Phát Hiện Khuôn Mặt Đeo Khẩu Trang Trực Tiếp Từ Webcam

Đinh Trọng Đạt 1914775, Nguyễn Trung Nguyên 1914626, Mai Thanh Lâm 1911162

*Lớp CTK43, Khoa Công nghệ Thông tin, Trường Đại học Đà Lạt*

Tóm tắt – Bài báo này đề xuất một mô hình hiệu quả để giải quyết bài toán nhận dạng khuôn mặt có đeo khẩu trang hay không, trực tiếp từ hệ thống webcam. Trong đó, bài báo tập trung chính vào công đoạn: Phát hiện và nhận dạng khuôn mặt từ màn hình webcam. Phương pháp phát hiện khuôn mặt được đề xuất sử dụng ngôn ngữ lập trình Python với thư viện Keras, OpenCV và phương pháp CNN. Mô hình nhận dạng khuôn mặt được đề xuất trên cơ sở mô hình mạng neural để trích chọn đặc trưng khuôn mặt. Hiệu quả của mô hình nhận dạng được kiểm nghiệm trên các tập cơ sở dữ liệu chuẩn. Các kết quả thực nghiệm cho thấy mô hình đề xuất đạt độ chính xác cao.

Từ khóa – **real-time facemask detector, face recognition, nhận dạng, khung hình, phát hiện khuôn mặt, phân tích tập mẫu.**

1. GIỚI THIỆU

Trên thế giới, bài toán nhận dạng sinh trắc học nói chung và nhận dạng khuôn mặt nói riêng đã được đầu tư nghiên cứu từ vài chục năm về trước và đã thu được nhiều kết quả về cả lý thuyết lẫn ứng dụng thực tiễn. Hiện nay các công nghệ nhận dạng sinh trắc học không chỉ sử dụng để xác nhận nhân thân mà còn được dung trong rất nhiều bài toán thực tiễn như kiểm soát vào / ra, kiểm soát truy cập mạng, đảm bảo mức độ an ninh, hỗ trợ tự động hóa chấm công,…

Hiện nay, đại dịch Covid19 hoành hành khiến nền kinh tế thế giới và tại Việt Nam bị ảnh hưởng nặng nề, nhiều người đã phải bỏ mạng vì dịch bệnh,… Bắt buộc chúng ta phải đeo khẩu trang khi đi tới những nơi công cộng để bảo vệ bản thân và bảo vệ những người xung quanh. Bài toán được đặt ra là mô hình phát hiện khuôn mặt của một người nào đó có đeo khẩu trang hay không.

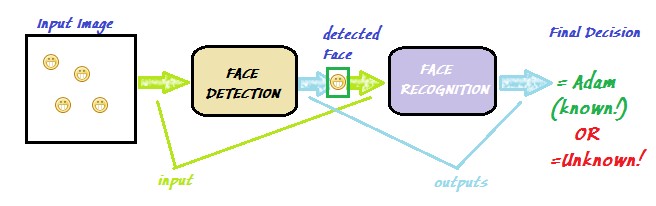
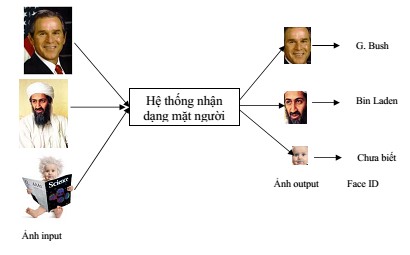
Trong bài báo này, chúng tôi đề xuất một giải pháp tổng thể để giải quyết bài toán phát hiện đeo khẩu trang trực tiếp từ các thiết bị camera/webcam, hướng tới mục tiêu xây dựng hệ thống camera giám sát, kiểm soát vào/ra, phát hiện đối tượng không đeo khẩu trang ở nơi công cộng,… Trong đó, việc cải thiện chất lượng nhân dạng được tập trung ở hai công đoạn chính là phát hiện khuôn mặt trực tiếp từ các khung hình và nhận dạng các khuôn mặt người đã được phát hiện.

Bài báo này được viết theo cấu trúc như sau:

* Chương 1:.Các hướng tiếp cận.
* Chương 2: Đề xuất mô hình phát hiện đeo khẩu trang từ webcam.
* Chương 3: Xây dựng chương trình-mô phỏng
* Chương 4: Đánh giá thực nghiệm.
* Chương 5: Kết luận.

1. CÁC HƯỚNG TIẾP CẬN

Nhận dạng khuôn mặt người là quá trình xác định danh tính tự động cho từng đối tượng người trong ảnh/video dựa vào nội dung. Rất nhiều hướng tiếp cận đã được đề xuất để giải quyết bài toán này [1],[2],[5],[8]. Nhìn chung, quy trình giải quyết bài toán thường bao gồm các công đoạn cơ bản như: (i) Thu thập hình ảnh; (ii) Tiền xử lý, tang cường chất lượng hình ảnh;(iii) Phát hiện, căn chỉnh, crop ảnh khuôn mặt; (iv) Nhận dạng (trích chọn đặc trưng và phân lớp) khuôn mặt



**Hình 1**.Mô hình nhận dạng khuôn mặt

Bài toán nhận dạng mặt người là bài toán đã được nghiên cứu từ những năm 70. Tuy nhiên, đây là một bài toán khó, nên những nghiên cứu hiện tại vẫn chưa đạt được những kết quả mong muốn. Chính vì thế, vấn đề này vẫn đang được nhiều nhóm trên thế giới quan tâm nghiên cứu. Khó khan của bài toán nhận dạng mặt người có thể kể đến như sau:

* Tư thế chụp, góc chụp: Ảnh chụp khuôn mặt có thể thay đổi rất nhiều bởi vì góc chụp giữa camera và khuôn mặt. Chẳng hạn như: chụp thẳng, chụp chéo bên trái 45° hay chụp chéo bên trái 45°, chụp từ trên xuống, chụp từ dưới lên, v.v… Với các tư thế khác nhau, các thành phần trên khuôn mặt như mắt có thể bị khuất một phần hoặc thậm chí khuất hết.
* Sự xuất hiện hoặc thiếu một số thành phần của khuôn mặt: Các đặc trưng như mắt kính, nón, v.v… có thể xuất hiện hoặc không. Vấn đề này làm cho bài toán càng trở nên khó hơn rất nhiều.
* Sự biểu cảm của khuôn mặt: Biểu cảm của khuôn mặt con người có thể làm ảnh hưởng đáng kể lên các thông số của khuôn mặt. Chẳng hạn, cùng một khuôn mặt một người, nhưng có thể sẽ rất khác khi họ cười hoặc sợ hãi, v.v…
* Sự che khuất: Khuôn mặt có thể bị che khuất bởi các đối tượng khác hoặc các khuôn mặt khác.
* Hướng của ảnh (pose variations): Các ảnh khuôn mặt có thể biến đổi rất nhiều với các góc quay khác nhau của trục camera. Chẳng hạn chụp với trục máy ảnh nghiêng làm cho khuôn mặt bị nghiêng so với trục của ảnh.
* Điều kiện của ảnh: Ảnh được chụp trong các điều kiện khác nhau về : chiếu sáng, về tính chất camera (máy kỹ thuật số, máy hồng ngoại, v.v…), ảnh có chất lượng thấp ảnh hưởng rất nhiều đến chất lượng ảnh khuôn mặt.
* Aging condition: Việc nhận dạng ảnh mặt thay đổi theo thời gian còn là một vấn đề khó khan, ngay cả đối với khả năng nhận dạng của con người.
* Các hệ thống cực lớn: Các CSDL ảnh mặt được test bởi các nhà nghiên cứu còn khá nhỏ (vài trăm tới vài chục nghìn ảnh mặt), tuy nhiên trên thực tế các CSDL có thể rất lớn, ví dụ CSLD ảnh mặt của cảnh sát của một đất nước có thể chứa từ hang triệu tới hơn 1 tỉ ảnh…

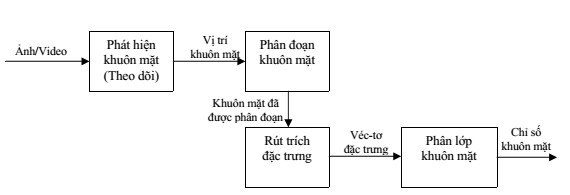
Một hệ thống nhận dạng mặt người thông thường bao gồm bốn bước xử lý sau:

1.Phát hiện khuôn mặt (Face Detection)

2.Phân đoạn khuôn mặt (Face Alignment hay Segmentation).

3.Trích chọn đặc trưng (Feature Extraction).

4.Nhận dạng (Recognition) hay Phân lớp khuôn mặt (Face Clasaification).



Phát hiện khuôn mặt dò tìm, định vị những vùng (vị trí) có thể là khuôn mặt xuất hiện trong ảnh hoặc các frame video. Các vùng này sẽ được tacgs riêng để xử lý. Phân đoạn khuôn mặt sẽ xác định vị trí mắt mũi, miệng và các thành phần khác của khuôn mặt và chuyển kết quả này cho bước trích chọn đặc trưng. Ở bước trích chọn đặc trưng, bằng một phương pháp trích chọn đặc điểm nào đó (mẫu nhị phân cục bộ-Local Binary Pattern-LBP, Gabor wavelets…) sẽ được sử dụng với ánh mặt để trích xuất các thông tin đặc trưng cho ảnh từ các thông tin về các thành phần trên khuôn mặt, kết quả là mỗi ảnh sẽ được biểu diễn dưới dạng một vector đặc trưng. Những vector đặc trưng này sẽ là dữ liệu đầu vào cho một mô hình đã được huấn luyện trước để nhận dạng khuôn mặt (Face Recognition), thường thì phương pháp k-láng giềng gần sẽ được sử dụng.

Bên cạnh những bước chính nêu trên, chúng ta còn có thẻ áp dụng them một số bước khác như tiền xử lý, hậu xử lý nhằm làm tang độ chính xác cho hệ thống. Ví dụ, sau bước phát hiện khuôn mặt, ta có thể thực hiện bước tiền xử lý (Preprocessing) bao gồm các bước căn chỉnh ảnh và chuẩn hóa ánh sáng.

Do một số thông số như: tư thế khuôn mặt, độ sáng, điều kiện ánh sáng,v.v… phát hiện khuôn mặt được đánh giá là bước khó khan và quan trọng nhất so với các bước còn lại của hệ thống. Các bước phát hiện khuôn mặt sẽ được tìm hiểu sâu trong phần sau.

Dữ liệu cho một hệ thống phát hiện khuôn mặt được chia làm 3 tập: tập huấn luyện (training set), tập tham chiếu (reference set) và tập để nhận dạng (probe set hay query set, đôi khi còn gọi là test set). Trong nhiều hệ thống, tập training trùng với tập reference. Tập training gồm các ảnh được dung để huấn luyện (hay học-learing), thông thường, tập này được dung để sinh ra một không gian con là một ma trận và phương pháp hay được sử dụng là PCA (Principal Component Analysis), WPCA (Whitened PCA), LDA (Linear Discriminant Analysis), KPCA (Kernel PCA). Tập reference gồm các ảnh đã biết danh tính được chiếu (projected) vào không gian con ở bước training. Bước training nhằm 2 mục đích: giảm số chiều (dimension reduction) của các vector đặc điểm (feature vector) vì các vector này thường có độ dài khá lớn (vài nghìn tới vài tram nghìn) nên nếu để nguyên thì việc tính toán sẽ rất rất lâu, thứ hai là làm tang tính phân biệt (discriminative) giữa các ảnh khác lớp (định danh khác nhau), ngoài ra có thể làm giảm tính phân biệt giữa các ảnh thuộc về mọt lớp (tùy theo phương pháp) ví dụ như Linear Discriminant Analysis LDA- còn gọi là *Fisher Linear Discriminant Analysis-Fisherface* là một phương pháp làm việc với tập training mà mỗi đối tượng có nhiều ảnh mặt ở các điều kiện khác nhau). Sau khi thực hiện chiếu tập reference vào không gian con, hệ thống lưu lại kết quả là một ma trận với mỗi cột của ma trận là một vector tương ứng với ảnh (định danh đã biết) để thực hiện nhận dạng (hay phân lớp). Nhận dạng (hay phân lớp) được thực hiện với tập các ảnh *probe*, sau khi tiền xử lý xong, mỗi ảnh sẽ được áp dụng phương pháp trích chọn đặc điểm (như với các ảnh thuộc tập *training* và *reference*) và được chiếu vào không gian con. Tiếp đến việc phân lớp sẽ dựa trên phương pháp k-NN, định danh của một ảnh cần xác định sẽ được gán là định danh của ảnh có khoảng cách *(distance)* gần với nó nhất. Ở đây cần lưu ý là mỗi ảnh là một vector nên có thể dùng khái niệm hàm khoảng cách giữa hai vector để đo sự khác biệt giữa các ảnh.

Các hướng tiếp cận trước đây chủ yếu dựa trên đặc trưng (feature-based) và luôn cố gắng đưa ra các định nghĩa tường minh để biểu diễn khuôn mặt dựa trên các tỷ lệ khoảng cách, diện tích và góc. Một biểu diễn khuôn mặt được định nghĩa tường minh hướng tới mục tiêu xây dựng một không gian đặc trưng trực quan. Tuy nhiên, trong thực tế các biểu diễn được định nghĩa tường minh thường không chính xác. Để khắc phục điều đó, các hướng tiếp cận sau này được đề xuất dựa trên ý tưởng sử dụng các mô hình học máy thống kê có khả năng học để lựa chọn các đặc trưng khuôn mặt từ một tập mẫu cho trước, điển hình như phương pháp sử dụng các mô hình mạng neural tích chập CNN (Convolutional Neural Network).

Hiện tại, hiệu quả của các mô hình nhận dạng khuôn mặt đã được cải thiện đáng kể dựa trên việc kết hợp sử dụng các mô hình học sâu để tự động phát hiện các đặc trưng trên khuôn mặt và các ký thuật phân lớp thống kê.

Nghiên cứu của Md.Shahriar Islam [4] về sử dụng CNN, phát triển nhận dạng khuôn mặt với khẩu trang ở trung tâm khuôn mặt, phát hiện khuôn mặt được che một cách an toàn và để thiết lập trình giám sát an ninh y tế tốt hơn. Sau đó Lars Ankile, Morgan Heggland, Kjartan [5] thiết kế một giải pháp cho vấn đề nhận dạng khuôn mặt bằng cách sử dụng mạng neural phức hợp, với mục đích áp dụng giải pháp này trong hệ thống kiểm soát truy cập vào nhà dựa trên camera, đạt được độ chính xác trên 99% đối với thử nghiệm. CUngx trong năm 2020, Hafidz Sanjaya [6] nghiên cứu việc ứng dụng hệ thống nhận dạng khuôn mặt trong thời gian thực bằng thuật toán Tiny Yolo V2 trong hệ thống chấm công trực tuyến với tỷ lệ thành công là 97%. Sau đó Alwin Fau [7] đã cải tiến nghiên cứu, sử dụng Machine Learning kết hợp với máy ảnh để chụp khuôn mặt của một người nào đó sau đó được so sánh với khuôn mặt đã được lưu trữ trước đó trong một cơ sở dữ liệu cụ thể. Tỷ lệ thành công là 99%.

Nhìn chung, các ứng dụng nhận dạng khuôn mặt thường mong muốn tìm được một biểu diễn ít chiều, có khả năng mới mà mạng chưa được huấn luyện bao giờ. Để giải quyết bài toán nhận dạng đeo khẩu trang sử dụng thư viện OpenCV, chúng ta có thể sử dụng các ngôn ngữ lập trình như: .NET C#, Python, Java, C++… Trong đồ án này, ngôn ngữ lập trình được sử dụng là Python, viết trên phần mềm Subline Text.

1. ĐỀ XUẤT MÔ HÌNH NHẬN DẠNG ĐEO KHẨU TRANG TỪ WEBCAM

Thực tế cho thấy, việc nhận dạng đối tượng nói chung và nhận dạng khuôn mặt đeo khẩu trang nói riêng trực tiếp từ hệ thống camera giám sát hoặc webcam hiện vẫn là một bài toán phức tạp, còn nhiều khó khan, thách thức

Webcam thu nhận hình ảnh

Phân đoạn video

Frame

Phát hiện mặt người

Khuôn mặt

Tiền xử lý,

Căn chỉnh khuôn mặt

Khuôn mặt sau căn chỉnh

Nhận dạng khuôn mặt, khẩu trang trên khuôn mặt

Hình 2

**Hình 2.**Phương pháp phát hiện đeo khẩu trang từ webcam

Một trong những thách thức điển hình của bài toán này là hình ảnh khuôn mặt của đối tượng chuyển động và thay đổi liên tục với nhiều tư thế góc nghiêng / xoay và trạng thái khác nhau. Điều này đòi hỏi các thuật toán nhận dạng phải có khả năng tổng quát hóa, không bị ảnh hưởng nhiều bởi độ nghiêng / xoay, và dịch chuyển của đối tượng. Ngoài ra, việc nhận dạng trực tiếp từ camera/webcam luôn đòi hỏi phải đáp ứng được tính thời gian thực (real time) từ webcam hoặc camera được đề xuất cụ thể trên Hình 2.

Từ tín hiệu video đầu vào, bước xử lý đầu tiên sẽ tiến hành phân đoạn video thành các khung hình (frame) riêng biệt. Việc phân đoạn video ở đây được tiến hành theo thời gian (ngưỡng được chọn hiện tại là 24 khung hình trên giây). Mỗi khung hình có thể không chứa, chứa một phần hoặc chứa toàn bộ khuôn mặt. Vì vậy, trong bước xử lý đầu tiên, thuật toán sẽ tiến hành phát hiện (face detection) và xác định vị trí các khuôn mặt (nếu có) trên ảnh. Các khuôn mặt phát hiện được sau đó sẽ tiếp tục được tiền xử lý nhằm tăng cường chất lượng hình ảnh (loại nhiễu, khử bóng/ mờ), chuẩn hóa kích thước và độ phân giải ảnh, căn chính khuôn mặt về hướng trực hiện (nhìn thẳng). Các khuôn mặt sau khi đã tiền xử lý sẽ được sử dụng làm đầu vào cho một mô hình mạng neral học sâu (DNN-Deep Neural Network). Mô hình này sẽ tự động học và trích chọn ra các đặc trưng để nhận dạng (phân lớp) khuôn mặt. Bước xử lý cuối của thuật toán sẽ tiến hành phân lớp (nhận diện) các khuôn mặt có đeo khẩu trang hay không. Bản chất của việc phân lớp khuôn mặt là tìm kiếm đối tượng người có mẫu khuôn mặt đeo khẩu trang tượng tự với khuôn mặt đang cần phát hiện khẩu trang. Để thực hiện được điều này, các mô hình phân lớp cần phải được huấn luyện với một tập mẫu cho trước. Trong đó, mỗi mẫu khuôn mặt được thể hiện đặc trưng DNN ở bước trên.

A.. Phát hiện khuôn mặt trên khung hình

Như đã đề cập ở trên, bản chất của việc phát hiện khuôn mặt là quá trình tìm kiếm và định vị khuôn mặt trên frame ảnh bất kỳ. Phương pháp phát hiện khuôn mặt ở đây được đề xuất sử dụng các đặc trưng HOG (Histograms of Oriented Gradients) và bộ phận lớp tuến tính SVM (Support Vector Machines).

Gom nhóm đặc trưng tại mỗi cell

Tính đạo hàm gradient

Frame

Có mặt

Không mặt

Phân lớp tuyến tính SVM

Tính đặc trưng trên các cửa sổ detector

Tính đặc trung cho khối, chuẩn hóa

Hình 3. Phương pháp phát hiện khuôn mặt

Ý tưởng chính của đặc trưng HOG là hình dạng và trạng thái của vật thể được đặc trưng bởi sự phân bố về gradient và hướng của cạnh. Đặc trưng này được phát triển dựa trên các đặc trưng SIFT (Scale-Invariant Feature Transform), đặc trưng HOG được tính trên cả một vùng, Do sự biến thiên màu sắc trong các vùng khác nhau nên mỗi vùng sẽ cho ta một vector đặc trưng của nó. Vì vậy để có được đặc trưng của toàn bộ cửa sổ (window) ta phải kết hợp nhiều vùng liên tiếp lại với nhau. Các bước cơ bản trong quy trình phát hiện khuôn ặt người trên các khung hình được mô tả cụ thể trên Hình 3.

Đầu vào của thuật toán là một frame ảnh bất kỳ thu được từ bước phân đoạn video. Bước xử lý đầu tiên sẽ tiến hành chuyển đổi ảnh trong không gian RGB (ảnh màu) sang ảnh đa cấp xám (gray scale), sau đó tiến hành cân bằng histogram trên ảnh gray scale để giảm sự nhạy cảm với nguồn sáng.

Việc lưu trữ chính xác từng giá trị góc của từng điểm ảnh (x,y) tốn nhiều chi phí và không mang lại nhiều kết quả, do vậy ta sẽ chia không gian góc ra thành các bin. Việc phân chia bin càng nhỏ sẽ càng làm tang độ chính xác, các kết quả thự nghiệm cho thấy kích thước bin khoảng 200 cho kết quả tốt nhất đối với việc phát hiện khuôn mặt người. Do đó với không gian hướng biến thiên trong miền từ 0° - 180° sẽ được chia thành 9 bi như sau: [0° -20°], [21° - 40°], [41° - 60°], [61° - 80°], [81° - 100°], [101° - 120°], [121° - 140°], [141° - 160°], [161° - 180°]. Uứng với mỗi bin trên, tiến hành thống kê biên độ (magnitude) tại từng vị trí, Với mỗi bin, tại vị trí (x,y) nếu góc (orientation) thuộc về bin đó thì giá trị của bin đó tại vị trí (x,y) bằng giá trị biên độ, ngược lại giá trị bin tại vị trí (x,y) bằng 0. Bước tiếp theo tiến hành tính toán vector đặc trưng cho từng cell (mỗi cell thường được chọn với kích thước 8x8 px). Vector đặc trưng của mỗi cell sẽ gồm 9 thành phần tương ứng với 9 bin và giá trị tại thành phần I bằng tổng giá trị của các điểm trong bin I mà có tọa độ nằm trong cell đó. Tiếp theo, tính toán vector đặc trưng của khối được tính bằng cách ghép vector đặc trưng của từng cell trong block lại với nhau.

Với giả thiết mỗi cell có kích thước 8x8 px, mỗi block có kích thước 2x2 cells (16x16 px), không gian hướng biến thiên xét trong miền từ 0° - 180° và được chia thành 9 bin thì số đặc trưng trong mỗi khối sẽ được tính bằng 4x9=36 thành phần. Từ đó, tiến hành tính toán vector đặc trưng của các cửa sổ trên toàn bộ ảnh đầu vào. Trong đó, một cửa số (Window) được tạo bởi các khối xếp gối nhau – overlapping. Đặc trưng của một cửa sổ sẽ được tính bằng cách ghép các vector đặc trưng của từng block tạo lên cửa sổ đó.

Ở bước xử lý cuối cùng, toàn bộ vector đặc trưng thu được trên mỗi cửa sổ sẽ được sử dụng làm đầu vào của bộ phận lớp tuyến tính SVM. Bộ phận lớp có nhiệm vụ xác định lớp mẫu (có chứa khuôn mặt hay không chứa khuôn măt) đối với mỗi ảnh đầu vào dựa trên các tri thức mà thuật toán đã được huấn luyện.

B. Phát hiện khuôn mặt đeo khẩu trang bằng phương pháp CNN

Công đoạn phát hiện nhận dạng thường gồm 2 bước xử lý chính là trích chọn đặc trưng và phân lớp khuôn mặt.Phương pháp trích chọn đặc trưng ở đây được đề xuất sử dụng các lớp mạng neural tích chập (CNN – Convolutional Neural Network), là một tập hợp con của học máy và chúng là trung tâm của các thuật toán học sâu. Chúng bao gồm các lớp nút, chứa một lớp đầu vào, một hoặc nhiều lớp ẩn và một lớp đầu ra. Mỗi nút kết nối với một nút khác và có trọng lượng và ngưỡng liên quan. Nếu đầu ra của bất kỳ nút riêng lẻ nào vượt quá giá trị ngưỡng được chỉ định, nút đó sẽ được kích hoạt, gửi dữ liệu đến lớp tiếp theo của mạng. Nếu không có dữ liệu nào được truyền đến lớp tiếp theo của mạng. Mạng neural tích chập thường được sử dụng để phân loại các tác vụ thị giác máy tính. Trước CNN, các phương pháp trích xuất tính năng thủ công, tốn thời gian đã được sử dụng để xác định các đối tượng trong hình ảnh. Tuy nhiên, các mạng neural tích chập hiện cung cấp một cách tiếp cận có thể mở rộng hơn đối với các tác vụ phân loại hình ảnh và nhận dạng đối tượng, tận dụng các nguyên tắc từ đại số tuyến tính, đặc biệt là phép nhân ma trận, để xác định các mẫu trong hình ảnh. Điều đó nói rằng, chúng có thể đòi hỏi tính toán, yêu cầu các đơn vị xử lý đồ họa (GPU) đê đào tạo các mô hình.

Đây là mô hình có khả năng học từ một tập mẫu cho trước, nhằm tự dộng phát hiện các ddwacj trưng quan trọng nhất để nhận dạng đối tượng(ở đây là khuôn mwajt đeo khẩu trang). Ý tưởng chính của hướng tiếp cận này dựa trên việc học một không gian Euclidean nhúng trong mỗi ảnh sử dụng một cấu hình mạng neural tích chập. Mạng được huấn luyện sao cho khoảng cách 1,2 bình phương trong không gian nhúng là tương ứng trực tiếp với độ tương tự của khuôn mặt.

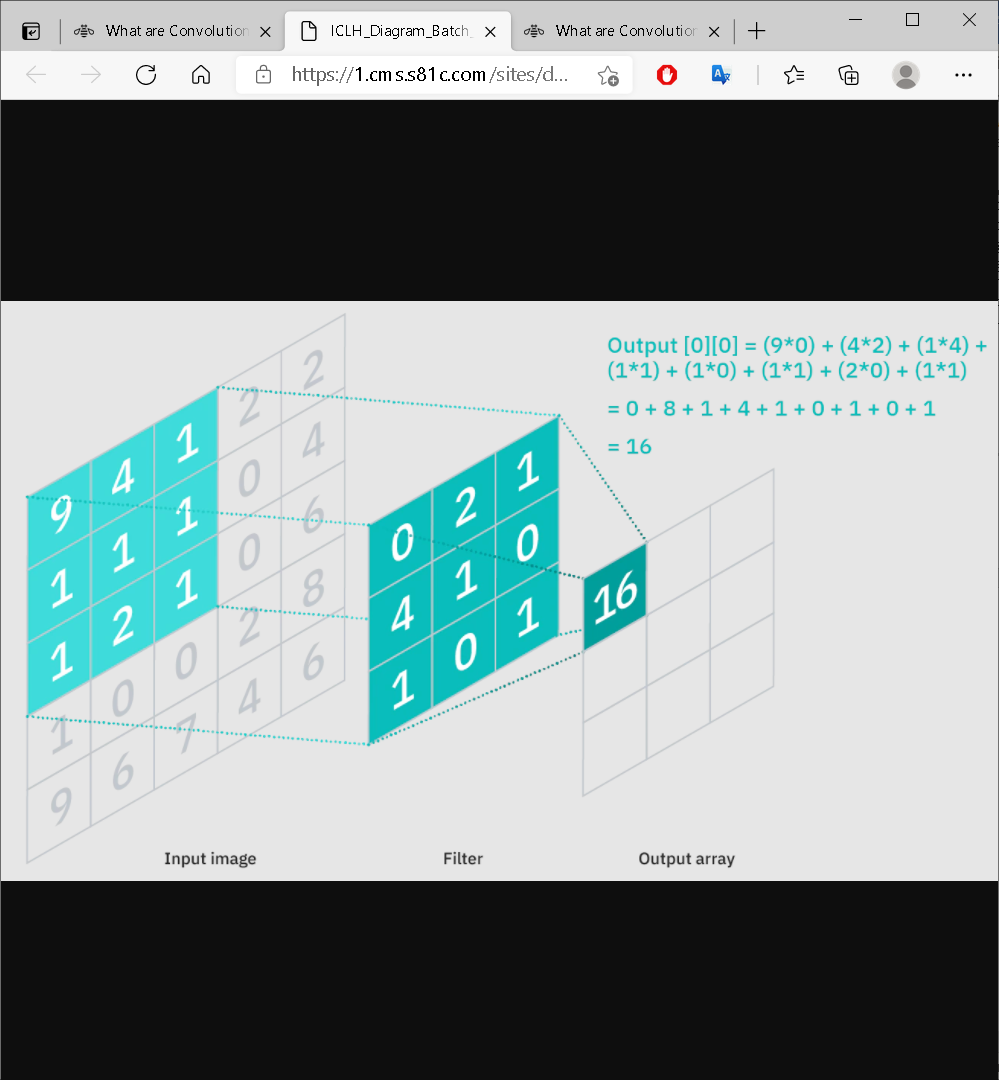
CNN được phân biệt với các mạng neural khác bởi hiệu suất vượt trội của chúng với đầu vào tín hiệu hình ảnh, lời nói hoặc âm thanh. Chúng có ba loại lớ chihs, đó là:

* Convolutional layer
* Pooling layer
* Fully-connected layer

Convolutional layer là lớp đầu tiên của mạng tích chập. Mặc dù các lớp tích chập có thể được theo sau bởi các lớp tích chập bổ sung hoặc các lớp gộp chung, lớp được kết nối đầy đủ là lớp cuối cùng. Với mỗi lớp, CNN tang độ phức tạp của nó, xác định các phần lớp hơn của hình ảnh. Các lớp trước đó tập trung vào các tính nawg đơn giản, chẳng hạn như màu sắc và cạnh. Khi dữ liệu hình ảnh tiến triển qua các lớp của CNN, nó bắt đầu nhận ra các phần tử hoặc hình dạng lớn hơn của đối tượng cho đến khi cuối cùng nó xác định đối tượng dự định.

Convolutional layer là khối xây dựng cốt lõi của CNN và đó là nơi xẩy ra phần lớn tính toán. Nó yêu cầu một vài thành phần, đó là dữ liệu đầu vào, bộ lọc và bản đồ tính năng. Gỉa sử rằng đầu vào sẽ là một hình ahr màu, được tạo thành từ một ma trận các điểm ảnh trong 3D. Điều này có nghĩa là đầu vào sẽ có ba chiều-chiều cao, chiều rộng và độ sâu-tương ứng với RGB trong ảnh. CHúng tôi cũng có một máy dò tính năng, còn được gọi là hạt nhân hoặc bộ lọc, sẽ di chuyển qua các trường tiếp nhận của hình ảnh, kiểm tra xem tính năng này có hiện diện không. Qúa trình này được gọi là một Convolutional.

Máy dò chức năng là một mảng trọng lượng hai chiều (2-D), đại diện cho một phần của hình ảnh. Mặc dù chúng có thể khác nhau về kích thước, kích thước bộ lọc thường là ma trận 3x3; điều này cũng xác định kích thước của trường tiếp nhận. Bộ lọc sau đó được áp dụng cho một khu vực của hình ảnh và một sản phẩm dấu chấm được tính toán giữa các điểm ảnh đầu vào và bộ lọc. Sản phẩm này sau đó được đưa cho một mảng đầu ra. Sau đó, bộ lọc thay đổi theo một sải bước, lặp lại quá trình cho đến khi hạt nhân quét qua toàn bộ hình ảnh. Đầu racuoois cùng từ loạt sản phẩm dấu chấm từ đầu vào và bộ lọc được gọi là bản đồ tính năng, bản đồ kích hoạt hoặc tính nawg xoay.



Hình 4.Mô hình CNN

Như bạn có thể thấy trong Hình 4, mỗi giá trị đầu ra trong bản đồ tính năng không phải kết nối với từng giá trị pixel trong hình ảnh đầu vào. Nó chỉ cần kết nối với trường tiếp nhận, nơi bộ lọc đang được áp dụng. Vì mảng đầu ra không cần ánh xạ trực tiếp đến từng giá trị đầu vào, các Convolutional layer thường được gọi là các lớp “được kết nối một phần”. Tuy nhiên, đặc điểm này cũng có thể được mô tả là kết nối cục bộ.

Lưu ý rằng trọng lượng trong máy dò tính năng vẫn cố định khi nó di chuyển trên hình ảnh, còn được gọi là chia sẻ tham số. Một số tham số, như giá trị trọng lượng, điều chỉnh trong quá trình đào tạo thông qua quá trình backpropagation và gradient descent. Ty nhiên, có ba siêu cận điểm ảnh hưởng đến kích thước âm lượng của đầu ra cần được đặt trước khi đào tạo mạng thần kinh bắt đầu. Chúng bao gồm:

1. Số lượng bộ lọc ảnh ảnh hưởng đến độ sâu của đầu ra. Ví dụ: ba bộ lọc riêng biệt sẽ mang lại ba bản đồ tính năng khác nhau, tạo ra độ sâu ba.
2. Sải (Stride) là khoảng cách, hoặc số điểm ảnh, mà hạt nhân di chuyển qua ma trận đầu vào. Trong khi các giá trị sải của hai hoặc lớn hơn là rất hiếm, một sải bước lớn hơn mang lại chất lượng thấp hơn.
3. Zero-padding thường được sử dụng khi các bộ lọc không phù hợp với hình ảnh đầu vào. Điều này đặt tất cả các phần tử nằm ngoài ma trận đầu vào về 0, tạo ra đầu ra lớn hơn hoặc có kích thước bằng nhau. Có ba loại padding: Valid padding, Same padding, Full padding.

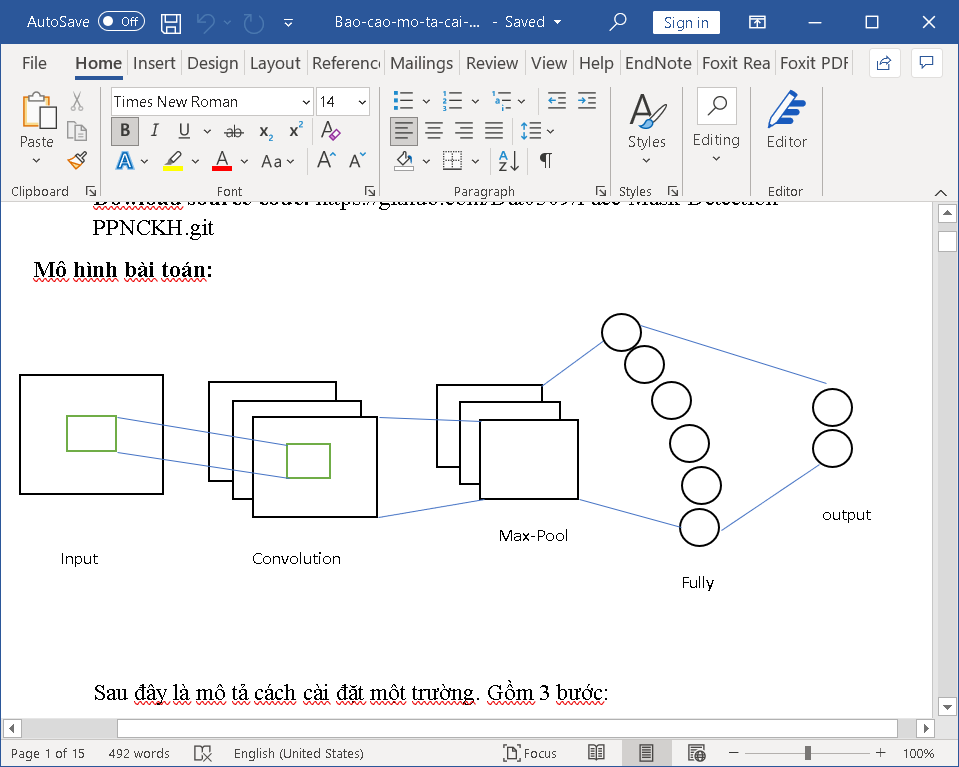
Cấu trúc cảu CNN có thể trở nên phân cấp vì ccas lớp sau có thể nhìn thấy các điểm ảnh trong các trường tiếp nhận của các lớp trước đó. Cuối cùng, Convolutional layer chuyển đổi hình ảnh thành các giá trị số, cho phép mạng neural diễn giải và trích xuất các mẫu có liên quan.

Pooling Layer, tiến hành giảm tính chiều hướng, giảm số lượng tham số trong đầu vào. Tương tự như Convolutional layer, hoạt động gộp chung quét một bộ lọc trên toàn bộ đầu vào, nhưng sự khác biệt là bộ lọc này không có bất kỳ trọng lượng nào. Thay vào đó, hạt nhân áp dụng hàm tổng hợp cho các giá trị trong trường tiếp nhận, điền mảng đầu ra. Mặc dù rất nhiều thông tin bị mất trong Pooling Layer, nhưng nó cũng có một số lợi ích cho CNN. Chúng giúp giảm độ phức tạp, cải thiện hiệu quả và hạn chế nguy cơ quá tải.

Fully-Connected layer, như đã đề cập trước đó, các giá trị pixel của hình ảnh đầu vào không được kết nối trực tiếp đến với lớp đầu ra trong các lớp được kết nối một phần. Tuy nhiên, trong layer được kết nối đầy đủ, mỗi nút trong layer output kết nối trực tiếp với một nút trong layer trước đó.

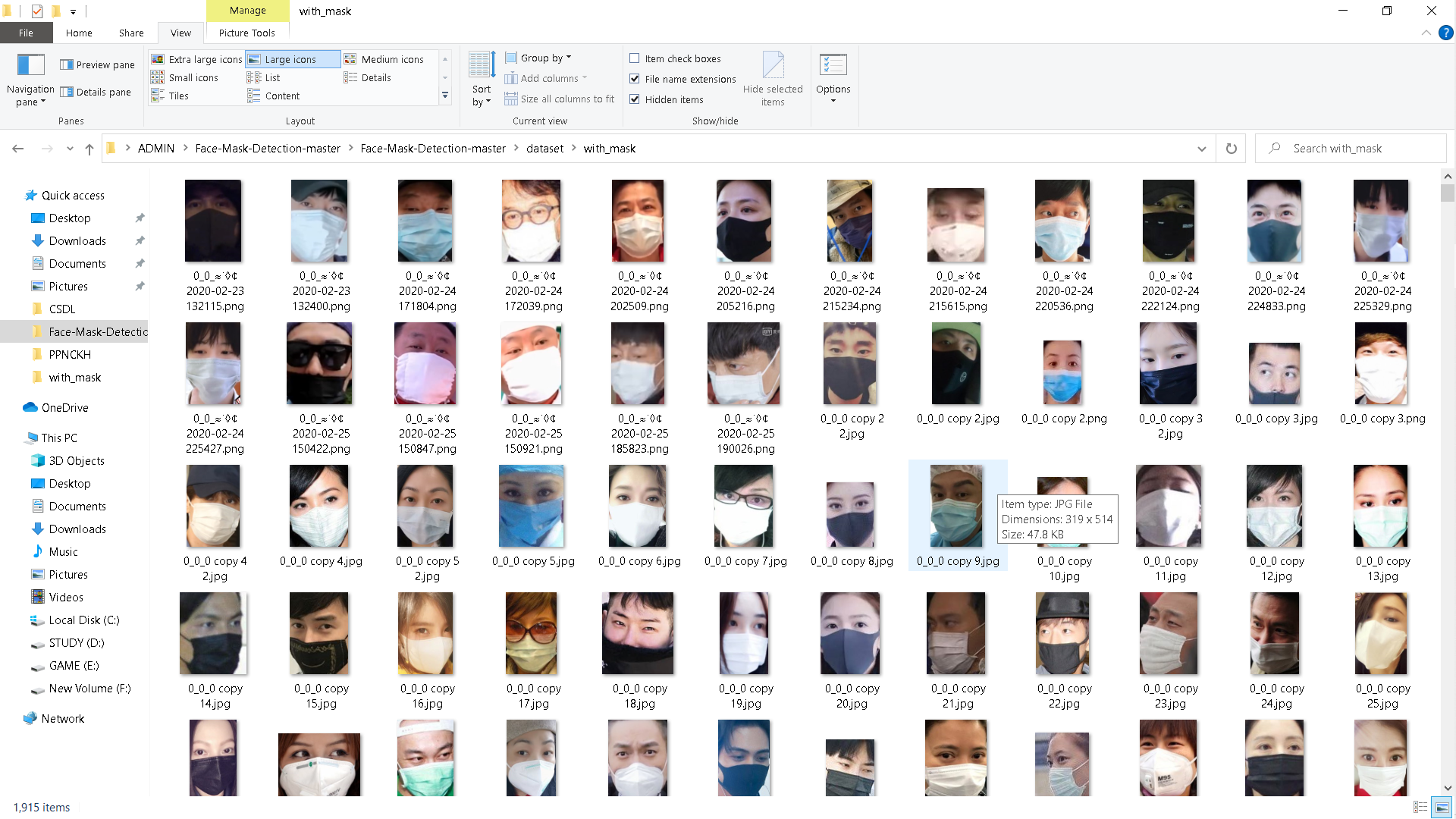
Lớp này thực hiện nhiệm vụ phân loại dựa trên các tính năng được trích xuất thông qua các lớp trước đó và các bộ lọc khác nhau của chúng. Trong khi các convolutional và pooling có xu hướng sử dụng các hàm ReLu, các lớp FC thường tận dụng chức năng kích hoạt softmax để phân loại đầu vào một các thích hợp, tạo ra xác suất từ 0 đến 1.

Đi sâu vào mô hình phát hiện khuôn mặt đeo khẩu trang, chúng tôi sử dụng phương pháp CNN để training dữ liệu cho máy. Dữ liệu được train cho máy gồm Những file hình ảnh có và không có khẩu trang. CNN sẽ tạo ra các lớp để phân tích, học tập các đặc trưng quan trọng nhất của file ảnh, để tiến hành quá trình nhận dạng.

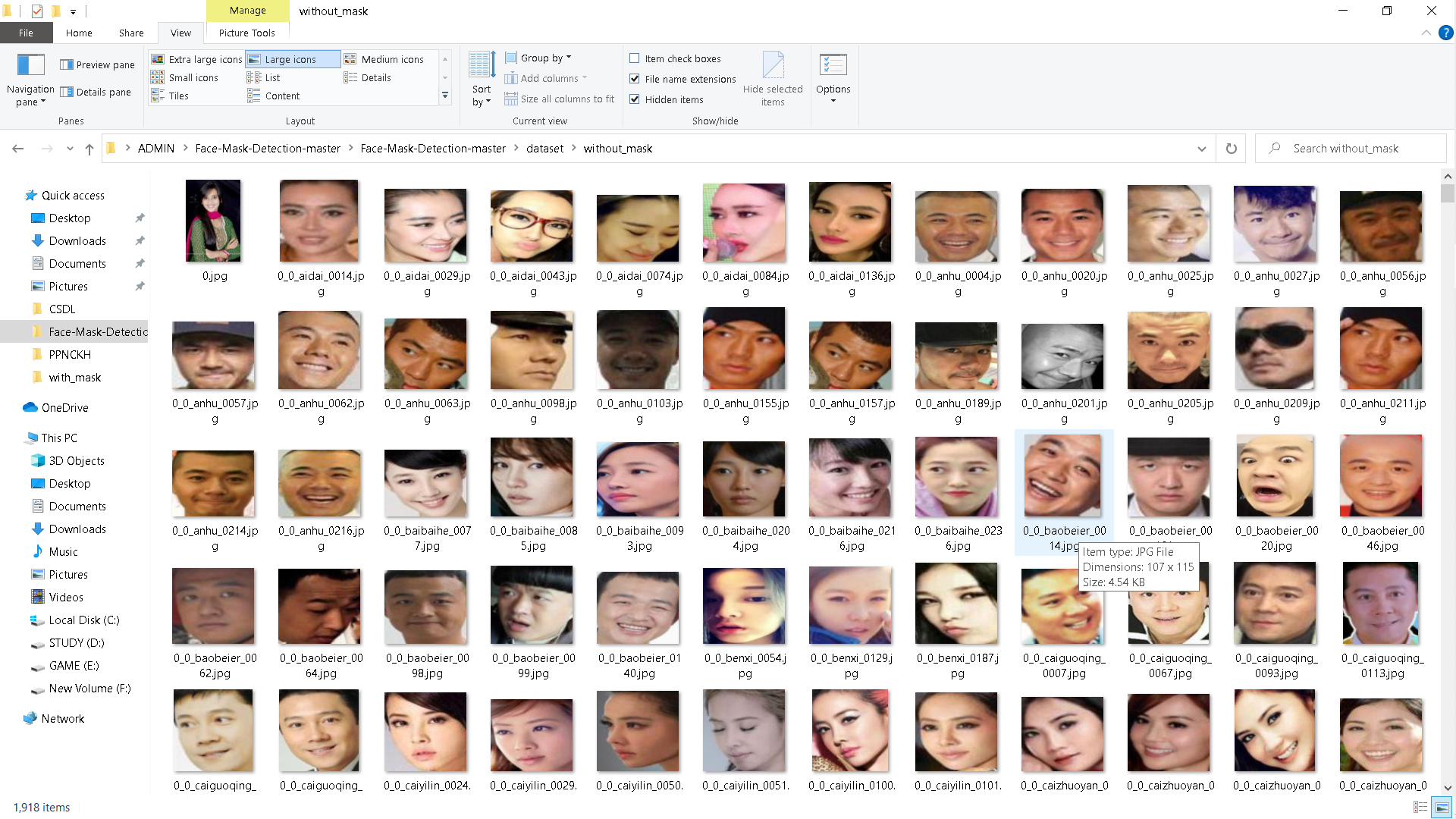


Hình 5. Mô hình CNN trong phát hiện khuôn mặt đeo khẩu trang.

Mạng neral được training một cách trực tiếp để đầu ra của nó trở thành một vector đặc trưng 128 chiều, sử dụng hàm chi phí bộ ba (tripletbased loss function). Một bộ ba (triplet) được định nghĩa bao gồm các khuôn mặt của nhiều người khác nhau. Mục tiêu của hàm là phân tách các khuôn mặt ra khỏi khuôn mặt negative sử dụng một lề khoảng cách – distance margin. Từ các độ đo thu được, thuật toán sẽ ước lượng giá trị của hàm chi phí dựa trên việc so sánh khoảng cách giữa các tập đặc trưng được sinh ra từ các ảnh khuôn mặt đeo khẩu trang khác nhau. Các giá trị ước lượng của hàm chi phí sau khi tính sẽ được lan truyền ngược từ lớp cuối cùng đến lớp đầu tiên của mạng để tính chi phí trọng số (cập nhật lại trọng số) trên các lớp mạng. Qúa trình tính toán, ước lượng và cập nhật trọng số của mạng được lặp đi lặp lại liên tục cho đến khi giá trị của hàm chi phí thỏa mãn điều kiện đã cho. Lặp lại các bước trên đối với toàn bộ tập dữ liệu huấn luyện cho đến khi thuật toán huấn luyện mạng hội tụ. Mô hình nhận dạng khuôn mặt được mô tả cụ thể trên Hình 5.



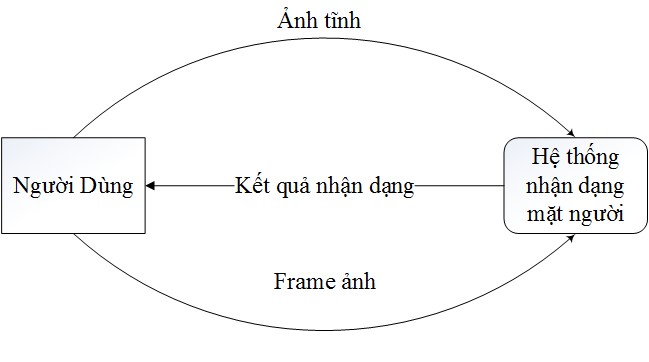
Hình 6. Tập ảnh dung để huấn luyện máy phát hiện khuôn mặt đeo khẩu trang.



Hình 7. Tập ảnh dung để huấn luyện máy phát hiện khuôn mặt không đeo khẩu trang

1. XÂY DỰNG CHƯƠNG TRÌNH – MÔ PHỎNG
   1. Xây dựng chương trình.

Nhiệm vụ chính của chương trình là phát hiện một hoặc nhiều khuôn mặt có đeo khẩu trang hay không. Ngoài ra, chương trình còn thực hiện công việc phát hiện, tách các khuôn mặt người (nếu có) từ một frame ảnh thu được từ camera.

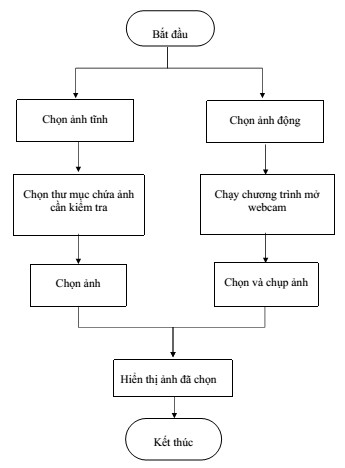


Hình 8. Sơ đồ ngữ cảnh của hệ thống.

Như vậy, các chức năng của chương trình bao gồm:

* Dựa trên dữ liệu đã được train là các tệp ảnh được lưu trong dataset như Hình 6, Hình 7. Chương tình sẽ kết nối đến webcam, hiển thị lên ImageBox.
* Thực hiện tách các khuôn mặt trên ImageBox
* So sánh các khuôn mặt vừa tách được với các khuôn mặt đã được train sẵn.
* Kết quả hiển thị “with\_mask” nếu phát hiện khẩu trang hoặc “no\_mask” nếu không phát hiện thấy khẩu trang.

Đầu vào của hệ thống là 1 ảnh chứa khuôn mặt cần xử lý, ảnh này có thể là ảnh tĩnh, lấy từ trong bộ nhớ máy tính hoặc là frame ảnh bắt được từ dòng hình ảnh của camera. Sau khi có được ảnh đầu vào thì bắt đầu tiến hành tìm kiếm, phát hiện các khuôn mặt trong ảnh có đeo khẩu trang hay không.

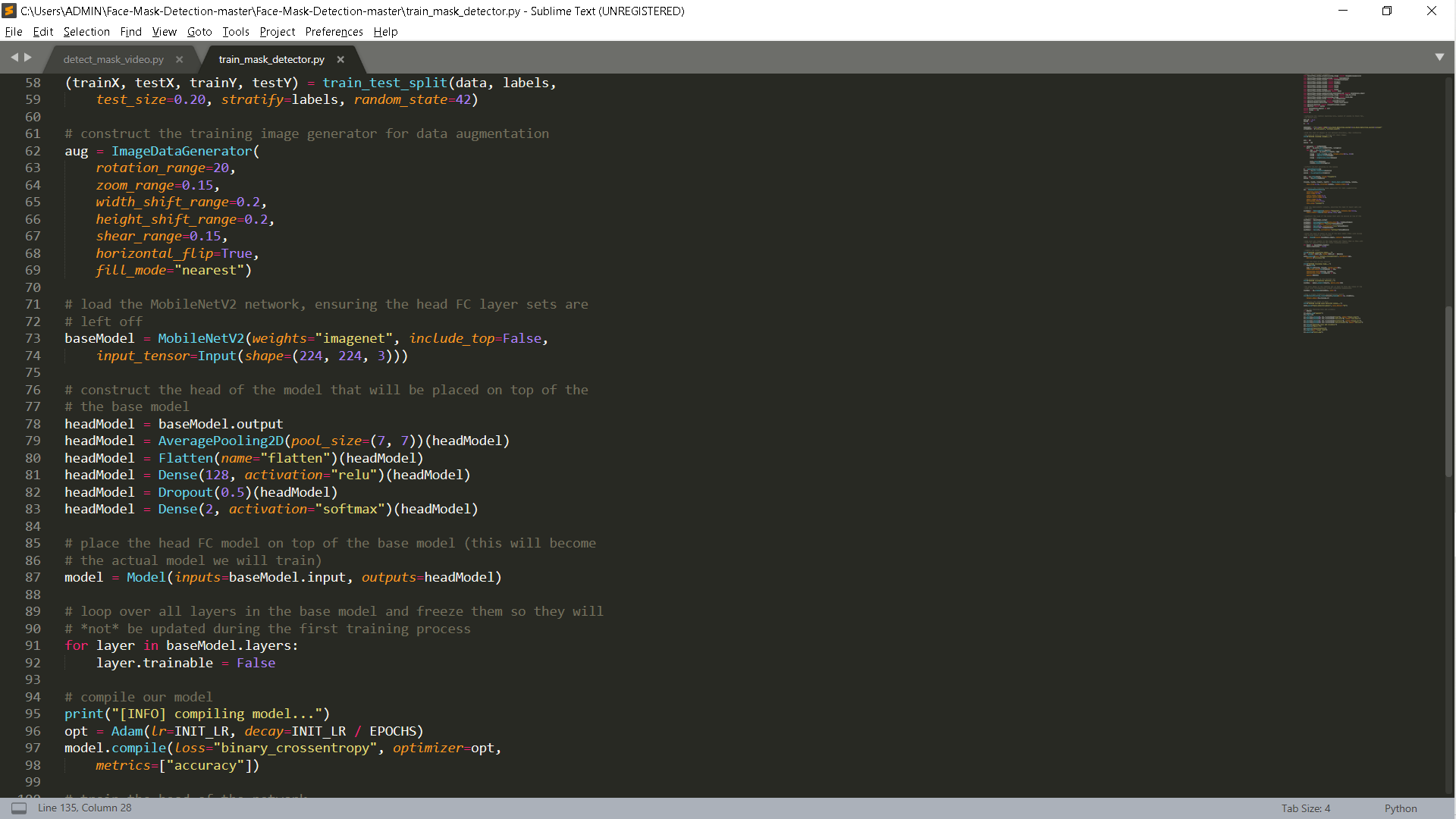


Hình 9. Lưu đồ giải thuật chọn ảnh đầu vào.

Sau khi nhận được ảnh đầu vào, hệ thống sẽ thực hiện chức năng phát hiện khuôn mặt có/không đeo khẩu trang trong ảnh. Như đã nói trong phần giới thiệu, bài toán phát hiện khuôn mặt đeo khẩu trang trong ảnh/video là một bài toán khó, nên ta không đi sâu tìm hiểu cách giải quyết bài toán này.

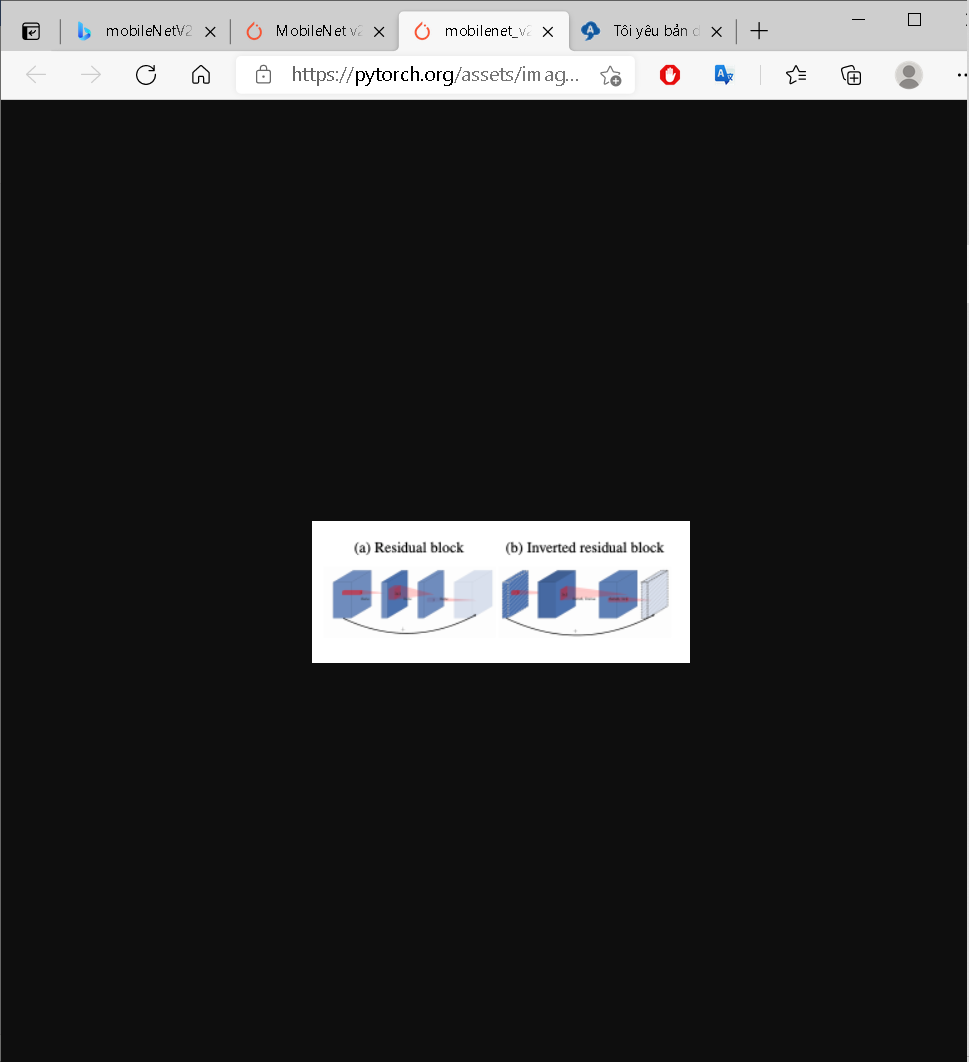
Ở đây ta sử dụng phương pháp tiếp cận theo hướng diện mạo (appearance-based) được thực hiện nhanh bằng thuật toán adaboost thông qua hàm cvHaarObjects() của bộ thư viện OpenCV. Hàm này thực hiện việc phát triển đối tượng dựa trên các đặc trưng haar-like, cụ thể là nhờ vào một bộ Cascade được truyền vào cho hàm. Bộ cascade được xây dựng theo dạng tree-node và đã được huấn luyện từ trước.

Việc huấn luyện bộ Cascade có thể thực hiện từ những dữ liệu thu thập được để phục vụ cho quá trình nhận dạng. Ví dụ, muốn nhận dạng một người A có đeo khẩu trang hay không, ta thu thập các ảnh khuôn mặt của những người đeo khẩu trang ở nhiều góc độ, tư thế và điều kiện ánh sáng khác nhau, sau đó cho bộ phận dạng học theo thuật toán của CNN training. Kết quả thu được sẽ có một mô hình nhận diện, được sử dụng để nhận dạng các đối tượng.



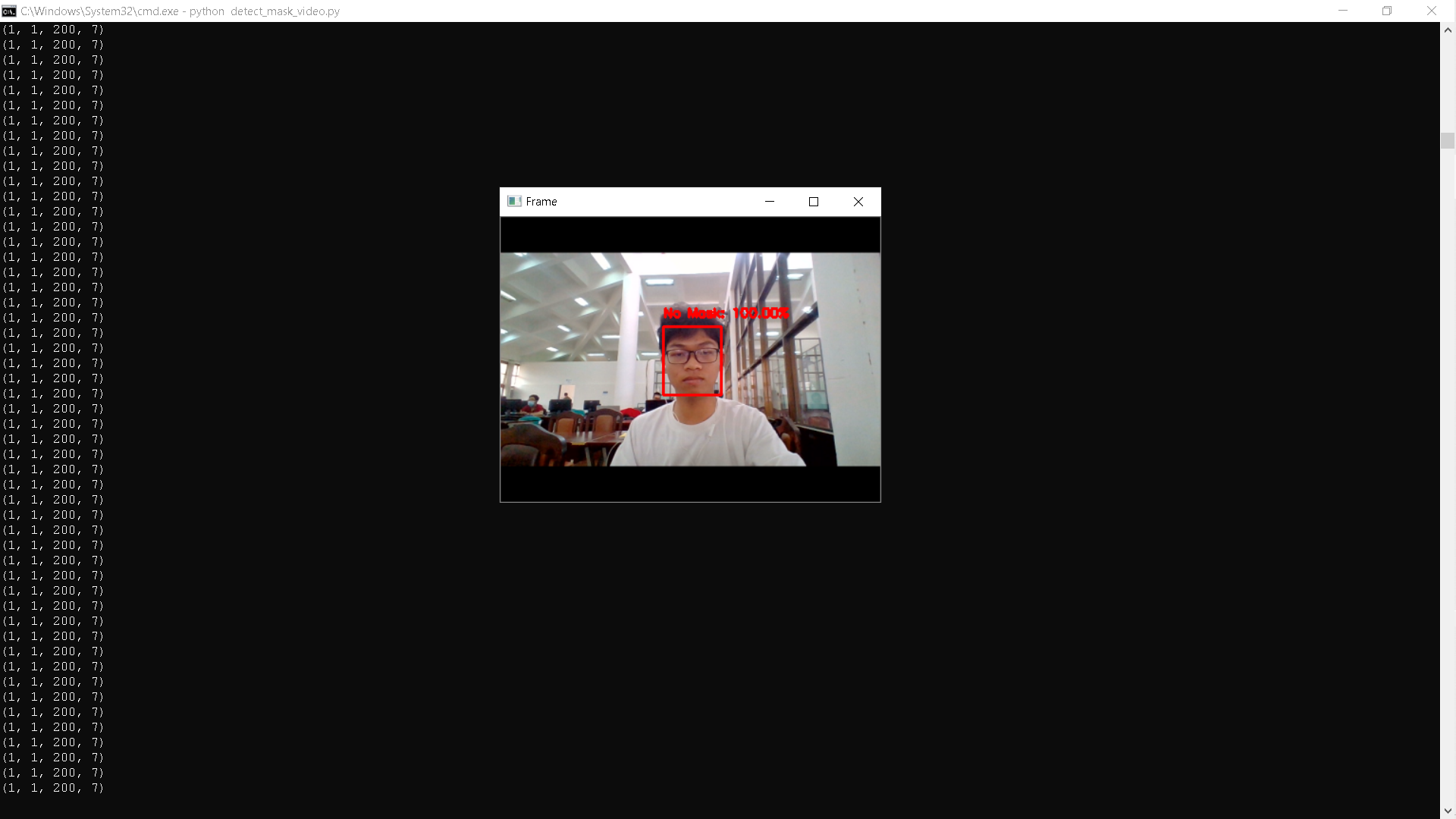
Hình 10. Cách học của mô hình CNN

Trong Hình 10, ta thấy xuất hiện MobileNetV2. Kiến trúc MobileNetV2 dựa trên cấu trúc phần dư ngược, trong đó đầu vào và đầu ra của khối dư là các lớp nút cổ chai mỏng đối diện với các mô hình phần dư truyền thống sử dụng các biểu diễn mở rộng trong đầu vào. MobileNetV2 sử dụng phức hợp theo chiều sâu nhẹ để lọc các tính anwng trong lớp mở rộng trung gian. Ngoài ra, các điểm không tuyến tính trong các lớp hẹp đã bị loại bỏ để duy trì sức manh đại diện.

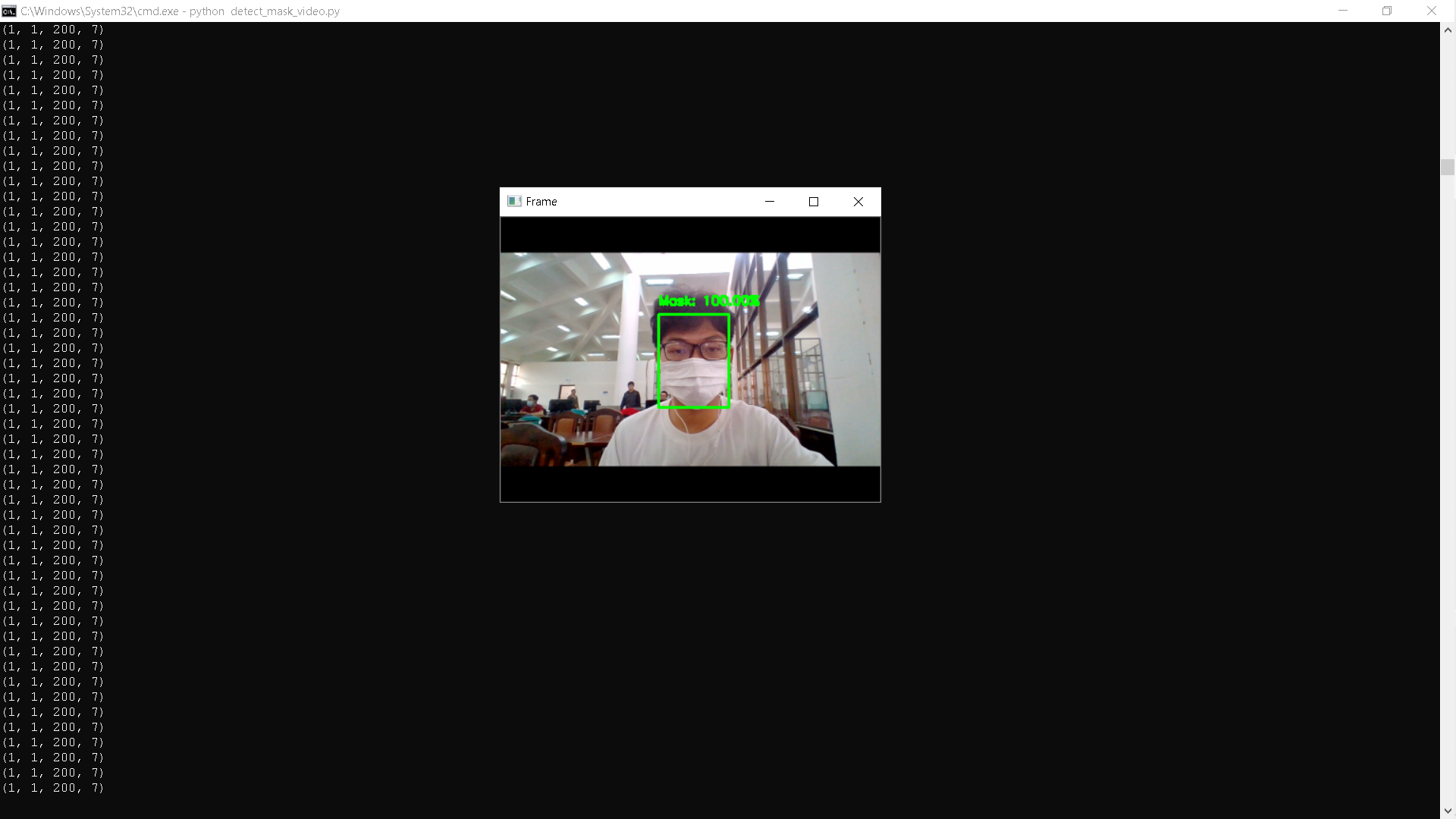


Hình 11. Cách hoạt động của MobileNetV2

AveragePooling2D là lớp hoạt động tổng hợp trung bình cho dữ liệu không gian. Đối số pool\_size là một số nguyên hoặc tuple của 2 số nguyên, các yếu tố để giảm quy mô (dọc, ngang). Sẽ giảm một nữa đầu vào trong cả hai kích thước không gian. Nếu chỉ có một số nguyên được chỉ định, cùng một độ dài cửa sổ sẽ được sử dụng cho cả 2 kích thước (7,7). Mô hình CNN chia thành nhiều lớp để tiến hành xử lý dữ liệu ảnh trong dataset như Flatten,Dense, Dropour.



Hình 12. Kết quả phát hiện khuôn mặt không có khẩu trang.



Hình 13. Kết quả phát hiện khuôn mặt có đeo khẩu trang.

Kết quả thực nghiệm (Hình 12, Hình 13) cho thấy việc sử dụng các lớp mạng tích chập để trích chọn đặc trưng cho độ chính xác cao. Do thuật toán được huấn luyện với tập dữ liệu lớn, đa dạng nên các đặc trưng phát hiện được thường ít bị ảnh hưởng bởi nhiễu và ccas tính chất nghiêng, xoay của ảnh. Tuy nhiên, do mạng được kiến trúc nhiều lớp và số liên kết giữa các lớp mạng rất lớn nên việc tính toán trên mạng thường mất nhiều thời gian. Điều này dẫn tới tốc độ tổng thể của thuật toán nhận dạng sẽ bị ảnh hưởng. Vì vậy, để đảm bảo thuật toán có thể đáp ứng tính thời gian thực (real-time) trong quá trình nhận dạng, chúng tôi đã tận dụng khả năng tính toán GPU (Graphic Proccessing Unit), cho phép việc tính toán trên các lớp mạng thực hiện theo cơ chế song song.

1. ĐÁNH GIÁ THỰC NGHIỆM

* Môi trường thực nghiệm

Chương trình thực nghiệm được cài đặt trong môi trường python, sử dụng các thư viện Keras cho việc biểu diễn, lưu trữ và thao tác dữ liệu, thư viện OpenCV để thực hiện các thao tác xử lý ảnh cơ bản, thử nghiệm các mô hình học máy (mạng neural, mô hình svm, v.v.). Chương trình được thử nghiệm trên hệ điều hành Windows 10, máy PC tốc độ 2,4GHz, bộ nhớ 4GB.

* Dữ liệu thử nghiệm

Hiệu quả của mô hình nhận dạng được đánh giá trên các bộ cơ sở dữ liệu chuẩn (chứa các khung hình được thu nhận từ các thiết bị camera, webcam khác nhau), được công bố dung chung cho các nhóm nghiên cứu trên thế giới. Gồm 2 tập dữ liệu: with\_mask, without\_mask. Anhr trong cơ sở dữ liệu là ảnh màu 24 bit, định dạng JPEG. Tập dữ liệu chứa một tập hợp các hình ảnh khuôn mặt có đeo khẩu trang và không đeo khẩu trang, tổng cộng có hơn 1200 hình ảnh. Tất cả khuôn mặt chủ yếu được thực hiện bới các sinh viên đại học năm đầu tiên có độ tuổi từ 18 đến 20 và một số người lớn tuổi, một số cá nhân đeo kính, có râu, thuộc nhiều chủng tộc khác nhau (Hinhf6, Hình 7).

* Kết quả thực nghiệm

Qúa trình đánh giá thực nghiệm được chia thành 2 công đoạn: Đánh giá hiệu quả của mô hình phát hiện khuôn mặt đeo khẩu trang được đánh giá dựa trên các độ đo đượ định nghĩa cụ thể trong phần sau đây:

* Độ chính xác phát hiện khuôn mặt DP (Detection Precision):

DP = Số vùng khuôn mặt phát hiện đúng trên tổng số khuôn mặt cần phát hiện

* Khả năng tìm hết DR (Detection Recall):

DR = Số vùng khuôn mặt phát hiện đúng / (số vùng khuôn mặt phát hiện đúng + Số vùng không được phát hiện)

* Độ trung bình điều hòa DM (Detection F - Measure):

DM = (2×FDP\*FDR)/(FDP+FDR)

Bên cạnh đó, để các kết quả thực nghiệm chính xác và trực quan, trong quá trình thử nghiệm, chúng tôi đã so sánh hiệu quả của mô hình phát hiện khuôn mặt đề xuất với mô hình phát hiện khuôn mặt sử dụng đặc trưng Haar-Like và bộ phân lớp AdaBoost (được quy ước gọi tên là phương pháp Haar-Like AdaBoost). Các kết quả thực nghiệm được mô tả cụ thể trên Bảng 1.

Bảng 1. ĐÁNH GIÁ HIỆU QUẢ PHÁT HIỆN KHUÔN MẶT ĐEO KHẨU TRANG

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dữ liệu thử nghiệm | Số mẫu | Phương pháp đề xuất | | | Phương pháp Haar-Like AdaBoost | | |
| DP | DR | DM | DP | DR | DM |
| With\_mask | 650 | 98.43 | 98.43 | 98.43 | 80.71 | 80.71 | 80.71 |
| Without\_mask | 700 | 97.43 | 97.43 | 97.43 | 79.43 | 79.43 | 79.43 |

Hiệu quả của mô hình nhận dạng tổng thể được đánh giá dựa trên độ chính xác nhận dạng, được định nghĩa cụ thể như sau:

R\_Precision = Số khuôn mặt nhận dạng đúng/ Tổng số khuôn mặt cần nhận dạng.

Qúa trình đánh giá thực nghiệm được thực hiện lần lượt trên từng tập dữ liệu. Mỗi tập dữ liệu được chia ngẫu nhiên thành 2 tập training và testing tho tỷ lệ 90/10 (90% số mẫu để huấn luyện mô hình và 10% số mẫu còn lại để kiểm thử).

Việc huấn luyện mô hình gồm 2 công đoạn: Huấn luyện bộ trích chọn đặc trưng (mô hình mạng neural tích chập) và huấn luyện bộ phân lớp SVM. Quy trình huấn luyện được tiến hành cụ thể như sau: Từ tập mẫu huấn luyện đầu vào, trước tiên bộ phát hiện khuôn mặt sẽ tiến hành tìm kiếm, định vị và crop vùng ảnh khuôn mặt trên mỗi khung hình. Toàn bộ tập ảnh khuôn mặt crop sau đó sẽ được sử dụng làm đầu vào để huấn luyện mô hình trích chọn đặc trưng. Tập đặc trưng đầu ra của mô hình trích chọn đặc trưng sẽ được sử dụng làm đầu vào để huấn luyện mô hình phân lớp SVM.

Các kết quả thực nghiệm được mô tả cụ thể trên Bảng 2. Trong đó, hiệu quả của mô hình đề xuất được so sánh với phương pháp phân lớp sử dụng đặc trưng PCA và bộ phân lớp Eigenface (được quy ước gọi tên là phương pháp PCA-Eigenface).

Bảng 2. ĐÁNH GIÁ ĐỘ CHÍNH XÁC NHẬN DẠNG

|  |  |  |  |
| --- | --- | --- | --- |
| Tập dữ liệu thử nghiệm | Số khuôn mặt cần nhận dạng | R\_Precision (%) | |
| PP đề xuất | PCA-Eigenface |
| With\_mask | 650 | 98.6 | 82.12 |
| Without\_mask | 700 | 97.55 | 75.43 |

Từ các kết quả thực nghiệm cho thấy phương pháp đề xuất đạt được độ chính xác cao (trên 95%) trên tất cả các tập dữ liệu thử nghiệm. Trong khi đó độ chính xác của phương pháp PCA – Eigenface bị ảnh hưởng nhiều bởi độ sáng và độ dịch chuyển của ảnh đầu vào.

1. KẾT LUẬN

Bài báo này đề xuất một mô hình tổng thể cho việc phát hiện khuôn mặt đeo khẩu trang từ webcam/camera. Trong đó tập trung chính vào công đoạn phát hiện và nhận dạng khuôn mặt. Hiệu quả của mô hình đã được đánh giá trên các tập dữ liệu chuẩn, dung chung cho cộng đồng nghiên cứu về nhận dạng khuôn mặt người trên thế giới. Qúa trình đánh giá thực nghiệm được chia thành 2 bước, trong đó hiệu quả của phương pháp phát triển khuôn mặt được đánh giá dựa trên 3 độ đo là độ chính xác (Precision), khả năng tìm hết (recall) và độ đo F-meansure, hiệu quả của mô hình nhận dạng khuôn mặt được đánh giá dựa trên độ chính xác nhận dạng. Các kết quả thực nghiệm cho thấy mô hình đề xuất đạt được độ chính xác cao và ổn định trong môi trường thực tế, có thể ứng dụng mô hình để giải quyết các bài toán ứng dụng điển hình như trên hệ thống camera giám sát cho phép phát hiện những người ra vào nơi công cộng mà không đeo khẩu trang, các khu vực đông người như nhà ga, sân bay, các cơ quan chính phủ, tòa nhà, chung cư, công trường,v.v.

TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | T. S. M. P. M. Karthikeyan, “Real Time Face Recognition based Smart Lab for Energy,” in *Webology*, India, 2020. |
| [2] | A. Roihan, “Perancangan Sistem Kehadiran Face Recognition Menggunakan,” in *TMJ*, Indonesia, 2021. |
| [3] | D. A. Ph.D. Candidate, “Comparative Analysis of Machine Learning Based Approaches for Face,” in *University of Tehran, Faculty of Management*, India, 2021. |
| [4] | M. V. M. A. N. G. K. Mr. Amar Sarode, “REVIEW PAPER: VEHICAL ANTI-THEFT FACE RECOGNITION,” in *IRJMETS*, Indie, 2020. |
| [5] | M. H. K. K. Lars Ankile, “Application of Facial Recognition using,” in *Department of Computer Science*, America, 2020. |
| [6] | M. S. Islam, “A Novel Approach to Detect Face Mask using CNN,” in *International University*, Bangladesh, 2020. |
| [7] | D. S. S. F. Hafidz Sanjaya, “Face Recognition Using Tiny Yolo V2 Algorithm as Attendance,” in *International Journal of Information System & Technology*, Indonesia, 2020. |
| [8] | A. Fau, “Face Recognition Application Using the Eigenface Method for,” in *International Journal of Information System & Technology*, Indonesia, 2020. |
| [9] | A. C. S. N. D. A. M. L. Devina Varshney, “MULTIMODAL EMOTION DETECTION USING,” in *IJARET*, India, 2020. |
| [10] | L. Bhadana, “REAL TIME FACIAL EMOTION RECOGNITION,” in *JOURNAL OF CRITICAL REVIEWS*, America, 2020. |