DeepFace: Thu hẹp khoảng cách đến hiệu suất nhận diện khuôn mặt ở mức độ con người

Yaniv Taigman Ming Yang Marc'Aurelio Ranzato

Facebook AI Research
Menlo Park, CA, Hoa Kỳ

{yaniv, mingyang, ranzato}@fb.com

Lior Wolf

Đại học Tel Aviv
Tel Aviv, Israel

wolf@cs.tau.ac.il

Tóm tắt

Trong nhận diện khuôn mặt hiện đại, quy trình thông thường bao gồm bốn giai đoạn: phát hiện □ căn chỉnh □ biểu diễn □ phân loại. Chúng tôi xem xét lại cả bước căn chỉnh và bước biểu diễn bằng cách sử dụng mô hình hóa khuôn mặt 3D rõ ràng nhằm áp dụng phép biến đổi affine từng phần, và trích xuất biểu diễn khuôn mặt từ một mạng nơ-ron sâu gồm chín lớp. Mạng sâu này bao gồm hơn 120 triệu tham số, sử dụng một số lớp kết nối cục bộ mà không chia sẻ trọng số, thay vì các lớp tích chập tiêu chuẩn. Do đó, chúng tôi đã huấn luyện nó trên bộ dữ liệu khuôn mặt lớn nhất cho đến nay, một bộ dữ liệu được gán nhãn danh tính gồm bốn triệu hình ảnh khuôn mặt thuộc về hơn 4.000 danh tính. Các biểu diễn học được, kết hợp giữa căn chỉnh dựa trên mô hình chính xác và cơ sở dữ liệu khuôn mặt lớn, tổng quát hóa rất tốt cho các khuôn mặt trong môi trường không bị ràng buộc, ngay cả với bộ phân loại đơn giản. Phương pháp của chúng tôi đạt độ chính xác 97,35% trên bộ dữ liệu Labeled Faces in the Wild (LFW), giảm sai số của phương pháp hiện tại hơn 27%, tiến gần đến hiệu suất ở mức con người.

1. Giới thiệu

Nhận diện khuôn mặt trong các hình ảnh không bị ràng buộc đang ở tuyến đầu của cuộc cách mạng nhận thức thuật toán. Các tác động xã hội và văn hóa của các công nghệ nhận diện khuôn mặt là rất sâu rộng, tuy nhiên khoảng cách hiệu suất hiện tại trong lĩnh vực này giữa máy móc và hệ thống thị giác của con người đóng vai trò như một lớp đệm giúp tránh phải đối mặt với những tác động này.

Chúng tôi giới thiệu một hệ thống (DeepFace) đã thu hẹp phần lớn khoảng cách còn lại trên bộ đánh giá phổ biến nhất về nhận diện khuôn mặt trong điều kiện không bị ràng buộc, và hiện đang tiến sát đến độ chính xác ở mức con người. Hệ thống được huấn luyện trên một tập dữ liệu lớn gồm các khuôn mặt được thu thập từ một quần thể rất khác biệt so với quần thể được sử dụng để xây dựng các bộ đánh giá, và có khả năng vượt trội hơn các hệ thống hiện có chỉ với sự thích nghi tối thiểu. Hơn nữa, hệ thống tạo ra một biểu diễn khuôn mặt cực kỳ nhỏ gọn, hoàn toàn trái ngược với sự chuyển dịch hướng tới hàng chục nghìn đặc trưng về ngoại hình trong các hệ thống gần đây khác [5, 7, 2].

Hệ thống được đề xuất khác với phần lớn các đóng góp trong lĩnh vực này ở chỗ nó sử dụng khung học sâu (deep learning - DL) [3, 21] thay cho các đặc trưng được thiết kế kỹ lưỡng. DL đặc biệt phù hợp để xử lý các bộ dữ liệu huấn luyện lớn, với nhiều thành công gần đây trong các lĩnh vực đa dạng như thị giác, nhận diện giọng nói và mô hình hóa ngôn ngữ. Cụ thể với khuôn mặt, sự thành công của mạng học được trong việc nắm bắt ngoại hình khuôn mặt một cách mạnh mẽ phụ thuộc rất nhiều vào bước căn chỉnh 3D cực kỳ nhanh. Kiến trúc mạng dựa trên giả định rằng một khi việc căn chỉnh đã hoàn tất, vị trí của từng vùng khuôn mặt sẽ được cố định ở cấp độ điểm ảnh. Do đó, có thể học từ các giá trị RGB của điểm ảnh thô, mà không cần phải áp dụng nhiều lớp tích chập như được thực hiện trong nhiều mạng khác [19, 21].

Tóm lại, chúng tôi đóng góp những điểm sau: (i) Phát triển một kiến trúc và phương pháp học mạng nơ-ron sâu (DNN) hiệu quả, tận dụng một bộ dữ liệu khuôn mặt được gán nhãn rất lớn để thu được biểu diễn khuôn mặt có khả năng tổng quát hóa tốt sang các bộ dữ liệu khác; (ii) Một hệ thống căn chỉnh khuôn mặt hiệu quả dựa trên mô hình hóa 3D rõ ràng của khuôn mặt; và (iii) Nâng cao đáng kể trình độ hiện tại trong (1) bộ chuẩn Labeled Faces in the Wild (LFW) [18], đạt gần mức hiệu suất của con người; và (2) bộ dữ liệu YouTube Faces (YTF) [30], giảm tỷ lệ lỗi ở đó hơn 50%.

1.1. Công trình liên quan

Dữ liệu lớn và học sâu Trong những năm gần đây, một số lượng lớn ảnh đã được các công cụ tìm kiếm thu thập và được tải lên các mạng xã hội, bao gồm nhiều loại nội dung không bị ràng buộc, như vật thể, khuôn mặt và cảnh vật. Khối lượng dữ liệu lớn này cùng với sự gia tăng về tài nguyên tính toán đã cho phép sử dụng các mô hình thống kê mạnh mẽ hơn. Những mô hình này đã cải thiện đáng kể độ vững chắc của các hệ thống thị giác đối với nhiều biến đổi quan trọng, như biến dạng không cứng nhắc, lộn xộn, che khuất và chiếu sáng, tất cả đều là những vấn đề cốt lõi của nhiều ứng dụng thị giác máy tính. Trong khi các phương pháp học máy truyền thống

các phương pháp học máy như Máy Hỗ Trợ Vector (Support Vector Machines), Phân Tích Thành Phần Chính (Principal Component Analysis) và Phân Tích Tuyến Tính Đa Dạng (Linear Discriminant Analysis), có khả năng hạn chế trong việc tận dụng khối lượng lớn dữ liệu, trong khi các mạng nơ-ron sâu đã cho thấy khả năng mở rộng tốt hơn. Gần đây, đã có một làn sóng quan tâm đến các mạng nơ-ron [19, 21]. Đặc biệt, các mạng sâu và lớn đã thể hiện kết quả ấn tượng khi: (1) chúng được áp dụng cho lượng lớn dữ liệu huấn luyện và (2) các tài nguyên tính toán có thể mở rộng như hàng ngàn lõi CPU [11] và/hoặc GPU [19] đã trở nên sẵn có. Đáng chú ý nhất, Krizhevsky et al. [19] đã chỉ ra rằng các mạng nơ-ron tích chập rất lớn và sâu [21] được huấn luyện bằng phương pháp lan truyền ngược tiêu chuẩn [25] có thể đạt được độ chính xác nhận diện xuất sắc khi được huấn luyện trên một bộ dữ liệu lớn.

Nhận diện khuôn mặt - trình độ hiện tại: Tỷ lệ lỗi trong nhận diện khuôn mặt đã giảm đi ba bậc trong hai mươi năm qua [12] khi nhận diện các khuôn mặt chính diện trong các ảnh tĩnh được chụp trong môi trường kiểm soát nhất quán (có ràng buộc). Nhiều nhà cung cấp đã triển khai các hệ thống tinh vi cho ứng dụng kiểm soát biên giới và nhận diện sinh trắc học thông minh. Tuy nhiên, các hệ thống này đã cho thấy sự nhạy cảm với nhiều yếu tố khác nhau, như ánh sáng, biểu cảm, che khuất và lão hóa, những yếu tố này làm giảm đáng kể hiệu suất nhận diện người trong các môi trường không bị ràng buộc như vậy.

Hầu hết các phương pháp xác thực khuôn mặt hiện nay sử dụng các đặc trưng được thiết kế thủ công. Hơn nữa, các đặc trưng này thường được kết hợp để cải thiện hiệu suất, ngay cả trong những đóng góp đầu tiên cho LFW. Các hệ thống hiện đang dẫn đầu bảng xếp hạng hiệu suất sử dụng hàng chục nghìn bộ mô tả hình ảnh [5, 7, 2]. Ngược lại, phương pháp của chúng tôi được áp dụng trực tiếp lên giá trị pixel RGB, tạo ra một bộ mô tả rất nhỏ gọn nhưng thưa.

Các mạng nơ-ron sâu cũng đã được áp dụng trước đây cho phát hiện khuôn mặt [24], căn chỉnh khuôn mặt [27] và xác thực khuôn mặt [8, 16]. Trong lĩnh vực không bị ràng buộc, Huang et al. [16] đã sử dụng đặc trưng LBP làm đầu vào và họ cho thấy sự cải thiện khi kết hợp với các phương pháp truyền thống. Trong phương pháp của chúng tôi, chúng tôi sử dụng hình ảnh thô làm biểu diễn cơ sở, và để nhấn mạnh đóng góp của công trình, chúng tôi tránh kết hợp các đặc trưng của mình với các bộ mô tả được thiết kế. Chúng tôi cũng cung cấp một kiến trúc mới, đẩy giới hạn của những gì có thể đạt được với các mạng này bằng cách tích hợp căn chỉnh 3D, tùy chỉnh kiến trúc cho các đầu vào đã căn chỉnh, mở rộng mạng gần hai bậc độ lớn và trình diễn một phương pháp chuyển giao tri thức đơn giản khi mạng đã được huấn luyện trên một bộ dữ liệu lớn có gán nhãn.

Các phương pháp học metric được sử dụng nhiều trong xác thực khuôn mặt, thường kết hợp với các mục tiêu đặc thù cho từng tác vụ [26, 29, 6]. Hiện tại, hệ thống thành công nhất sử dụng một bộ dữ liệu lớn các khuôn mặt có gán nhãn [5] áp dụng một kỹ thuật học chuyển giao thông minh, điều chỉnh mô hình Joint Bayesian [6] được học trên một bộ dữ liệu chứa 99.773 ảnh từ 2.995 đối tượng khác nhau, sang miền ảnh LFW. Ở đây, để

(a) (b) (c) (d)

(e) (f) (g) (h)

Hình 1. Quy trình căn chỉnh. (a) Khuôn mặt được phát hiện, với 6 điểm mốc ban đầu. (b) Ảnh cắt căn chỉnh 2D được tạo ra. (c) 67 điểm mốc trên ảnh cắt căn chỉnh 2D cùng với phép phân chia tam giác Delaunay tương ứng, chúng tôi đã thêm các tam giác ở viền ngoài để tránh sự gián đoạn. (d) Hình dạng 3D tham chiếu được biến đổi lên mặt phẳng ảnh của ảnh cắt căn chỉnh 2D. (e) Độ hiển thị của các tam giác so với camera 3D-2D đã được khớp; các tam giác tối hơn thì ít hiển thị hơn. (f) 67 điểm mốc được tạo ra bởi mô hình 3D, được sử dụng để điều hướng quá trình biến dạng từng phần theo phép biến đổi affine. (g) Ảnh cắt chính diện cuối cùng. (h) Một góc nhìn mới được tạo ra bởi mô hình 3D (không được sử dụng trong bài báo này).

để chứng minh hiệu quả của các đặc trưng, chúng tôi giữ cho bước học khoảng cách đơn giản.

2. Căn chỉnh khuôn mặt

Các phiên bản đã được căn chỉnh sẵn của một số cơ sở dữ liệu khuôn mặt (ví dụ: LFW-a [29]) giúp cải thiện các thuật toán nhận diện bằng cách cung cấp đầu vào đã được chuẩn hóa [26]. Tuy nhiên, việc căn chỉnh khuôn mặt trong các tình huống không bị ràng buộc vẫn được xem là một vấn đề khó, phải tính đến nhiều yếu tố như tư thế (do khuôn mặt không phẳng) và các biểu cảm không cứng nhắc, vốn rất khó tách biệt khỏi hình thái khuôn mặt mang đặc trưng nhận diện. Các phương pháp gần đây đã cho thấy những cách thành công để bù đắp cho những khó khăn này bằng cách sử dụng các kỹ thuật căn chỉnh tinh vi. Những phương pháp này có thể sử dụng một hoặc nhiều trong các cách sau: (1) sử dụng mô hình 3D phân tích của khuôn mặt [28, 32, 14], (2) tìm kiếm các cấu hình điểm chuẩn tương tự từ một tập dữ liệu bên ngoài để suy ra [4], và (3) các phương pháp không giám sát nhằm tìm ra phép biến đổi tương đồng cho các điểm ảnh [17, 15].

Mặc dù căn chỉnh được sử dụng rộng rãi, hiện tại vẫn chưa có giải pháp hoàn chỉnh và chính xác về mặt vật lý trong bối cảnh xác thực khuôn mặt không bị ràng buộc. Các mô hình 3D đã không còn được ưa chuộng trong những năm gần đây, đặc biệt là trong các môi trường không bị ràng buộc. Tuy nhiên, vì khuôn mặt là các đối tượng 3D, nếu được thực hiện đúng cách, chúng tôi tin rằng đây là hướng đi đúng. Trong bài báo này, chúng tôi mô tả một hệ thống bao gồm mô hình hóa 3D phân tích của khuôn mặt dựa trên các điểm chuẩn, được sử dụng để biến đổi một vùng khuôn mặt đã được phát hiện sang chế độ chính diện 3D (frontalization).

Tương tự như phần lớn các tài liệu căn chỉnh gần đây, phương pháp căn chỉnh của chúng tôi dựa trên việc sử dụng các bộ phát hiện điểm chuẩn để điều hướng quá trình căn chỉnh. Chúng tôi sử dụng một bộ điểm chuẩn tương đối đơn giản

bộ phát hiện điểm, nhưng áp dụng nó qua nhiều lần lặp để tinh chỉnh đầu ra của nó. Ở mỗi lần lặp, các điểm chuẩn được trích xuất bởi một Bộ Hồi Quy Hỗ Trợ Vector (SVR) được huấn luyện để dự đoán cấu hình điểm từ một mô tả ảnh. Mô tả ảnh của chúng tôi dựa trên LBP Histograms [1], nhưng cũng có thể xem xét các đặc trưng khác. Bằng cách biến đổi ảnh sử dụng ma trận tương đồng T được sinh ra thành một ảnh mới, chúng ta có thể chạy bộ phát hiện điểm chuẩn một lần nữa trên không gian đặc trưng mới và tinh chỉnh vị trí.
Căn chỉnh 2D Chúng tôi bắt đầu quá trình căn chỉnh bằng cách phát hiện 6 điểm chuẩn bên trong vùng cắt phát hiện, tập trung tại trung tâm của mắt, chóp mũi và vị trí miệng như minh họa ở Hình 1(a). Chúng được sử dụng để xấp xỉ tỷ lệ, xoay và tịnh tiến ảnh vào sáu vị trí neo bằng cách khớp T i 2d := (si, Ri, ti) trong đó: xj anchor := si[Ri|ti]□xj source cho các điểm j = 1..6 và lặp lại trên ảnh đã biến dạng mới cho đến khi không còn thay đổi đáng kể, cuối cùng tạo thành phép biến đổi tương đồng 2D cuối cùng: T2d := T 1 2d □... □T k 2d. Phép biến đổi tổng hợp này tạo ra một vùng cắt đã căn chỉnh 2D, như thể hiện ở Hình 1(b). Phương pháp căn chỉnh này tương tự như phương pháp được sử dụng trong LFW-a, vốn thường được dùng để tăng độ chính xác nhận diện. Tuy nhiên, phép biến đổi tương đồng không thể bù cho xoay ngoài mặt phẳng, điều này đặc biệt quan trọng trong các điều kiện không bị ràng buộc.
Căn chỉnh 3D Để căn chỉnh các khuôn mặt bị xoay ngoài mặt phẳng, chúng tôi sử dụng một mô hình hình dạng 3D tổng quát và đăng ký một camera affine 3D, được sử dụng để biến dạng vùng cắt đã căn chỉnh 2D lên mặt phẳng ảnh của hình dạng 3D. Điều này tạo ra phiên bản đã căn chỉnh 3D của vùng cắt như minh họa ở Hình 1(g). Việc này được thực hiện bằng cách xác định thêm 67 điểm chuẩn x2d trong vùng cắt đã căn chỉnh 2D (xem Hình 1(c)), sử dụng một SVR thứ hai. Là mô hình hình dạng 3D tổng quát, chúng tôi đơn giản lấy trung bình các bản quét 3D từ cơ sở dữ liệu USF Human-ID, đã được xử lý hậu kỳ để biểu diễn dưới dạng các đỉnh đã căn chỉnh vi = (xi, yi, zi)n i=1. Chúng tôi đặt thủ công 67 điểm neo trên hình dạng 3D, và bằng cách này đạt được sự tương ứng đầy đủ giữa 67 điểm chuẩn được phát hiện và các điểm tham chiếu 3D của chúng.
Một camera affine 3D-2D P sau đó được khớp bằng cách sử dụng nghiệm bình phương tối thiểu tổng quát cho hệ phương trình tuyến tính x2d = X3d □P với ma trận hiệp phương sai Σ đã biết, tức là, □P tối thiểu hóa hàm mất mát sau: loss(□P) = rT Σ−1r trong đó r = (x2d −X3d □P) là vector dư và X3d là ma trận (67 □2)×8 được tạo thành bằng cách xếp chồng các ma trận (2×8) [x□ 3d(i), 1,□0;□0, x□ 3d(i), 1], với □0 là một vector hàng gồm bốn số 0, cho mỗi điểm chuẩn tham chiếu x3d(i). Camera affine P kích thước 2×4 được biểu diễn bởi vector gồm 8 ẩn số □P. Hàm mất mát có thể được tối thiểu hóa bằng phân tích Cholesky của Σ, giúp biến đổi bài toán thành bình phương tối thiểu thông thường. Vì, ví dụ, các điểm được phát hiện trên đường viền khuôn mặt thường nhiễu hơn, do vị trí ước lượng của chúng bị ảnh hưởng nhiều bởi độ sâu so với góc camera, chúng tôi sử dụng ma trận hiệp phương sai (67 □2)×(67 □2) Σ được xác định bởi các hiệp phương sai ước lượng của sai số các điểm chuẩn.
Làm thẳng mặt (Frontalization) Do các phép chiếu phối cảnh đầy đủ và các biến dạng phi tuyến không được mô hình hóa, camera được khớp P chỉ là một xấp xỉ. Để giảm sự sai lệch của các yếu tố quan trọng mang đặc trưng nhận diện đến phép biến dạng cuối cùng, chúng tôi cộng các phần dư tương ứng trong r vào các thành phần x-y của mỗi điểm chuẩn tham chiếu x3d, ký hiệu là f x3d. Sự nới lỏng này là hợp lý nhằm mục đích biến dạng ảnh 2D với ít biến dạng hơn đối với đặc trưng nhận diện. Nếu không có bước này, các khuôn mặt sẽ bị biến dạng thành cùng một hình dạng trong 3D, làm mất đi các yếu tố phân biệt quan trọng. Cuối cùng, việc làm thẳng mặt được thực hiện bằng phép biến đổi từng phần tuyến tính (piece-wise affine transformation) T từ x2d (nguồn) sang f x3d (đích), được dẫn hướng bởi phép tam giác hóa Delaunay dựa trên 67 điểm chuẩn1. Ngoài ra, các tam giác không nhìn thấy được so với camera P có thể được thay thế bằng cách trộn ảnh với các tam giác đối xứng tương ứng của chúng.

3. Đại diện

Trong những năm gần đây, tài liệu về thị giác máy tính đã thu hút nhiều nỗ lực nghiên cứu trong việc thiết kế bộ mô tả đặc trưng. Các bộ mô tả này, khi được áp dụng cho nhận diện khuôn mặt, phần lớn sử dụng cùng một toán tử cho tất cả các vị trí trên ảnh khuôn mặt. Gần đây, khi có nhiều dữ liệu hơn, các phương pháp dựa trên học máy đã bắt đầu vượt trội hơn các đặc trưng được thiết kế thủ công, vì chúng có thể khám phá và tối ưu hóa các đặc trưng cho nhiệm vụ cụ thể [19]. Ở đây, chúng tôi học một biểu diễn tổng quát của ảnh khuôn mặt thông qua một mạng sâu lớn.
Kiến trúc và huấn luyện DNN Chúng tôi huấn luyện DNN của mình trên một nhiệm vụ nhận diện khuôn mặt đa lớp, cụ thể là phân loại danh tính của một ảnh khuôn mặt. Kiến trúc tổng thể được thể hiện trong Hình 2. Một ảnh khuôn mặt đã được căn chỉnh 3D với 3 kênh màu (RGB) có kích thước 152 x 152 pixel được đưa vào một lớp tích chập (C1) với 32 bộ lọc có kích thước 11x11x3 (chúng tôi ký hiệu là 32x11x11x3@152x152). 32 bản đồ đặc trưng thu được sau đó được đưa vào một lớp gộp cực đại (M2), lớp này lấy giá trị lớn nhất trên các vùng lân cận không gian 3x3 với bước nhảy là 2, riêng biệt cho từng kênh. Tiếp theo là một lớp tích chập khác (C3) với 16 bộ lọc có kích thước 9x9x16. Mục đích của ba lớp này là để trích xuất các đặc trưng mức thấp, như các cạnh đơn giản và kết cấu. Các lớp gộp cực đại giúp đầu ra của mạng tích chập trở nên bền vững hơn với các phép tịnh tiến cục bộ. Khi áp dụng cho các ảnh khuôn mặt đã căn chỉnh, chúng giúp mạng trở nên bền vững hơn với các lỗi căn chỉnh nhỏ. Tuy nhiên, nhiều mức gộp cực đại sẽ khiến mạng mất thông tin về vị trí chính xác của các cấu trúc khuôn mặt chi tiết và vi kết cấu. Do đó, chúng tôi chỉ áp dụng gộp cực đại cho lớp tích chập đầu tiên. Chúng tôi xem các lớp đầu tiên này như một giai đoạn tiền xử lý thích nghi ở phía trước. Mặc dù chúng chịu trách nhiệm cho phần lớn tính toán, chúng giữ

1T2d có thể được sử dụng ở đây để tránh phải qua quá trình biến dạng mất mát 2D.

Hình 2. Phác thảo kiến trúc DeepFace. Phần đầu gồm một lớp lọc chập-kết hợp gộp-chập đơn trên đầu vào đã được chỉnh hình, tiếp theo là ba lớp kết nối cục bộ và hai lớp kết nối đầy đủ. Màu sắc minh họa các bản đồ đặc trưng được tạo ra ở mỗi lớp. Mạng này bao gồm hơn 120 triệu tham số, trong đó hơn 95% đến từ các lớp kết nối cục bộ và kết nối đầy đủ.

rất ít tham số. Các tầng này chỉ đơn giản mở rộng đầu vào thành một tập hợp các đặc trưng cục bộ đơn giản. Các tầng tiếp theo (L4, L5 và L6) thay vào đó được kết nối cục bộ [13, 16], giống như một tầng tích chập, chúng áp dụng một bộ lọc, nhưng mỗi vị trí trong bản đồ đặc trưng sẽ học một tập hợp bộ lọc khác nhau. Vì các vùng khác nhau của một ảnh đã được căn chỉnh có các thống kê cục bộ khác nhau, giả định về tính đồng nhất không gian của tích chập không còn đúng. Ví dụ, các vùng giữa mắt và lông mày có ngoại hình rất khác biệt và có khả năng phân biệt cao hơn nhiều so với các vùng giữa mũi và miệng. Nói cách khác, chúng tôi tùy chỉnh kiến trúc của DNN bằng cách tận dụng thực tế rằng các ảnh đầu vào của chúng tôi đã được căn chỉnh. Việc sử dụng các tầng cục bộ không ảnh hưởng đến gánh nặng tính toán của việc trích xuất đặc trưng, nhưng lại ảnh hưởng đến số lượng tham số cần huấn luyện. Chỉ vì chúng tôi có một bộ dữ liệu lớn đã được gán nhãn, chúng tôi mới có thể sử dụng ba tầng kết nối cục bộ lớn. Việc sử dụng các tầng kết nối cục bộ (không chia sẻ trọng số) cũng có thể được lý giải bởi thực tế rằng mỗi đơn vị đầu ra của một tầng kết nối cục bộ bị ảnh hưởng bởi một vùng rất lớn của đầu vào. Ví dụ, đầu ra của L6 bị ảnh hưởng bởi một vùng 74x74x3 ở đầu vào, và hầu như không có sự chia sẻ thống kê nào giữa các vùng lớn như vậy trên các khuôn mặt đã căn chỉnh. Cuối cùng, hai tầng trên cùng (F7 và F8) là các tầng kết nối đầy đủ: mỗi đơn vị đầu ra được kết nối với tất cả các đầu vào. Các tầng này có khả năng nắm bắt các mối tương quan giữa các đặc trưng được trích xuất ở các phần xa nhau trên ảnh khuôn mặt, ví dụ, vị trí và hình dạng của mắt và vị trí, hình dạng của miệng. Đầu ra của tầng kết nối đầy đủ đầu tiên (F7) trong mạng sẽ được sử dụng làm vector đặc trưng biểu diễn khuôn mặt thô của chúng tôi trong suốt bài báo này. Về mặt biểu diễn, điều này trái ngược với các phương pháp biểu diễn dựa trên LBP hiện có được đề xuất trong tài liệu, vốn thường tổng hợp các mô tả cục bộ (bằng cách tính histogram) và sử dụng chúng làm đầu vào cho bộ phân loại. Đầu ra của tầng kết nối đầy đủ cuối cùng được đưa vào một softmax K lớp (trong đó K là số lượng lớp) để tạo ra một phân phối trên các nhãn lớp. Nếu chúng ta ký hiệu ok là đầu ra thứ k của mạng trên một đầu vào nhất định, xác suất gán cho lớp thứ k là đầu ra của hàm softmax: pk = exp(ok)/ P

h exp(oh).

Mục tiêu của quá trình huấn luyện là tối đa hóa xác suất của lớp đúng (face id). Chúng tôi đạt được điều này bằng cách tối thiểu hóa hàm mất mát cross-entropy cho mỗi mẫu huấn luyện. Nếu k là chỉ số của nhãn đúng cho một đầu vào nhất định, thì hàm mất mát là: L = –log pk. Hàm mất mát này được tối thiểu hóa đối với các tham số bằng cách tính gradient của L theo các tham số và cập nhật các tham số đó bằng phương pháp descent gradient ngẫu nhiên (SGD). Các gradient được tính toán bằng phương pháp lan truyền ngược lỗi tiêu chuẩn [25, 21]. Một đặc điểm thú vị của các đặc trưng được tạo ra bởi mạng này là chúng rất thưa (sparse). Trung bình, 75% thành phần đặc trưng ở các lớp trên cùng có giá trị đúng bằng 0. Điều này chủ yếu là do việc sử dụng hàm kích hoạt ReLU [10]: max(0, x). Hàm phi tuyến ngưỡng mềm này được áp dụng sau mỗi lớp tích chập, lớp kết nối cục bộ và lớp kết nối đầy đủ (trừ lớp cuối cùng), khiến toàn bộ chuỗi mạng tạo ra các đặc trưng phi tuyến và thưa mạnh. Độ thưa cũng được khuyến khích bởi phương pháp regularization gọi là dropout [19], phương pháp này đặt ngẫu nhiên các thành phần đặc trưng về 0 trong quá trình huấn luyện. Chúng tôi chỉ áp dụng dropout cho lớp fully-connected đầu tiên. Do bộ dữ liệu huấn luyện lớn, chúng tôi không quan sát thấy hiện tượng overfitting đáng kể trong quá trình huấn luyện2.
Với một ảnh I, biểu diễn G(I) sau đó được tính toán bằng mạng feed-forward đã mô tả ở trên. Bất kỳ mạng neural feed-forward nào với L lớp đều có thể được xem như một tổ hợp các hàm gl ϕ. Trong trường hợp của chúng tôi, biểu diễn là: G(I) = gF7 ϕ (gL6 ϕ (...gC1 ϕ (T(I, θT))...)) với các tham số của mạng là pa-

rameters ϕ = {C1, ..., F7} và θT = {x2d, P,r} như đã mô tả trong Phần 2. Chuẩn hóa Ở giai đoạn cuối cùng, chúng tôi chuẩn hóa các đặc trưng để nằm trong khoảng từ 0 đến 1 nhằm giảm độ nhạy với sự thay đổi về chiếu sáng: Mỗi thành phần của vector đặc trưng được chia cho giá trị lớn nhất của nó trên toàn bộ tập huấn luyện. Sau đó, chúng tôi thực hiện chuẩn hóa L2: f(I) := ¯G(I)/|| ¯G(I)||2, trong đó ¯G(I)i = G(I)i/ max(Gi, ϵ) 3. Vì chúng tôi sử dụng các hàm kích hoạt ReLU, hệ thống của chúng tôi không bất biến với việc thay đổi tỷ lệ cường độ ảnh. Nếu không có bi-

2Xem tài liệu bổ sung để biết thêm chi tiết. 3ϵ = 0,05 để tránh chia cho một số nhỏ.

trong các trường hợp trong DNN, tính tương đương hoàn hảo sẽ đạt được.

4. Chỉ số xác minh

Xác minh xem hai mẫu đầu vào có thuộc cùng một lớp (danh tính) hay không đã được nghiên cứu rộng rãi trong lĩnh vực nhận diện khuôn mặt không bị ràng buộc, với các phương pháp có giám sát cho thấy lợi thế hiệu suất rõ rệt so với các phương pháp không giám sát. Bằng cách huấn luyện trên tập huấn luyện của miền mục tiêu, người ta có thể tinh chỉnh một vector đặc trưng (hoặc bộ phân loại) để hoạt động tốt hơn trong phân phối cụ thể của bộ dữ liệu. Ví dụ, LFW có khoảng 75% là nam giới, những người nổi tiếng được chụp ảnh chủ yếu bởi các nhiếp ảnh gia chuyên nghiệp. Như đã được chứng minh trong [5], việc huấn luyện và kiểm tra trong các phân phối miền khác nhau làm giảm hiệu suất đáng kể và đòi hỏi phải tinh chỉnh thêm cho biểu diễn (hoặc bộ phân loại) để cải thiện khả năng tổng quát hóa và hiệu suất của chúng. Tuy nhiên, việc khớp một mô hình với một bộ dữ liệu tương đối nhỏ sẽ làm giảm khả năng tổng quát hóa của nó đối với các bộ dữ liệu khác. Trong nghiên cứu này, chúng tôi hướng tới việc học một metric không giám sát có khả năng tổng quát hóa tốt cho nhiều bộ dữ liệu. Độ tương đồng không giám sát của chúng tôi đơn giản là tích vô hướng giữa hai vector đặc trưng đã được chuẩn hóa. Chúng tôi cũng đã thử nghiệm với một metric có giám sát, độ tương đồng χ2 và mạng Siamese.

4.1. Khoảng cách χ2 có trọng số

Vector đặc trưng DeepFace đã được chuẩn hóa trong phương pháp của chúng tôi có một số điểm tương đồng với các đặc trưng dựa trên histogram, chẳng hạn như LBP [1]: (1) Nó chứa các giá trị không âm, (2) nó rất thưa, và (3) các giá trị của nó nằm trong khoảng [0, 1]. Do đó, tương tự như [1], chúng tôi sử dụng độ tương đồng weighted-χ2: $\sum_i w_i(f1[i] - f2[i])^2/(f1[i] + f2[i])$ trong đó f1 và f2 là các biểu diễn DeepFace. Các tham số trọng số được học bằng cách sử dụng SVM tuyến tính, áp dụng cho các vector của các phần tử $(f1[i] - f2[i])^2/(f1[i] + f2[i])$.

4.2. Mạng Siamese

Chúng tôi cũng đã thử nghiệm một phương pháp học metric end-to-end, được gọi là mạng Siamese [8]: sau khi được huấn luyện, mạng nhận diện khuôn mặt (không bao gồm lớp trên cùng) được nhân đôi (mỗi bản cho một ảnh đầu vào) và các đặc trưng được sử dụng để dự đoán trực tiếp liệu hai ảnh đầu vào có thuộc về cùng một người hay không. Điều này được thực hiện bằng cách: a) lấy giá trị tuyệt đối của hiệu giữa các đặc trưng, sau đó b) sử dụng một lớp fully connected trên cùng để ánh xạ thành một đơn vị logistic duy nhất (cùng/không cùng). Mạng này có số lượng tham số xấp xỉ bằng mạng gốc, vì phần lớn các tham số được chia sẻ giữa hai bản sao, nhưng yêu cầu gấp đôi lượng tính toán. Lưu ý rằng để tránh overfitting trên tác vụ xác minh khuôn mặt, chúng tôi chỉ cho phép huấn luyện hai lớp trên cùng. Khoảng cách do mạng Siamese sinh ra là: $d(f1, f2) = P$ $i \alpha_i |f1[i] - f2[i]|$, trong đó α_i là

Hình 3. Các đường cong ROC trên bộ dữ liệu LFW. Xem tốt nhất ở chế độ màu.

các tham số có thể huấn luyện được. Các tham số của mạng Siamese được huấn luyện bằng hàm mất mát cross entropy tiêu chuẩn và lan truyền ngược của lỗi.

5. Thí nghiệm

Chúng tôi đánh giá hệ thống DeepFace được đề xuất bằng cách học biểu diễn khuôn mặt trên một bộ dữ liệu khuôn mặt có gán nhãn quy mô rất lớn được thu thập trực tuyến. Trong phần này, trước tiên chúng tôi giới thiệu các bộ dữ liệu được sử dụng trong các thí nghiệm, sau đó trình bày đánh giá chi tiết và so sánh với các phương pháp tiên tiến nhất, cũng như một số nhận định và phát hiện về việc học và chuyển giao các biểu diễn khuôn mặt sâu.

5.1. Bộ dữ liệu

Đại diện khuôn mặt được đề xuất được học từ một bộ sưu tập lớn các bức ảnh từ Facebook, được gọi là bộ dữ liệu Phân loại Khuôn mặt Xã hội (SFC).
Các đại diện này sau đó được áp dụng cho cơ sở dữ liệu Labeled Faces in the Wild (LFW), vốn là bộ dữ liệu chuẩn de facto cho việc xác thực khuôn mặt trong môi trường không bị ràng buộc, và bộ dữ liệu YouTube Faces (YTF), được xây dựng tương tự như LFW nhưng tập trung vào các đoạn video.
Bộ dữ liệu SFC bao gồm 4,4 triệu khuôn mặt đã được gán nhãn từ 4.030 người, mỗi người có từ 800 đến 1.200 khuôn mặt, trong đó 5% hình ảnh khuôn mặt gần đây nhất của mỗi danh tính được để lại cho việc kiểm tra. Việc này được thực hiện dựa trên dấu thời gian của các hình ảnh nhằm mô phỏng việc nhận diện liên tục theo thời gian (do lão hóa).
Số lượng lớn hình ảnh trên mỗi người mang lại cơ hội đặc biệt để học được tính bất biến cần thiết cho bài toán cốt lõi của nhận diện khuôn mặt. Chúng tôi đã xác thực bằng nhiều phương pháp tự động rằng các danh tính được sử dụng cho huấn luyện không trùng lặp với bất kỳ danh tính nào trong các bộ dữ liệu được đề cập bên dưới, bằng cách kiểm tra nhãn tên của họ.

Bộ dữ liệu LFW [18] bao gồm 13.323 ảnh web của 5.749 người nổi tiếng, được chia thành 6.000 cặp khuôn mặt trong 10 phần. Hiệu suất được đo bằng độ chính xác nhận diện trung bình sử dụng A) giao thức hạn chế, trong đó chỉ có nhãn giống và không giống được sử dụng trong huấn luyện; B) giao thức không hạn chế, nơi các cặp huấn luyện bổ sung có thể truy cập được trong quá trình huấn luyện; và C) thiết lập không giám sát, trong đó không thực hiện bất kỳ huấn luyện nào trên các ảnh LFW.
Bộ dữ liệu YTF [30] thu thập 3.425 video YouTube của 1.595 đối tượng (một tập con của những người nổi tiếng trong LFW). Những video này được chia thành 5.000 cặp video và 10 phần, được sử dụng để đánh giá xác thực khuôn mặt ở cấp độ video.
Các danh tính khuôn mặt trong SFC được gán nhãn bởi con người, thường có khoảng 3% lỗi. Ảnh khuôn mặt trên mạng xã hội có sự biến đổi lớn hơn nữa về chất lượng hình ảnh, ánh sáng và biểu cảm so với các ảnh web của người nổi tiếng trong LFW và YTF, vốn thường được chụp bởi các nhiếp ảnh gia chuyên nghiệp thay vì điện thoại thông minh4.

5.2. Đào tạo về SFC

Chúng tôi đầu tiên huấn luyện mạng nơ-ron sâu trên SFC như một bài toán phân loại đa lớp sử dụng một công cụ dựa trên GPU, thực hiện lan truyền ngược tiêu chuẩn trên các mạng truyền thẳng bằng phương pháp giảm dần theo gradient ngẫu nhiên (SGD) với động lượng (đặt là 0.9). Kích thước mini-batch của chúng tôi là 128, và chúng tôi đã đặt cùng một tốc độ học cho tất cả các lớp có thể huấn luyện là 0.01, tốc độ này được giảm thủ công, mỗi lần giảm một bậc khi lỗi xác thực không còn giảm nữa, cho đến tốc độ cuối cùng là 0.0001. Chúng tôi khởi tạo trọng số ở mỗi lớp từ một phân phối Gauss có trung bình bằng 0 với $\sigma = 0.01$, và các bias được đặt là 0.5. Chúng tôi đã huấn luyện mạng trong khoảng 15 vòng lặp (epoch) trên toàn bộ dữ liệu, mất khoảng 3 ngày. Như đã mô tả ở Mục 3, các phản hồi của lớp fully connected F7 được trích xuất để làm biểu diễn khuôn mặt.
Chúng tôi đã đánh giá các lựa chọn thiết kế khác nhau của DNN dựa trên lỗi phân loại trên 5% dữ liệu của SFC làm tập kiểm tra. Điều này xác nhận sự cần thiết của việc sử dụng một bộ dữ liệu khuôn mặt quy mô lớn và một kiến trúc sâu. Đầu tiên, chúng tôi thay đổi kích thước tập huấn luyện/kiểm tra bằng cách sử dụng một tập con của các đối tượng trong SFC. Các tập con có kích thước 1.5K, 3K và 4K đối tượng (tương ứng 1.5M, 3.3M và 4.4M khuôn mặt) được sử dụng. Sử dụng kiến trúc trong Hình 2, chúng tôi huấn luyện ba mạng, ký hiệu là DF-1.5K, DF-3.3K và DF-4.4K. Bảng 1 (cột bên trái) cho thấy lỗi phân loại chỉ tăng nhẹ từ 7.0% trên 1.5K đối tượng lên 7.2% khi phân loại 3K đối tượng, điều này cho thấy khả năng của mạng có thể đáp ứng tốt với quy mô 3 triệu ảnh huấn luyện. Tỷ lệ lỗi tăng lên 8.7% cho 4K đối tượng với 4.4 triệu ảnh, cho thấy mạng vẫn mở rộng tốt với số lượng đối tượng lớn hơn. Chúng tôi cũng đã thay đổi tổng số mẫu trong SFC lên 10%, 20%, 50%, giữ nguyên số lượng danh tính, được ký hiệu là DF-10%, DF-20%, DF-50% ở cột giữa của Bảng 1. Chúng tôi quan sát thấy lỗi kiểm tra tăng lên đến 20,7%, do hiện tượng overfitting trên tập huấn luyện bị giảm. Vì hiệu suất chưa đạt mức bão hòa ở 4 triệu ảnh, điều này cho thấy mạng lưới sẽ được hưởng lợi từ các bộ dữ liệu lớn hơn nữa.
Chúng tôi cũng thay đổi độ sâu của các mạng bằng cách cắt bỏ lớp C3, hai lớp cục bộ L4 và L5, hoặc cả 3 lớp này, lần lượt được gọi là DF-sub1, DF-sub2 và DF-sub3. Ví dụ, chỉ còn bốn lớp có thể huấn luyện trong DF-sub3, đây là một cấu trúc nông hơn đáng kể so với 9 lớp của mạng được đề xuất trong Hình 2. Khi huấn luyện các mạng này với 4,4 triệu khuôn mặt, lỗi phân loại ngừng giảm sau vài epoch và duy trì ở mức cao hơn so với mạng sâu, như có thể thấy ở Bảng 1 (cột bên phải). Điều này xác nhận sự cần thiết của độ sâu mạng khi huấn luyện trên một bộ dữ liệu khuôn mặt lớn.

5.3. Kết quả trên bộ dữ liệu LFW

Cộng đồng thị giác máy tính đã đạt được tiến bộ đáng kể trong việc xác minh khuôn mặt trong các môi trường không bị ràng buộc trong những năm gần đây. Độ chính xác nhận diện trung bình trên LFW [18] đang tiến dần đến mức hiệu suất của con người là trên 97,5% [20]. Với một số trường hợp rất khó do ảnh hưởng của lão hóa, sự thay đổi lớn về ánh sáng và tư thế khuôn mặt trong LFW, bất kỳ sự cải thiện nào so với phương pháp hiện đại đều rất đáng chú ý và hệ thống phải được cấu thành từ các mô-đun được tối ưu hóa cao. Có một hiệu ứng lợi nhuận giảm dần mạnh mẽ và bất kỳ tiến bộ nào hiện nay đều đòi hỏi nỗ lực đáng kể để giảm số lượng lỗi của các phương pháp hiện đại. DeepFace kết hợp các mô hình lớn dựa trên feedforward với căn chỉnh 3D tinh vi.
Về tầm quan trọng của từng thành phần: 1) Không có frontalization: khi chỉ sử dụng căn chỉnh 2D, độ chính xác đạt được "chỉ" là 94,3%. Nếu không căn chỉnh gì cả, tức là chỉ sử dụng phần trung tâm của khuôn mặt được phát hiện, độ chính xác là 87,9% vì một số phần của vùng khuôn mặt có thể bị cắt ra ngoài. 2) Không học máy: khi chỉ sử dụng frontalization và kết hợp đơn giản LBP/SVM, độ chính xác là 91,4%, điều này đã khá đáng chú ý với sự đơn giản của bộ phân loại này.
Tất cả các ảnh LFW đều được xử lý qua cùng một quy trình đã được sử dụng để huấn luyện trên bộ dữ liệu SFC, được gọi là DeepFace-single. Để đánh giá khả năng phân biệt của biểu diễn khuôn mặt một cách độc lập, chúng tôi theo thiết lập không giám sát để so sánh trực tiếp tích trong của một cặp đặc trưng đã được chuẩn hóa. Đáng chú ý, điều này đạt được độ chính xác trung bình 95,92%, gần như ngang bằng với hiệu suất tốt nhất cho đến nay, đạt được bằng học chuyển giao có giám sát [5]. Tiếp theo, chúng tôi huấn luyện một kernel SVM (với C=1) trên vector khoảng cách $\chi 2$ (Mục 4.1) theo giao thức hạn chế, tức là chỉ có 5.400 nhãn cặp cho mỗi lần chia được sử dụng để huấn luyện SVM. Điều này đạt được độ chính xác 97,00%, giảm đáng kể lỗi của các phương pháp hiện đại [7, 5], xem Bảng 3.

4Xem tài liệu bổ sung để biết thêm chi tiết về SFC.

Mạng Lỗi Mạng Lỗi Mạng Lỗi DF-1.5K 7,00% DF-10% 20,7% DF-sub1 11,2% DF-3.3K 7,22% DF-20% 15,1% DF-sub2 12,6% DF-4.4K 8,74% DF-50% 10,9% DF-sub3 13,5%

Bảng 1. So sánh sai số phân loại trên SFC theo kích thước tập dữ liệu huấn luyện và độ sâu mạng. Xem Mục 5.2 để biết chi tiết.

Mạng lưới Lỗi (SFC) Độ chính xác ± SE (LFW) DeepFace-align2D 9,5% 0,9430 ±0,0043 DeepFace-gradient 8,9% 0,9582 ±0,0037 DeepFace-Siamese NA 0,9617 ±0,0038

Bảng 2. Hiệu suất của các mạng DeepFace riêng lẻ khác nhau và mạng Siamese.

Tập hợp các DNN
Tiếp theo, chúng tôi kết hợp nhiều mạng được huấn luyện bằng cách cung cấp các loại đầu vào khác nhau cho DNN: 1) Mạng DeepFace-single đã mô tả ở trên dựa trên đầu vào RGB căn chỉnh 3D; 2) Ảnh mức xám cộng với độ lớn và hướng gradient của ảnh; và 3) Ảnh RGB căn chỉnh 2D. Chúng tôi kết hợp các khoảng cách này bằng một SVM phi tuyến (với C=1) với tổng đơn giản của các kernel CPD lũy thừa: KCombined := Ksingle + Kgradient + Kalign2d, trong đó K(x, y) := −||x −y||2, và theo giao thức bị giới hạn, đạt được độ chính xác 97,15%. Giao thức không giới hạn cung cấp cho người vận hành kiến thức về các danh tính trong tập huấn luyện, do đó cho phép tạo ra nhiều cặp huấn luyện hơn để thêm vào tập huấn luyện. Chúng tôi tiếp tục thử nghiệm với việc huấn luyện một Mạng Siamese (Mục 4.2) để học một metric xác thực bằng cách tinh chỉnh bộ trích xuất đặc trưng (chia sẻ) đã được huấn luyện trước của Siamese. Theo quy trình này, chúng tôi đã quan sát thấy hiện tượng overfitting đáng kể với dữ liệu huấn luyện. Các cặp huấn luyện được tạo ra bằng dữ liệu huấn luyện LFW là dư thừa vì chúng được tạo ra từ khoảng 9K ảnh, điều này không đủ để ước lượng đáng tin cậy hơn 120 triệu tham số. Để giải quyết các vấn đề này, chúng tôi đã thu thập thêm một bộ dữ liệu bổ sung theo cùng quy trình như với SFC, chứa thêm 100K danh tính mới, mỗi danh tính chỉ có 30 mẫu để tạo ra các cặp cùng và không cùng. Sau đó, chúng tôi huấn luyện Mạng Siamese trên bộ dữ liệu này, tiếp theo là 2 epoch huấn luyện trên các tập huấn luyện không giới hạn của LFW để điều chỉnh một số thiên lệch phụ thuộc vào bộ dữ liệu. Biểu diễn được tinh chỉnh nhẹ này được xử lý tương tự như trước.
Kết hợp nó vào tập hợp đã đề cập ở trên, tức là KCombined += KSiamese, đạt được độ chính xác 97,25% theo giao thức không giới hạn. Bằng cách thêm bốn mạng DeepFace-single bổ sung vào tập hợp, mỗi mạng được huấn luyện lại từ đầu với các seed ngẫu nhiên khác nhau, tức là KCombined += □ KDeepFace-Single, độ chính xác thu được là 97,35%. Hiệu suất của từng mạng riêng lẻ, trước khi kết hợp, được trình bày trong Bảng 2.
Các so sánh với các phương pháp tiên tiến gần đây...

Phương pháp Độ chính xác ± SE Giao thức Joint Bayesian [6] 0.9242 ±0.0108 bị giới hạn Tom-vs-Pete [4] 0.9330 ±0.0128 bị giới hạn High-dim LBP [7] 0.9517 ±0.0113 bị giới hạn TL Joint Bayesian [5] 0.9633 ±0.0108 bị giới hạn DeepFace-single 0.9592 ±0.0029 không giám sát DeepFace-single 0.9700 ±0.0028 bị giới hạn DeepFace-ensemble 0.9715 ±0.0027 bị giới hạn DeepFace-ensemble 0.9735 ±0.0025 không bị giới hạn Con người, đã cắt 0.9753

Bảng 3. So sánh với các phương pháp hiện đại trên bộ dữ liệu LFW.

Phương pháp Độ chính xác (%) AUC EER MBGS+SVM [31] 78.9 ±1.9 86.9 21.2 APEM+FUSION [22] 79.1 ±1.5 86.6 21.4 STFRD+PMML [9] 79.5 ±2.5 88.6 19.9 VSOF+OSS [23] 79.7 ±1.8 89.4 20.0 DeepFace-single 91.4 ±1.1 96.3 8.6

Bảng 4. So sánh với các phương pháp hiện đại nhất trên bộ dữ liệu YTF.

Các phương pháp được trình bày về độ chính xác trung bình và các đường cong ROC trong Bảng 3 và Hình 3, bao gồm cả hiệu suất của con người trên các khuôn mặt đã được cắt. Phương pháp DeepFace được đề xuất đã nâng cao trình độ hiện tại, tiến gần đến hiệu suất của con người trong việc xác minh khuôn mặt.

5.4. Kết quả trên bộ dữ liệu YTF

Chúng tôi tiếp tục kiểm chứng DeepFace trên bộ dữ liệu xác thực khuôn mặt ở cấp độ video gần đây. Chất lượng hình ảnh của các khung hình video YouTube thường kém hơn so với ảnh trên web, chủ yếu do hiện tượng mờ chuyển động hoặc khoảng cách quan sát. Chúng tôi sử dụng trực tiếp biểu diễn DeepFace-single bằng cách tạo, cho mỗi cặp video huấn luyện, 50 cặp khung hình, mỗi khung hình từ một video, và gán nhãn các cặp này là cùng hoặc không cùng theo cặp video huấn luyện. Sau đó, một mô hình $\chi 2$ có trọng số được học như trong Mục 4.1. Với một cặp kiểm tra, chúng tôi lấy mẫu ngẫu nhiên 100 cặp khung hình, mỗi khung hình từ một video, và sử dụng giá trị trung bình của độ tương đồng có trọng số đã học được.
So sánh với các phương pháp gần đây được thể hiện trong Bảng 4 và Hình 4. Chúng tôi báo cáo độ chính xác đạt 91,4%, giảm sai số của các phương pháp tốt nhất trước đó hơn 50%. Lưu ý rằng có khoảng 100 nhãn sai cho các cặp video, gần đây đã được cập nhật trên trang web YTF. Sau khi các nhãn này được sửa, DeepFace-single thực tế đạt 92,5%. Thí nghiệm này một lần nữa xác nhận rằng phương pháp DeepFace dễ dàng tổng quát hóa sang một miền mục tiêu mới.

5.5. Hiệu quả tính toán

Chúng tôi đã triển khai hiệu quả một toán tử lan truyền tiến dựa trên CPU, tận dụng cả các lệnh SIMD (Single Instruction Multiple Data) của CPU và bộ nhớ đệm của nó bằng cách khai thác tính cục bộ của các phép tính số thực (floating-point).

Hình 4. Các đường cong ROC trên bộ dữ liệu YTF. Xem tốt nhất ở chế độ màu.

trên toàn bộ các kernel và ảnh. Sử dụng một CPU Intel 2.2GHz lõi đơn, phép toán này mất 0,18 giây để trích xuất đặc trưng từ các pixel đầu vào thô. Các kỹ thuật biến dạng hiệu quả đã được triển khai để căn chỉnh; riêng việc căn chỉnh mất khoảng 0,05 giây. Tổng thể, DeepFace chạy mất 0,33 giây cho mỗi ảnh, bao gồm giải mã ảnh, phát hiện và căn chỉnh khuôn mặt, mạng lan truyền tiến, và đầu ra phân loại cuối cùng.

6. Kết luận

Một bộ phân loại khuôn mặt lý tưởng sẽ nhận diện khuôn mặt với độ chính xác chỉ có thể sánh ngang với con người. Bộ mô tả khuôn mặt cơ bản cần phải bất biến đối với tư thế, điều kiện chiếu sáng, biểu cảm và chất lượng hình ảnh. Nó cũng nên mang tính tổng quát, nghĩa là có thể áp dụng cho nhiều nhóm đối tượng khác nhau mà chỉ cần rất ít, hoặc thậm chí không cần, chỉnh sửa. Ngoài ra, các bộ mô tả ngắn được ưu tiên hơn, và nếu có thể, nên sử dụng các đặc trưng thưa. Chắc chắn, thời gian tính toán nhanh cũng là một mối quan tâm. Chúng tôi tin rằng công trình này, đi ngược lại xu hướng gần đây là sử dụng nhiều đặc trưng hơn và áp dụng kỹ thuật học metric mạnh mẽ hơn, đã giải quyết được thách thức này, thu hẹp phần lớn khoảng cách về hiệu suất. Công trình của chúng tôi chứng minh rằng việc kết hợp căn chỉnh dựa trên mô hình 3D với các mô hình truyền thẳng có dung lượng lớn có thể học hiệu quả từ nhiều ví dụ để vượt qua những nhược điểm và hạn chế của các phương pháp trước đây. Khả năng mang lại sự cải thiện rõ rệt trong nhận diện khuôn mặt cho thấy tiềm năng của sự kết hợp này có thể trở nên quan trọng trong các lĩnh vực thị giác máy tính khác.

Tài liệu tham khảo

[1] T. Ahonen, A. Hadid, và M. Pietik¨ainen. Mô tả khuôn mặt bằng các mẫu nhị phân cục bộ: Ứng dụng trong nhận diện khuôn mặt. PAMI, 2006. 3, 5 [2] O. Barkan, J. Weill, L. Wolf, và H. Aronowitz. Nhận diện khuôn mặt bằng phép nhân vectơ không gian chiều cao nhanh. Trong ICCV, 2013. 1, 2 [3] Y. Bengio. Học các kiến trúc sâu cho AI. Foundations and Trends in Machine Learning, 2009. 1 [4] T. Berg và P. N. Belhumeur. Bộ phân loại Tom-vs-pete và căn chỉnh bảo toàn danh tính cho xác thực khuôn mặt. Trong BMVC, 2012. 2, 7

[5] X. Cao, D. Wipf, F. Wen, G. Duan, và J. Sun. Một thuật toán học chuyển giao thực tiễn cho xác thực khuôn mặt. Trong ICCV, 2013. 1, 2, 5, 6, 7 [6] D. Chen, X. Cao, L. Wang, F. Wen, và J. Sun. Bayesian face revisited: Một công thức kết hợp. Trong ECCV, 2012. 2, 7 [7] D. Chen, X. Cao, F. Wen, và J. Sun. Phép lành của chiều không gian: Đặc trưng không gian cao và nén hiệu quả của nó cho xác thực khuôn mặt. Trong CVPR, 2013. 1, 2, 6, 7 [8] S. Chopra, R. Hadsell, và Y. LeCun. Học một metric tương đồng một cách phân biệt, với ứng dụng cho xác thực khuôn mặt. Trong CVPR, 2005. 2, 5 [9] Z. Cui, W. Li, D. Xu, S. Shan, và X. Chen. Kết hợp các mô tả vùng khuôn mặt mạnh mẽ thông qua học nhiều metric cho nhận diện khuôn mặt trong môi trường thực tế. Trong CVPR, 2013. 7 [10] G. E. Dahl, T. N. Sainath, và G. E. Hinton. Cải thiện mạng nơ-ron sâu cho LVCSR bằng cách sử dụng các đơn vị tuyến tính chỉnh lưu và dropout. Trong ICASSP, 2013. 4 [11] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, và A. Ng. Mạng nơ-ron sâu phân tán quy mô lớn. Trong NIPS, 2012. 2 [12] P. J. P. et al. Giới thiệu về bài toán thách thức nhận diện khuôn mặt tốt, xấu & xấu xí. Trong FG, 2011. 2 [13] K. Gregor và Y. LeCun. Sự xuất hiện của các tế bào giống phức hợp trong mạng sản phẩm thời gian với trường tiếp nhận cục bộ. arXiv:1006.0448, 2010. 4 [14] T. Hassner. Xem khuôn mặt thực tế trong 3D. Trong Hội nghị Quốc tế về Thị giác Máy tính (ICCV), Tháng 12, 2013. 2 [15] G. B. Huang, V. Jain, và E. G. Learned-Miller. Căn chỉnh kết hợp không giám sát các ảnh phức tạp. Trong ICCV, 2007. 2 [16] G. B. Huang, H. Lee, và E. Learned-Miller. Học các biểu diễn phân cấp cho xác thực khuôn mặt với mạng niềm tin sâu tích chập. Trong CVPR, 2012. 2, 4 [17] G. B. Huang, M. A. Mattar, H. Lee, và E. G. Learned-Miller. Học căn chỉnh từ đầu. Trong NIPS, trang 773–781, 2012. 2 [18] G. B. Huang, M. Ramesh, T. Berg, và E. Learned-miller. Labeled faces in the wild: Một cơ sở dữ liệu để nghiên cứu nhận diện khuôn mặt trong môi trường không bị ràng buộc. Trong Hội thảo ECCV về Khuôn mặt trong Ảnh thực tế, 2008. 1, 6 [19] A. Krizhevsky, I. Sutskever, và G. Hinton. Phân loại ImageNet với mạng nơ-ron tích chập sâu. Trong ANIPS, 2012. 1, 2, 3, 4 [20] N. Kumar, A. C. Berg, P. N. Belhumeur, và S. K. Nayar. Bộ phân loại thuộc tính và so sánh cho xác thực khuôn mặt. Trong ICCV, 2009. 6 [21] Y. LeCun, L. Bottou, Y. Bengio, và P. Haffner. Học dựa trên gradient áp dụng cho nhận diện tài liệu. Proc. IEEE, 1998. 1, 2, 4 [22] H. Li, G. Hua, Z. Lin, J. Brandt, và J. Yang. Ghép đàn hồi xác suất cho xác thực khuôn mặt biến đổi tư thế. Trong CVPR, 2013. 7 [23] H. Mendez-Vazquez, Y. Martinez-Diaz, và Z. Chai. Đặc trưng thứ tự có cấu trúc thể tích với đo lường tương đồng nền cho nhận diện khuôn mặt trong video. Trong Hội nghị Quốc tế về Sinh trắc học, 2013. 7 [24] M. Osadchy, Y. LeCun, và M. Miller. Phát hiện khuôn mặt và ước lượng tư thế cộng hưởng với các mô hình dựa trên năng lượng. JMLR, 2007. 2 [25] D. Rumelhart, G. Hinton, và R. Williams. Học các biểu diễn bằng cách lan truyền ngược lỗi. Nature, 1986. 2, 4 [26] K. Simonyan, O. M. Parkhi, A. Vedaldi, và A. Zisserman. Fisher vector faces in the wild. Trong BMVC, 2013. 2 [27] Y. Sun, X. Wang, và X. Tang. Chuỗi mạng tích chập sâu cho phát hiện điểm khuôn mặt. Trong CVPR, 2013. 2 [28] Y. Taigman và L. Wolf. Tận dụng hàng tỷ khuôn mặt để vượt qua rào cản hiệu suất trong nhận diện khuôn mặt không bị ràng buộc. arXiv:1108.1122, 2011. 2 [29] Y. Taigman, L. Wolf, và T. Hassner. Nhiều one-shot để tận dụng thông tin nhãn lớp. Trong BMVC, 2009. 2 [30] L. Wolf, T. Hassner, và I. Maoz. Nhận diện khuôn mặt trong video không bị ràng buộc với độ tương đồng nền được ghép nối. Trong CVPR, 2011. 1, 6 [31] L. Wolf và N. Levy. Điểm tương đồng SVM-minus cho nhận diện khuôn mặt trong video. Trong CVPR, 2013. 7 [32] D. Yi, Z. Lei, và S. Z. Li. Hướng tới nhận diện khuôn mặt bền vững với tư thế. Trong CVPR, 2013. 2