mô phỏng hoạt động của phương pháp huấn luyện mô hình nhận dạng người dựa trên khuôn mặt và khung xương.

Mô tả chi tiết sáng chế

Các bước và trình tự các bước của phương pháp so sánh ảnh khuôn mặt giữa ảnh khuôn mặt và ảnh hình dáng thu được từ các camera và ảnh chụp của sinh viên phục vụ cho các hệ thống định danh tự động được mô tả trong tham chiếu Hình 1. Dữ liệu đầu vào là các ảnh thu được từ các camera đặt ở nhiều vị trí khác nhau và ảnh chụp của sinh viên khi nhập học được chụp từ nhiều thiết bị đa dạng và từ nhiều góc độ, cự li khác nhau.

Tuy nhiên, cũng cần phải nói thêm rằng trong sáng chế này, các thuật ngữ “DeepFace”, “Insightface”, “IresNet”, “MobileFaceNet”, “MobileNet”, “DenseNet”, “MxNet”, “Entropy”, “ResNet”, “ResNet100”, “ArcFace”, “AnchorFace”, “Sigmoid”, “DenseLayer”, “Cross-entropy”, “Adam” là các

danh từ riêng (tên của mô hình, thuật toán, hàm số, vv) chưa có từ tiếng Việt thay thế nên xin được để ở dạng chuẩn quốc tế.

Các bước của phương pháp được thực hiện như sau:

Bước 1: tìm kiếm khuôn mặt và hình dáng của sinh viên trong ảnh. Nhiệm vụ của bước này là tìm kiếm được vùng chứa khuôn mặt và vùng chứa hình dáng của sinh viên trong ảnh và cắt ra vùng chứa khuôn mặt và hình dáng đó. Bằng việc sử dụng mô hình học sâu RetinaFace trong việc đưa ra các đặc trưng dạng tháp để tìm kiếm ra vùng chứa khuôn mặt (bounding box).

Do dữ liệu đầu vào là các ảnh từ các camera và ảnh chụp sinh viên được chụp từ nhiều thiết bị đa dạng và từ nhiều góc độ, cự li khác nhau nên chất lượng ảnh đầu vào là rất đa dạng tùy thuộc vào thiết bị sử dụng, điều kiện chụp ảnh (phông nền, độ sáng, vv), thêm vào đó là số lượng dữ liệu đầu vào hạn chế, vì vậy, cần kết hợp với các kỹ thuật tiền xử lý ảnh đầu vào và các kỹ thuật làm dày dữ liệu phù hợp, đặc biệt là xử lý ảnh với các ảnh thu được từ các camera do ảnh này chứa khuôn mặt có thể rất nhỏ và mờ và hình dáng người dễ bị nhầm lẫn bởi các đối tượng khác. Với các ảnh như vậy, kỹ thuật tiền xử lý được mô tả như sau:

- Kỹ thuật thứ nhất sử dụng cho ảnh khuôn mặt tự chụp: với mục đích giảm thời gian xử lý và đạt được độ chính xác cao hơn trong việc xác định vị trí khuôn mặt với nhiều cự li chụp khác nhau, kỹ thuật lựa chọn ngẫu nhiên (random search) kích cỡ ảnh đầu vào phù hợp với mô hình RetinaFace được sử dụng. Cụ thể, một số kích cỡ (multi-scale) chiều rộng ảnh được đề xuất như [4096; 1024; 512; 256] điểm ảnh (pixels) để thay đổi (resize) kích thước ảnh đầu vào và lựa chọn kích cỡ tốt nhất.

- Đối với việc làm dày dữ liệu phù hợp: với mục đích làm tăng số lượng ảnh dữ liệu đầu vào giúp nâng cao độ chính xác của mô hình. Bước này sẽ sử dụng một số kỹ thuật biến đổi ảnh như thay đổi màu sắc (color transform), làm mờ ảnh (blur image),... với một số tham số biến đổi ảnh (transform image parameter) phù hợp với mô hình mạng MxNet được áp dụng với bộ tham số

gồm: hệ số tương phản contrast = 0.1, độ bão hoà saturation = 0.1, độ sáng brightness = 0.1.

Bước 2: chuyển đổi ảnh chụp khuôn mặt từ các góc khác nhau về ảnh chụp chính diện khuôn mặt và xác định khung xương của sinh viên.

Trong thực tế ảnh thu được từ các camera thường chụp nhiều góc độ khác nhau của khuôn mặt dẫn đến việc nhận dạng sẽ gặp nhiều khó khăn sau này. Để giải quyết vấn đề này, sáng chế sử dụng mô hình AnchorFace, với mỗi ảnh từ tập dữ liệu đầu vào mô hình sẽ thực hiện sinh ra các ma trận neo (anchor matrix) và xây dựng ma trận độ tin cậy (confident matrix) của từng điểm ảnh làm cơ sở để tổng hợp kết quả nhận diện. Cụ thể, sau quá trình thực nghiệm mô hình sẽ thu được vectơ chứa 68 điểm mốc (landmark) đại diện cho ảnh khuôn mặt từ đó sẽ biến đổi ảnh chụp từ các góc khác nhau về góc chính diện. Đối với việc xác định khung xương, sáng chế sử dụng mô hình DualHeadNet với mỗi ảnh hình dáng thu được từ bước 1 mô hình này sẽ đưa ra các ma trận không gian (Spatial) và ma trận không gian (Temporal) từ đó đưa ra khung xương của sinh viên dưới dạng ma trận có kích thước trùng với kích thước ảnh đầu vào.

Bước 3: trích xuất đặc trưng từ khuôn mặt và khung xương.

Nhiệm vụ chính của bước này là biến đổi dữ liệu ảnh khuôn mặt và khung xương tại bước số 1 và 2 thành các vectơ đặc trưng đầu ra để phục vụ vấn đề đối sánh khuôn mặt sử dụng các mạng học sâu. Hai mạng nơron tích chập (Convolutional neural network - CNN) được đề xuất sử dụng ở bước này là mô hình học sâu nén khuôn mặt và khung xương Insightface và DeepFace. Cả hai mô hình InsightFace và DeepFace được xây dựng dựa trên mô hình cơ sở là ResNet100 tuy nhiên InsightFace với lớp sau của tầng CNN cuối cùng của mạng ResNet100, được áp dụng kỹ thuật chuẩn hóa theo lô (batch normalization) và học bỏ qua (drop-out), còn DeepFace sử dụng một lớp kết nối đầy đủ (fully-connected) thay cho tầng cuối cùng của mạng ResNet100. Kết quả đầu ra của hai mô hình này là vectơ đặc trưng khuôn mặt có kích thước là 2×256 . Đặc biệt, mô hình InsightFace sử dụng hàm giá trị ArcFace (ArcFace loss) để trích chọn được những đặc trưng tốt nhất dùng để phân loại trong các bài toán nhận diện

khuôn mặt. Mô hình InsightFace sẽ được sử dụng đối với các ảnh thu được từ các camera còn DeepFace sẽ được sử dụng đối với các ảnh chụp của sinh viên khi nhập học.

Bước 4: chuẩn hóa vectơ đặc trưng và tính toán sự khác biệt giữa các vectơ đặc trưng.

Với việc sử dụng hai mô hình mạng khác nhau cho dữ liệu đầu vào các vectơ đặc trưng khuôn mặt cần được chuẩn hóa theo Entropy như công thức sau:

$$H(X) = - \sum \pi(x) \log(\pi(x)) \quad (1)$$

trong đó: x là vectơ đặc trưng, $x \in \mathbb{R}^{2 \times n}$, và $n = 256$ là kích thước chiều rộng của vectơ đặc trưng thu được từ các mô hình biến đổi ảnh thành các vectơ đặc trưng tại bước 3. Sau khi các vectơ đặc trưng được chuẩn hóa, sự khác biệt giữa các khuôn mặt và khung xương được tính toán dựa trên khoảng cách của hai vectơ đặc trưng giữa ảnh từ các camera và ảnh chụp sinh viên của “cùng người” và “khác người”. Giả sử, hai vectơ đặc trưng là a and b , khoảng cách được tính theo công thức sau:

$$f_{ab} = |a_{norm} - b_{norm}| \quad (2)$$

trong đó a_{norm} và b_{norm} là các vectơ sau khi chuẩn hóa, f_{ab} sẽ được dùng làm đầu vào của bước 5.

Quá trình ghép cặp để tính hiệu số f_{ab} yêu cầu ghép cặp ngẫu nhiên vectơ đặc trưng của ảnh từ các camera và ảnh khuôn mặt tự chụp để đảm bảo độ chính xác và đa dạng của mô hình. Hiệu số f_{ab} sẽ được sử dụng cho việc huấn luyện mô hình phân loại ở bước 5 dựa trên hàm mất mát tính độ tương tự cosin. Ý tưởng sử dụng thuật toán này dựa trên việc giảm thiểu giá trị f_{ab} đối với các cặp ảnh được gán nhãn là “cùng người” và giá tăng giá trị f_{ab} đối với các cặp ảnh được gán nhãn là “khác người” quá trình này sẽ giúp phân nhóm rõ ràng các cặp dữ liệu đầu vào và là cơ sở cho việc phân loại hai nhãn “cùng người” và “khác người” thêm vào đó kỹ thuật này đòi hỏi sự cân bằng trong các cặp dữ liệu đầu vào, trong thực tế dữ liệu về các cặp “cùng người” thường ít hơn nhiều so với

“khác người” điều này có thể dẫn đến mô hình đạt hiệu quả cao trong thí nghiệm nhưng lại có hiệu quả kém khi áp dụng vào thực tế chính vì vậy kỹ thuật làm dày dữ liệu tại bước 1 đã trình bày sẽ giúp hạn chế vấn đề này. Thuật toán tính độ tương tự cosin đo độ giống nhau giữa hai vectơ của một không gian tích bên trong. Độ tương tự giữa chúng được đo dựa trên tính toán cosin góc giữa hai vectơ đó. Cho hai vectơ A và B , độ tương tự cosin được biểu diễn theo công thức sau:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (3)$$

trong đó $\cos(\theta)$ là độ tương tự cosin, và A_i và B_i lần lượt là các thành phần thứ i của hai vectơ A và B .

Bước 5: phân loại đặc trưng và dự đoán hai khuôn mặt là “cùng người” hay “khác người”. Sau khi đã ghép cặp các vectơ dựa trên thuật toán đo độ tương tự cosin và tính toán các độ lớn hiệu số f_{ab} , các hiệu số này sẽ được dùng để huấn luyện mô hình phân loại. Một mạng nơron sâu (DNN) với năm tầng được đề xuất sử dụng ở bước này, mô tả của mạng này tham chiếu ở Hình 2. Đầu vào của mạng các vectơ có chiều 2×256 . Tầng thứ hai của mạng được xây dựng với 256 nơron để học các đặc trưng từ đầu vào, tầng thứ ba của mạng được xây dựng với kích thước 128 nơron để học các đặc trưng tối ưu với đầu vào là kết quả thu được từ tầng thứ hai, tầng thứ tư sử dụng dữ liệu đầu vào là kết quả thu được của tầng thứ ba với kích thước 64 nơron nhằm mục đích học các đặc trưng tốt nhất, tầng cuối cùng là một lớp DenseLayer với hàm kích hoạt softmax trả về xác suất so sánh khuôn mặt và khung xương thuộc về lớp “cùng người” hay “khác người”. Các tham số và siêu tham số (hyper-parameters) của mô hình huấn luyện lần lượt là 100 (epoch size), tốc độ học (learning rate) là 10^{-4} , kích thước lô khi huấn luyện (batch size) là 32, hàm mất mát (loss) là binary Cross-entropy, hàm tối ưu hóa (optimizer) là Adadelata với các tham số: tốc độ học (learning rate) là 10^{-4} , hệ số phân rã (decay rate) $\rho=0.90$, sai số cho phép $\epsilon = 10^{-7}$

Mạng nơron sâu (DNN) sẽ trả về xác suất hai khuôn mặt khớp nhau và không khớp nhau là bao nhiêu phần trăm (%), tại đây cần xác định một ngưỡng phân loại phù hợp để xác định hai khuôn mặt là “cùng người” hay “khác người”. Sau khi áp dụng thực nghiệm sáng chế lựa chọn ngưỡng này là 0.7.

Yêu cầu bảo hộ

1. Phương pháp sử dụng máy tính để huấn luyện mô hình nhận dạng người dựa trên khuôn mặt và khung xương, bao gồm các bước sau:

thu tập dữ liệu ảnh chứa khuôn mặt và tập ảnh chứa hình dáng người, trong đó tập dữ liệu ảnh là dữ liệu ảnh thu từ camera được đặt ở nhiều vị trí khác nhau và ảnh khuôn mặt tự chụp từ trước, thực hiện tiền xử lý tập ảnh chứa khuôn mặt và tập ảnh chứa hình dáng người này để tìm kiếm vùng ảnh chứa khuôn mặt và vùng ảnh chứa hình dáng của người trong ảnh và cắt ra vùng ảnh chứa khuôn mặt và hình dáng đó, trong đó, tại bước này sử dụng mô hình học sâu RetinaFace trong việc đưa ra các đặc trưng dạng tháp để tìm kiếm ra vùng chứa khuôn mặt (bounding box) và các điểm mốc (landmark) đại diện cho ảnh khuôn mặt;

chuyển đổi ảnh chứa khuôn mặt từ các góc khác nhau về ảnh chính diện khuôn mặt và xác định khung xương của người từ ảnh chứa hình dáng người;

trích xuất đặc trưng từ vùng khuôn mặt và khung xương để đưa ra các vectơ đặc trưng khuôn mặt và khung xương;

chuẩn hóa các vectơ đặc trưng khuôn mặt và khung xương và tính toán khoảng cách giữa các vectơ đặc trưng này tính theo công thức sau:

$$f_{ab} = |a_{norm} - b_{norm}| \quad (2)$$

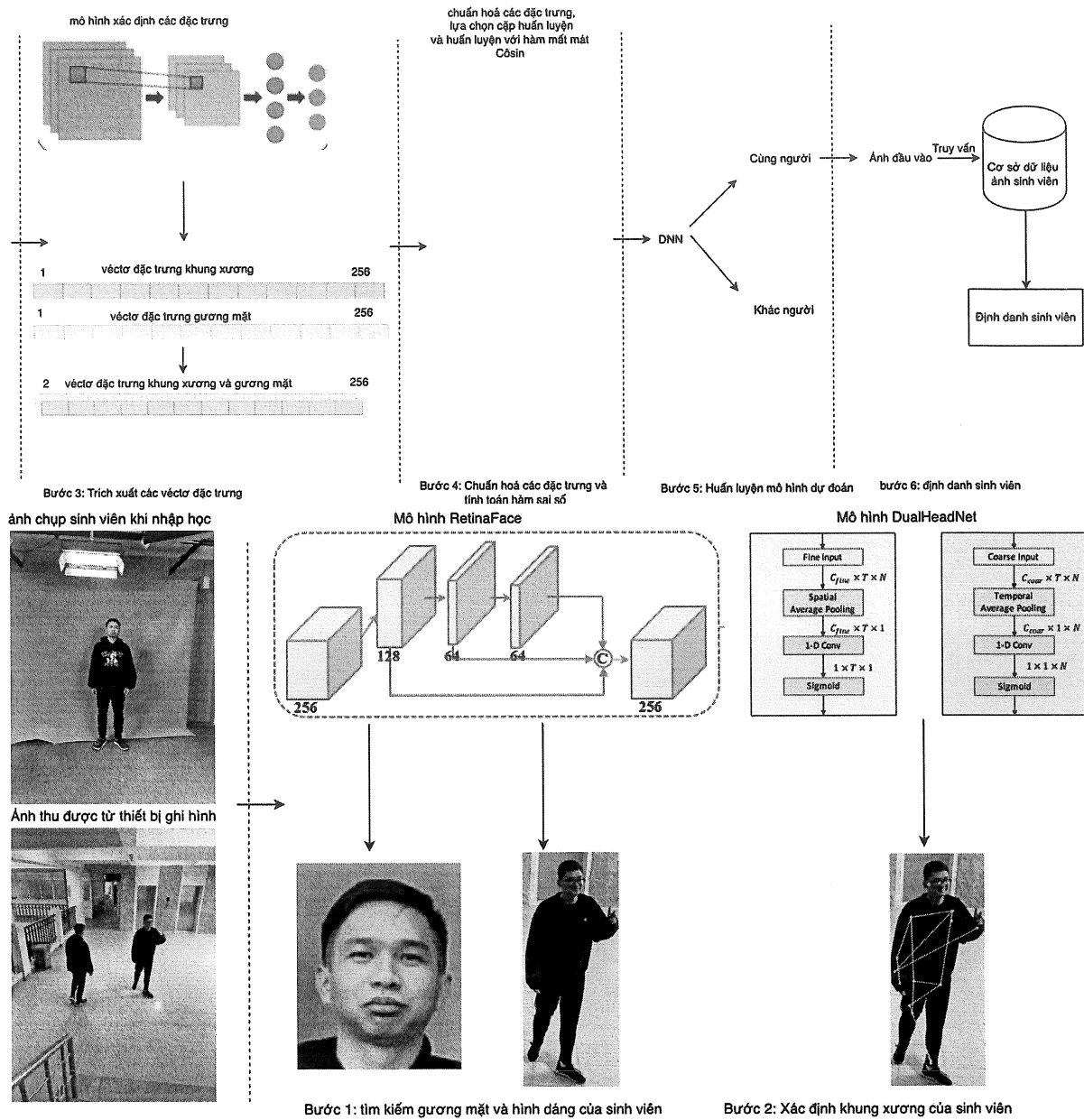
trong đó a_{norm} và b_{norm} là các vectơ sau khi chuẩn hóa và việc ghép cặp để tính hiệu số f_{ab} là ghép cặp ngẫu nhiên vectơ đặc trưng khuôn mặt và khung xương từ các camera và ảnh khuôn mặt tự chụp dựa trên độ tương tự cosin được biểu diễn theo công thức sau:

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (3)$$

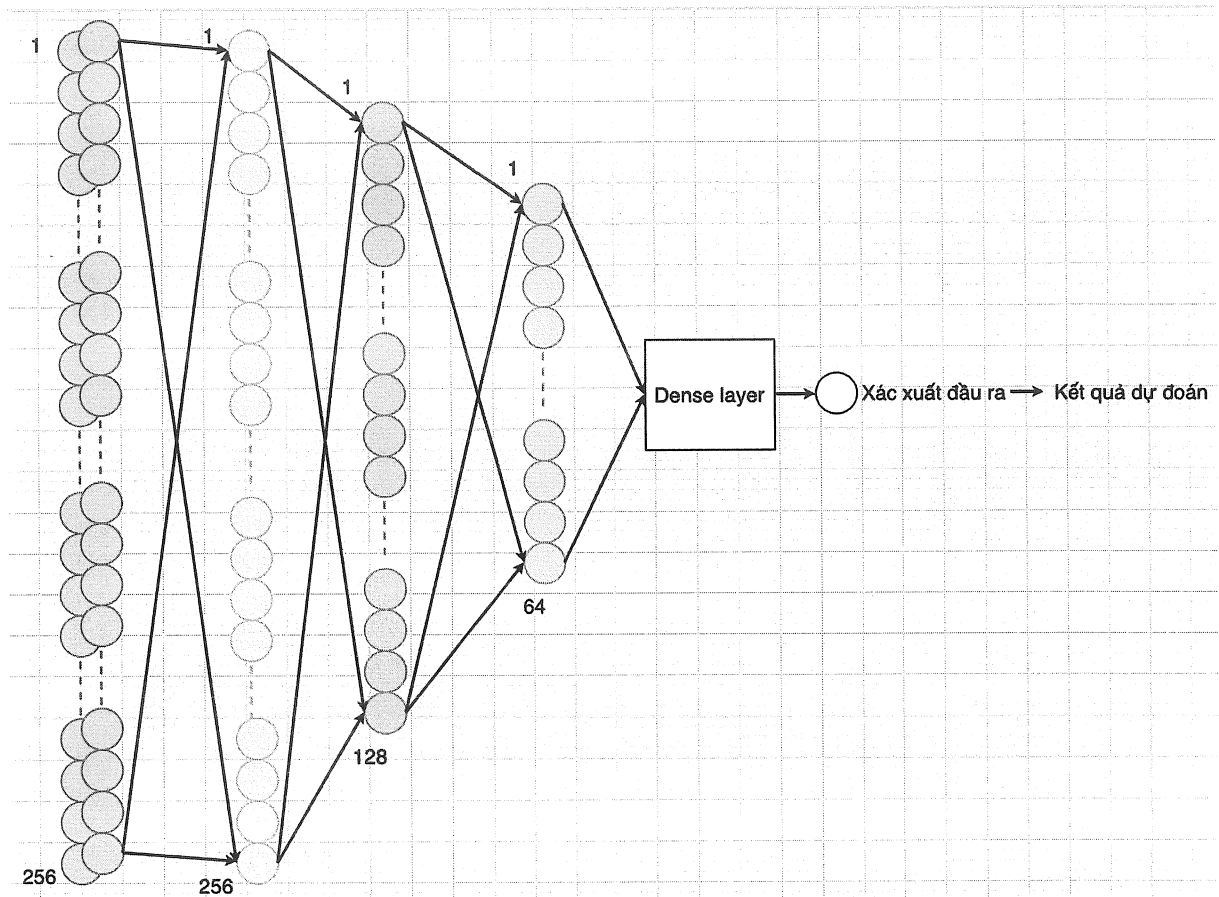
trong đó $\cos(\theta)$ là độ tương tự cosin, và A_i và B_i lần lượt là các thành phần thứ i của hai vectơ A và B ;

phân loại đặc trưng và dự đoán hai khuôn mặt là “cùng người” hay “khác người”, trong đó, sử dụng mạng nơron sâu (DNN) với năm tầng để tính xác suất so sánh khuôn mặt và khung xương thuộc về lớp “cùng người” hay “khác người” dựa trên ngưỡng phân loại được thiết lập trước.

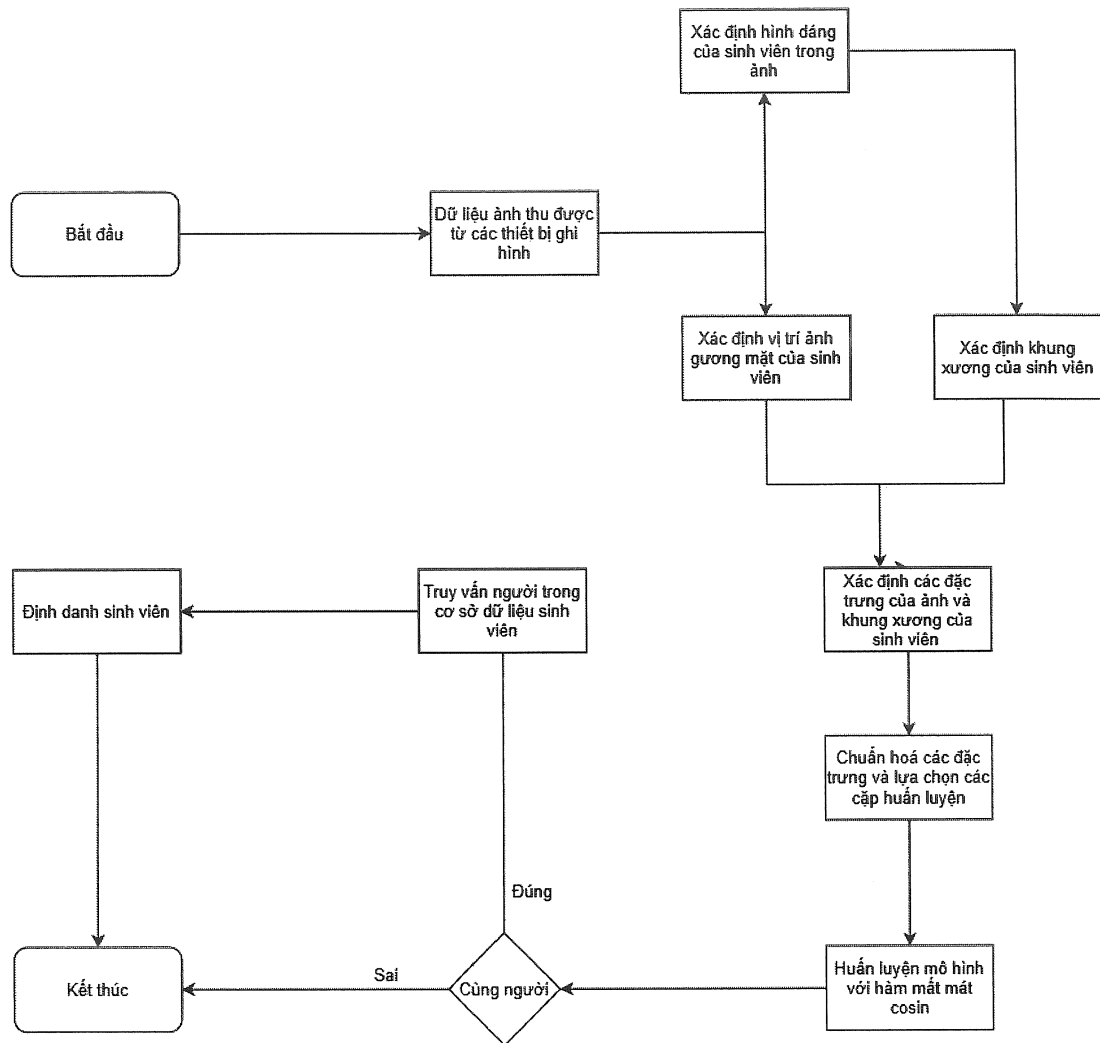
2. Phương pháp sử dụng máy tính để huấn luyện mô hình nhận dạng người dựa trên khuôn mặt và khung xương theo điểm 1, trong đó: ở bước phân loại đặc trưng và dự đoán hai khuôn mặt, để phân loại đặc trưng và dự đoán hai khuôn mặt có phải cùng một người hay không; mạng nơron sâu (DNN) gồm năm tầng được sử dụng với các tham số (parameters) và siêu tham số (hyper-parameters) như sau: 100 (epoch size), tốc độ học (learning rate) là 10^{-4} , kích thước lô khi huấn luyện (batch size) là 32, hàm mất mát (loss) là binary Cross-entropy, hàm tối ưu hóa (optimizer) là Adadelta với các tham số: tốc độ học (learning rate) là 10^{-4} , hệ số phân rã (decay rate) $\rho=0.90$, sai số cho phép $\epsilon = 10^{-7}$.



Hình 1



Hình 2



Hình 3