



(12) **BẢN MÔ TẢ GIẢI PHÁP HỮU ÍCH THUỘC BẢNG ĐỘC QUYỀN
GIẢI PHÁP HỮU ÍCH**

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

(11)



2-0003870

(51)

(13) **Y**

(21) 2-2024-00031

(22) 24/12/2021

(67) 1-2021-08365

(45) 27/01/2025 442

(43) 25/02/2022 407

(73) **TẬP ĐOÀN CÔNG NGHIỆP - VIỄN THÔNG QUÂN ĐỘI (VN)**

Lô D26 Khu đô thị mới Cầu Giấy, phường Yên Hòa, quận Cầu Giấy, thành phố Hà Nội.

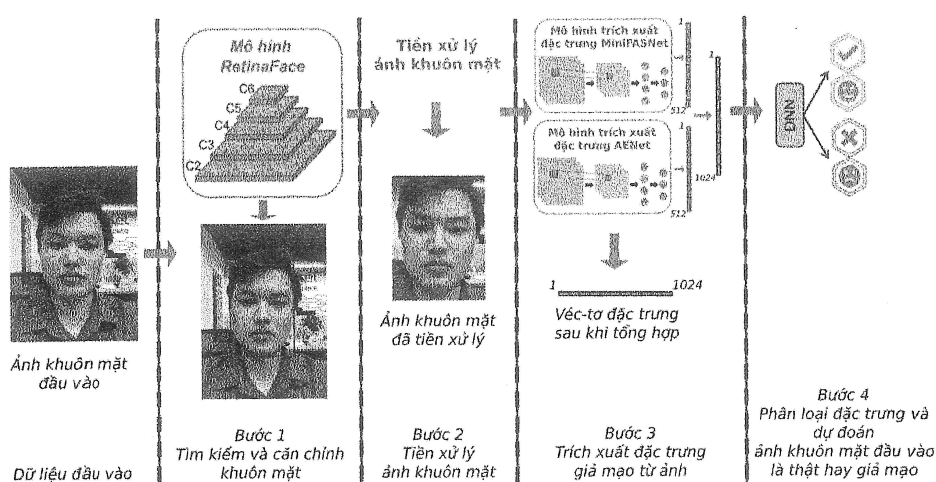
(72) **Đào Thanh Tùng (VN); Vũ Thị Hạnh (VN); Nguyễn Mạnh Quý (VN); Hoàng Ngọc Dương (VN); Nguyễn Quốc Khánh (VN); Phạm Văn Đan (VN).**

(74) **Công ty TNHH NACILAW (NACILAW)**

(54) **PHƯƠNG PHÁP PHÁT HIỆN GIẢ MẠO KHUÔN MẶT QUA HÌNH ẢNH SỬ DỤNG CÔNG NGHỆ TRÍ TUỆ NHÂN TẠO**

(21) 2-2024-00031

(57) Sáng chế đề cập đến phương pháp phát hiện giả mạo khuôn mặt qua hình ảnh sử dụng công nghệ trí tuệ nhân tạo. Phương pháp bao gồm các bước: bước 1: tìm kiếm và căn chỉnh khuôn mặt; bước 2: tiền xử lý ảnh khuôn mặt; bước 3: trích xuất đặc trưng giả mạo khuôn mặt bằng cách sử dụng các mô hình học sâu khác nhau để kết hợp nhiều loại đặc trưng đa dạng; bước 4: phân loại đặc trưng và dự đoán ảnh khuôn mặt đầu vào có phải ảnh giả mạo hay không. Phương pháp phục vụ cho các hệ thống liên quan tới nhận dạng khuôn mặt và xác minh danh tính với mục đích để chống các loại tấn công giả mạo vào hệ thống.



Hình 1

Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến phương pháp phát hiện giả mạo khuôn mặt qua hình ảnh. Cụ thể, phương pháp phát hiện giả mạo khuôn mặt qua hình ảnh được đề cập trong sáng chế sử dụng công nghệ trí tuệ nhân tạo và được ứng dụng trong lĩnh vực xác thực sinh trắc học nhận dạng khuôn mặt (*facial recognition-based biometric authentication*) nhằm chống tấn công giả mạo khuôn mặt (*face anti-spoofing*).

Tình trạng kỹ thuật của sáng chế

Xác thực sinh trắc học (*biometric authentication*) là một quá trình bảo mật dựa trên các đặc điểm sinh trắc học riêng biệt của con người nhằm mục đích xác thực một người có phải là chính họ hay không. Hiện nay, xác thực sinh trắc học được phát triển trên sự đa dạng các yếu tố sinh trắc học của con người như: dấu vân tay (*fingerprints*), võng mạc (*retina*), mống mắt (*iris*), v.v. Các hệ thống xác thực sinh trắc học đang được sử dụng rộng rãi trong các ứng dụng thực tế, ngày càng phải đối mặt với những cuộc tấn công giả mạo nhằm thực hiện hành vi truy cập, xâm phạm trái phép vào hệ thống.

Các hệ thống xác thực khuôn mặt đang trở nên phổ biến trong đời sống nhờ vào trải nghiệm tiện lợi, tiết kiệm thời gian của người sử dụng, tiết kiệm thời gian, nhân lực và chi phí cho doanh nghiệp. Một số sản phẩm xác thực khuôn mặt phổ biến như: chấm công tự động, định danh người dùng điện tử (*electronic Know Your Customer - eKYC*), kiểm soát an ninh, v.v. Việc xác thực khuôn mặt có một số khó khăn nhất định bao gồm điều kiện môi trường, vị trí đặt camera, hướng khuôn mặt, thiết bị thu ảnh, v.v. Đặc biệt, song song cùng với công nghệ xác thực giả mạo khuôn mặt là sự phát hiện đa dạng của công nghệ tấn công giả mạo khuôn mặt, bao gồm tấn công ảnh in màu (*photo*), tấn công hiển thị qua thiết bị điện tử, tấn công video (*replay attack*), tấn công mặt nạ ba chiều (3D), v.v. Trong sáng chế này, chúng tôi đề xuất một phương pháp mới sử dụng công nghệ trí tuệ nhân tạo cho vấn đề phát hiện giả mạo khuôn mặt cải thiện được các khó khăn nhằm đạt được kết quả chính xác cao, thời gian xử lý nhanh.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế là đề xuất phương pháp phát hiện giả mạo khuôn mặt qua hình ảnh phục vụ cho các hệ thống liên quan tới nhận dạng khuôn mặt và xác minh danh tính sử dụng công nghệ trí tuệ nhân tạo. Để đạt được mục đích trên, phương pháp phát hiện giả mạo khuôn mặt qua hình ảnh sử dụng mô hình học sâu bao gồm các bước:

Bước 1: tìm kiếm và căn chỉnh khuôn mặt trong ảnh; bước này thực hiện dựa trên mô hình học sâu RetinaFace nhằm xác định vị trí gương mặt trong ảnh và căn chỉnh gương mặt về phương thẳng đứng.

Bước 2: tiền xử lý ảnh khuôn mặt; bước này nhận đầu vào là tọa độ khuôn mặt được trả về sau khi xử lý bước 1, dựa trên các tọa độ, khuôn mặt sẽ được mở rộng ra theo những tỷ lệ đã xác định trước và cắt ra nhằm phù hợp với từng mô hình tại bước 3.

Bước 3: trích xuất đặc trưng giả mạo khuôn mặt; bước này thực hiện dựa trên sự kết hợp của hai mô hình học sâu phát hiện giả mạo khuôn mặt là MiniFASNet và AENet (*Auxiliary Information Embedding Network*) nhằm trích xuất đa dạng nhiều loại đặc trưng giả mạo trên ảnh.

Bước 4: phân loại đặc trưng và dự đoán khuôn mặt trong ảnh là khuôn mặt thật được chụp thực tế hay khuôn mặt giả mạo; bước này thực hiện dựa trên một mạng nơ-ron học sâu (*Deep neuron network - DNN*) để học các đặc trưng đã trích xuất từ bước 3; khi đưa ảnh khuôn mặt đầu vào bước 1, đầu ra tại bước 4 trả về xác suất giả mạo (%) của khuôn mặt và đưa ra dự đoán.

Mô tả vắn tắt các hình vẽ

Hình 1: là hình ảnh mô tả các bước trong phương pháp phát hiện giả mạo khuôn mặt qua hình ảnh.

Hình 2: là hình ảnh mô tả bước phân loại đặc trưng và dự đoán ảnh khuôn mặt đầu vào có phải ảnh giả mạo hay không.

Mô tả chi tiết sáng chế

Các bước và quy trình các bước của phương pháp phát hiện giả mạo khuôn mặt qua hình ảnh nhằm phục vụ cho các hệ thống xác thực sinh trắc học khuôn mặt được mô tả trong tham chiếu Hình 1. Dữ liệu đầu vào là các ảnh chứa khuôn mặt cần xác thực.

Tuy nhiên, cũng cần phải nói thêm rằng trong sáng chế này, các thuật ngữ “RetinaFace”, “Softmax”, “Cross-entropy”, “Adam”, “MiniFASNet”, “AENet”, “CelebA-Spoof” là các danh từ riêng (tên của mô hình, thuật toán, hàm số, v.v.) 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 và căn chỉnh khuôn mặt trong ảnh.

Bước này được thực hiện dựa trên mô hình học sâu RetinaFace với mục đích tìm kiếm được vùng chứa gương mặt trong ảnh (*bounding box*) và xác định năm vị trí mốc (*landmarks*) của khuôn mặt bao gồm hai mắt, mũi và hai khóe miệng; các vị trí mốc này được sử dụng nhằm xác định hướng xoay hiện tại của khuôn mặt, từ đó xoay mặt về phương thẳng đứng và trả về tọa độ vùng khuôn mặt đã xoay.

Bước 2: tiền xử lý ảnh khuôn mặt;

Mục đích của bước này là xử lý ảnh khuôn mặt sau khi đã cắt và căn chỉnh từ bước 1 sao cho phù hợp làm đầu vào cho từng mô hình trích xuất đặc trưng tại bước 3, do các mô hình ở bước 3 nhận ảnh đầu vào với tỷ lệ kích thước khuôn mặt khác nhau (*multi-scale*). Cụ thể, các ảnh khuôn mặt này được mở rộng ra theo các tỷ lệ khác nhau theo cả hai chiều ảnh để thu thập được nhiều đặc trưng hơn; trong đó, đối với mạng MiniFASNet thì ảnh được mở rộng theo hai tỷ lệ 2,7 và 4,0 theo cả hai chiều, đối với mạng AENet thì ảnh được mở rộng theo một tỉ lệ là 1,17 theo cả hai chiều (các tỷ lệ này được đo bằng thực nghiệm và đã cho kết quả tốt nhất). Kết quả từ bước 2 là các ảnh khuôn mặt đã được mở rộng về không gian theo các tỷ lệ nêu trên, và sẽ là đầu vào ở bước 3.

Bước 3: trích xuất đặc trưng giả mạo khuôn mặt;

Nhiệm vụ của bước này là trích xuất các đặc trưng liên quan đến giả mạo trong ảnh khuôn mặt bao gồm đặc trưng không gian (*spatial features*), đặc trưng tần số (*spectrum features*), đặc trưng hình học (*geometric features*), đặc trưng ngữ nghĩa (*semantic features*) nhằm phục vụ việc phát hiện giả mạo trong ảnh. Đặc trưng không gian là các đặc trưng thể hiện trên ảnh RGB thông thường; đặc trưng tần số là đặc trưng thể hiện trên ảnh sau khi đã phân tích phổ Fourier (*Fourier spectrum*); đặc trưng hình học là những đặc trưng liên quan đến hình khối khuôn mặt được thể hiện qua các phân tích biểu đồ phản xạ (*reflection map*)

và biểu đồ chiều sâu (*depth map*) của khuôn mặt; đặc trưng ngữ nghĩa là những đặc trưng liên quan tới tác động ngoại cảnh sinh ra giả mạo, bao gồm loại giả mạo (ví dụ: ảnh in, mặt nạ 2D, ảnh chụp từ thiết bị khác) và điều kiện ánh sáng và môi trường (ví dụ: trong nhà, ngoài trời, tối, sáng, v.v). Bước này được thực hiện dựa trên hai mạng nơ-ron tích chập (*Convolutional neural network - CNN*) là MiniFASNet và AENet.

Thứ nhất, MiniFASNet là một mạng học sâu được xây dựng với mục đích phát hiện giả mạo khuôn mặt. Mô hình này có cấu trúc rẽ nhánh với sự giám sát của phân tích Fourier. Ngoài nhánh trích xuất đặc trưng chính (*spatial features*), MiniFASNet tách một nhánh riêng nhằm mục đích tái tạo kết quả phân tích phổ Fourier (*Fourier spectrum*) từ ảnh khuôn mặt đầu vào. Phổ Fourier mang lại thông tin đặc trưng giả mạo của ảnh trên miền tần số (*spectrum features*).

Thứ hai, AENet là một mạng học sâu đa tác vụ được sinh ra với mục đích phục vụ bài toán phát hiện giả mạo khuôn mặt. AENet được thiết kế bao gồm ba khối: khối trích xuất đặc trưng hình học (*geometric features*) của khuôn mặt bao gồm các lớp tích chập, khối trích xuất đặc trưng ngữ nghĩa (*semantic features*) của ảnh bao gồm các lớp nơ-ron FC (*fully-connected*), khối phân loại (*classification*) để dự đoán được ảnh là giả mạo hay ảnh thật dựa trên những đặc trưng đã trích xuất ở hai khối trước.

Mô hình MiniFASNet là mô hình chống giả mạo hiệu quả, đã được ứng dụng trên nhiều sản phẩm về nhận dạng khuôn mặt với mục tiêu độ chính xác cao và tốc độ xử lý nhanh. Trong khi đó, AENet là một mô hình được xác minh hiệu quả trên bộ dữ liệu giả mạo khuôn mặt tốt nhất hiện nay CelebA-Spoof Dataset. Nhờ vào thế mạnh trong sự đa dạng của bộ dữ liệu CelebA-Spoof, AENet được xây dựng với mục tiêu khai thác yếu tố ngữ nghĩa, cấu trúc hình học của khuôn mặt trong ảnh. Với mục đích xây dựng phương pháp phát hiện giả mạo khuôn mặt có độ chính xác cao, tốc độ xử lý nhanh, tập trung vào ngữ nghĩa cho các sản phẩm liên quan đến định danh khách hàng, kiểm soát an ninh, nên sáng chế này đề xuất việc kết hợp hai mô hình MiniFASNet và AENet với nhau để phục vụ cho việc trích xuất đa dạng nhiều đặc trưng giả mạo từ ảnh khuôn mặt. Hai mô hình này đã được ứng dụng hiệu quả trong sản phẩm thực tế và đạt kết quả cao trên bộ dữ liệu tốt nhất hiện nay, tuy nhiên chưa có sự kết hợp nào giữa hai mô hình này để gia tăng đặc trưng quan trọng cho việc phát hiện giả mạo. Ảnh khuôn mặt đầu vào đưa lần lượt qua các bước

1, bước 2, bước 3, tại bước 3, hai mô hình MiniFASNet và AENet trả về hai véc-tơ đặc trưng, mỗi véc-tơ là một véc-tơ 512 chiều; hai véc-tơ sau đó được nối thành một véc-tơ đặc trưng 1024 chiều duy nhất.

Bước 4: phân loại đặc trưng và dự đoán ảnh khuôn mặt có phải là ảnh giả mạo hay không;

Sau khi qua bước 3, véc-tơ đặc trưng 1024 chiều sẽ được sử dụng để phân loại ảnh khuôn mặt đầu vào là ảnh thật hay ảnh giả mạo. Một mạng học sâu (*DNN*) với ba tầng được đề xuất dùng để huấn luyện mô hình phân loại, tham chiếu Hình 2. Đầu vào của mạng là véc-tơ với kích thước 1×1024 . Lớp thứ hai của mạng được thiết kế với 32 nơ-ron nhằm nén các đặc trưng đầu vào từ véc-tơ đặc trưng với hàm kích hoạt (*activation function*) tuyến tính hiệu chỉnh (*ReLU*), lớp cuối cùng bao gồm hai nơ-ron tương ứng với hai khả năng xảy ra của kết quả dự đoán: ảnh thật, ảnh giả mạo. Đầu ra của mạng phân loại này là giá trị của hai nơ-ron tại lớp cuối cùng, sau đó được đưa qua hàm kích hoạt Softmax, tại nơ-ron đầu tiên ta nhận về xác suất khuôn mặt trong ảnh đầu vào là ảnh thật. Các tham số và siêu tham số (*hyper-parameters*) bao gồm số lượng vòng lặp huấn luyện là 100 (*epoch*), kích thước lô huấn luyện (*batch size*) là 128, hàm mất mát (*loss function*) là Cross-Entropy được tối ưu hóa bởi thuật toán tối ưu Adam với các tham số: tốc độ học (*learning rate*) là 5×10^{-4} , các tốc độ phân rã (*exponential decay*) lần lượt là $\beta_1 = 0,9$, $\beta_2 = 0,999$, hệ số chính quy hóa L_2 (L_2 regularization) là 5×10^{-4} , tham số ổn định là $\varepsilon = 10^{-8}$.

Sau khi mạng học sâu (*DNN*) trả về đầu ra là giá trị của hai nơ-ron tại lớp cuối, dựa vào giá trị của nơ-ron đầu tiên, ta nhận được xác suất (%) khuôn mặt trong ảnh đầu vào là ảnh thật. Một ngưỡng phân loại được đặt ra để quyết định ảnh khuôn mặt đầu vào là ảnh thật hay ảnh giả mạo. Trong sáng chế này, ngưỡng được đề xuất dựa trên thực nghiệm cho kết quả tốt nhất là 65%.

Mặc dù các mô tả nêu trên chứa nhiều chi tiết cụ thể, chúng không được coi là giới hạn về phương án thực thi sáng chế, mà chỉ nhằm mục đích minh họa một số phương án thực thi được ưu tiên.

Yêu cầu bảo hộ

1. Phương pháp phát hiện giả mạo khuôn mặt qua hình ảnh sử dụng công nghệ trí tuệ nhân tạo, bao gồm các bước sau:

bước 1: tìm kiếm và căn chỉnh khuôn mặt;

tại bước này, mô hình học sâu RetinaFace được sử dụng để tìm ra vùng chứa khuôn mặt (*bounding box*) và xác định năm vị trí mốc (*landmarks*) của khuôn mặt; từ đó căn chỉnh khuôn mặt về phương thẳng đứng và trả về tọa độ vùng chứa khuôn mặt đã căn chỉnh để phục vụ cho bước sau;

bước 2: tiền xử lý ảnh khuôn mặt;

tại bước này, vùng chứa khuôn mặt được mở rộng theo nhiều tỷ lệ (*multi-scale*) nhằm tăng cường vùng ảnh chứa đặc trưng và phù hợp với từng mô hình được sử dụng tại bước 3;

bước 3: trích xuất đặc trưng giả mạo khuôn mặt;

tại bước này, ảnh khuôn mặt sau khi tiền xử lý được đưa qua hai mô hình học sâu MiniFASNet và AENet nhằm mục đích trích xuất các đặc trưng giả mạo trong ảnh; bước này kết hợp hai mô hình MiniFASNet và AENet nhằm mục đích trích xuất đa dạng các loại đặc trưng bao gồm đặc trưng không gian (*spatial features*), đặc trưng tần số (*spectrum features*), đặc trưng hình học (*geometric features*), đặc trưng ngữ nghĩa (*semantic features*); hai mô hình trả về hai véc-tơ đặc trưng cùng có 512 chiều, véc-tơ đặc trưng 1024 chiều duy nhất được tạo ra bằng cách nối hai véc-tơ đặc trưng đó; cụ thể: nhằm trích xuất được đặc trưng giả mạo từ ảnh đầu vào một cách phong phú, phương pháp đề xuất kết hợp hai mô hình MiniFASNet và AENet; MiniFASNet là một mạng học sâu được thiết kế với mục đích nâng cao khả năng phát hiện giả mạo khuôn mặt bằng cách sử dụng đặc trưng giả mạo trên miền tần số; ảnh khuôn mặt được đưa về miền tần số sẽ làm nổi bật những đặc trưng giả mạo mà không gian ảnh thông thường khó có thể nhận biết; AENet là một mạng học sâu đa tác vụ được thiết kế nhằm thử nghiệm trên bộ dữ liệu tốt nhất hiện nay CelebA-Spoof, AENet khai thác các đặc trưng trên không gian ảnh, đặc trưng về hình học của khuôn mặt, đặc trưng về ngữ nghĩa của ảnh và đã có kết quả tốt trên bộ dữ liệu CelebA-

Spoof; nhận đầu vào là ảnh khuôn mặt, kết quả tại bước 3 sau khi đưa qua hai mô hình trích xuất đặc trưng MiniFASNet và AENet thu về hai véc-tơ đặc trưng 512 chiều; hai véc-tơ này sau đó được nối lại thành một véc-tơ đặc trưng 1024 chiều duy nhất nhằm sử dụng cho việc phân loại đặc trưng;

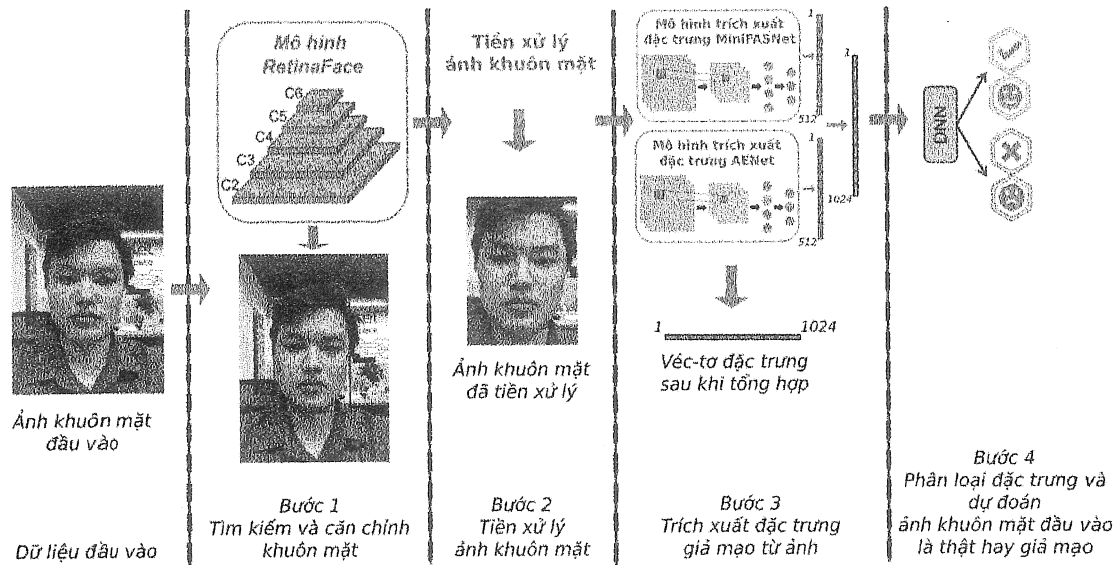
bước 4: phân loại đặc trưng và dự đoán ảnh khuôn mặt đầu vào có phải là ảnh giả mạo hay không;

tại bước này, một mạng học sâu (DNN) với ba lớp được đề xuất sử dụng: đầu vào của mạng là các véc-tơ đặc trưng 1024 chiều, tầng thứ hai của mạng được xây dựng với 32 nơ-ron nhằm nén các đặc trưng từ véc-tơ đặc trưng đầu vào với hàm kích hoạt tuyến tính (*activation function*) hiệu chỉnh (*ReLU*), đầu ra của mạng là giá trị của hai nơ-ron tại lớp cuối với hàm kích hoạt Softmax tương ứng với khả năng xảy ra của hai trường hợp ảnh thật và ảnh giả mạo; từ giá trị của nơ-ron đầu tiên, xác suất khuôn mặt là ảnh thật được đưa ra (%), một ngưỡng phân loại sẽ được đặt ra để quyết định ảnh khuôn mặt đầu vào là ảnh thật hay ảnh giả mạo.

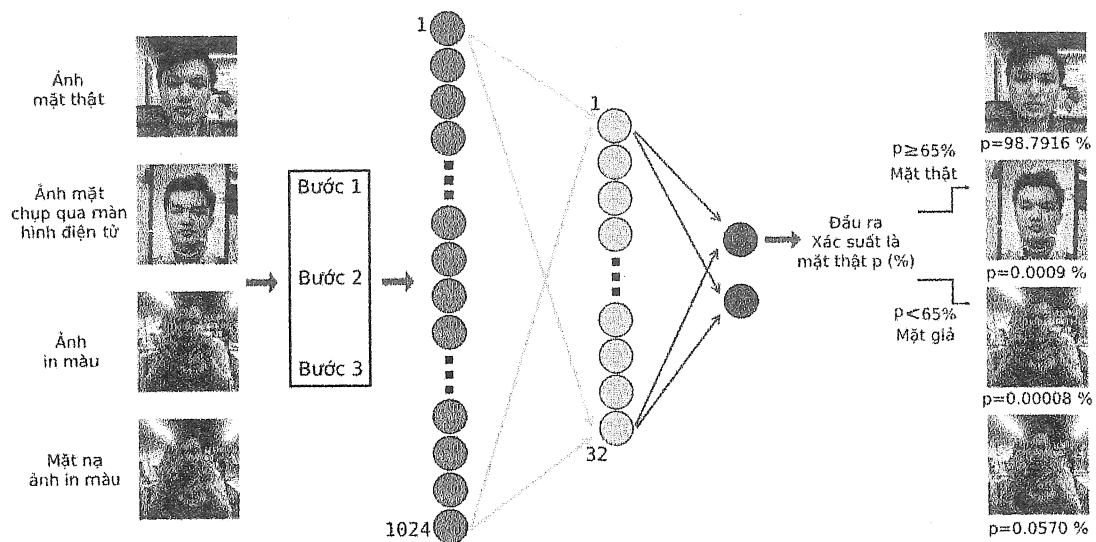
2. Phương pháp phát hiện giả mạo khuôn mặt qua hình ảnh sử dụng công nghệ trí tuệ nhân tạo theo điểm 1, trong đó:

ở bước 4, để phân loại đặc trưng giả mạo thu được từ bước 3, một mạng học sâu (DNN) gồm ba lớp được sử dụng với các tham số và siêu tham số như sau: số vòng lặp huấn luyện 100 (*epoch*), kích thước lô huấn luyện (*batch size*) là 128, hàm mất mát (*loss function*) là cross-entropy, thuật toán tối ưu hóa adam: tốc độ học (*learning rate*) là 5×10^{-4} , các tốc độ phân rã (*exponential decay*) lần lượt là $\beta_1 = 0,9$, $\beta_2 = 0,999$, hệ số chính quy hóa L_2 (L_2 regularization) là 5×10^{-4} , tham số ổn định là $\epsilon = 10^{-8}$.

3. Phương pháp phát hiện giả mạo khuôn mặt qua hình ảnh sử dụng công nghệ trí tuệ nhân tạo theo điểm 1, trong đó, tại bước 3, ngưỡng phân loại được đề xuất dựa trên thực nghiệm cho kết quả tốt nhất là 65%.



Hình 1



Hình 2