TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

**LỚP CỬ NHÂN TÀI NĂNG**

NGUYỄN THÀNH AN - NGUYỄN PHÁT TÀI

HỆ THỐNG DỊCH VỤ TỔNG HỢP VÀ TÌM KIẾM TRÊN VIDEO DỰA VÀO PHÁT HIỆN VÀ NHẬN BIẾT MẶT NGƯỜI

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN CNTT

TP.HCM, 2017

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

**LỚP CỬ NHÂN TÀI NĂNG**

NGUYỄN THÀNH AN 1312016

NGUYỄN PHÁT TÀI 1312504

HỆ THỐNG DỊCH VỤ TỔNG HỢP VÀ TÌM KIẾM TRÊN VIDEO DỰA VÀO PHÁT HIỆN VÀ NHẬN BIẾT MẶT NGƯỜI

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN TIN HỌC

GIÁO VIÊN HƯỚNG DẪN

PGS.TS.TrẦn Minh TriẾt – THS. NGUYỄN VINH TIỆP

NIÊN KHÓA 2013– 2017

|  |
| --- |
| **NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  Khóa luận đáp ứng yêu cầu của LV cử nhân tin học.  TpHCM, ngày …… tháng …… năm 2017  Giáo viên hướng dẫn |
| **NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN**  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  ………………………………………………………………………………  Khóa luận đáp ứng yêu cầu của LV cử nhân tin học.  TpHCM, ngày …… tháng …… năm 2017  Giáo viên phản biện |

LỜI CÁM ƠN

Chúng em xin chân thành cảm ơn Khoa Công Nghệ Thông Tin, trường Đại Học Khoa Học Tự Nhiên, Tp.HCM đã tạo điều kiện tốt cho chúng em thực hiện đề tài này.

Chúng em xin chân thành cảm ơn Thầy Trần Minh Triết, là người đã truyền đạt cho chúng em những kiến thức quý báu trong suốt quá trình học tập và luôn tận tình hướng dẫn, chỉ bảo chúng em trong suốt thời gian thực hiện đề tài. Chúng em xin cảm ơn Thầy Nguyễn Vinh Tiệp đã có những trao đổi, những chỉ dẫn giúp chúng em giải quyết các vấn đề và hoàn thiện đề tài.

Chúng em cũng xin gửi lời cảm ơn sâu sắc đến quý Thầy Cô trong Khoa đã tận tình giảng dạy, trang bị cho chúng em những kiến thức quí báu trong những năm học vừa qua.

Chúng em xin gửi lòng biết ơn sâu sắc đến Mẹ, Ba, các anh chị và bạn bè đã ủng hộ, giúp đỡ và động viên chúng em trong những lúc khó khăn cũng như trong suốt thời gian học tập và nghiên cứu.

Mặc dù chúng em đã cố gắng hoàn thành đề tài trong phạm vi và khả năng cho phép, nhưng chắc chắn sẽ không tránh khỏi những thiếu sót, kính mong sự cảm thông và tận tình chỉ bảo của quý Thầy Cô và các bạn.

Nhóm thực hiện

Nguyễn Thành An & Nguyễn Phát Tài

ĐỀ CƯƠNG CHI TIẾT

|  |  |
| --- | --- |
| **Tên Đề Tài**: Hệ thống dịch vụ tổng hợp và tìm kiếm trên video dựa vào phát hiện và nhận biết mặt người. | |
| **Giáo viên hướng dẫn**: PGS.TS. Trần Minh Triết – ThS. Nguyễn Vinh Tiệp | |
| **Thời gian thực hiện**: Từ ngày 15/11/2016 đến ngày 15/07/2017 | |
| **Sinh viên thực hiện:**  Nguyễn Thành An (1312016) – Nguyễn Phát Tài (1312504) | |
| **Loại đề tài:** Nghiên cứu lý thuyết, giải pháp kỹ thuật và xây dựng môi trường tương tác thông minh hỗ trợ xem video cho người dùng. | |
| **Nội Dung Đề Tài** (mô tả chi tiết nội dung đề tài, yêu cầu, phương pháp thực hiện, kết quả đạt được, …):  Mục tiêu của đề tài nhằm ***nghiên cứu để phát triển một API có khả năng nhận diện 500-1000 nhân vật nổi tiếng*** (nghệ sĩ, chính trị gia, doanh nhân,…). Đồng thời, đề tài trình bày một phương pháp tương tác thông minh mới, trong đó người dùng có thể ***xem các video được tóm lược và điều hướng theo khuôn mặt của các nhân vật chính*** xuất hiện trong video đó.  Đề tài cũng xây dụng một ***ứng dụng nhằm phân tích, tóm lược và chỉ mục video*** theo từng đoạn tương ứng vợi sự xuất hiện của khuôn mặt từng nhân vật. Từ đó, ứng dụng hỗ trợ người dùng ***nắm bắt nhanh nội dung video, xem những đoạn video ứng với sự xuất hiện của nhân vật yêu thích, cũng như truy vấn những video liên quan*** đến nhân vật đó.  **Nội dung thực hiện chi tiết bao gồm:**   * Nghiên cứu các công trình về phát hiện và nhận biết khuôn mặt được đề xuất và đạt độ chính xác vượt bậc trong năm 2015 – 2016. * Nghiên cứu, cài đặt lại và tinh chỉnh thuật toán SSD300 được đề xuất bởi Wei Liu et al năm 2016 để phát hiện khuôn mặt. * Nghiên cứu và phân tích cấu trúc mạng nơ-ron được đề xuất bởi Karen và Andrew năm 2015 để chỉnh sửa và huấn luyện lại trên tập dữ liệu FaceScrub đề xuất bởi Hong-Wei Ng và Stefan Winkler năm 2014. * Xây dựng server và hiện thực hóa thành Web API với hai chức năng phát hiện và nhận diện khuôn mặt. * Nghiên cứu và kỹ thuật tracking sử dụng DeepMatching đề xuất trong công trình của Philippe Weinzaepfel et al năm 2013. * Xây dựng ứng dụng phân tích, tóm lược và chỉ mục video dựa vào khuôn mặt của các nhân vật. * Thu thập các video mẫu, chạy thực nghiệm và xây dựng ứng dụng hỗ trợ người dùng xem video trong môi trường tương tác thông minh. | |
| **Kế Hoạch Thực Hiện:**  15/11/2016-30/12/2016: Nghiên cứu các công trình về phát hiện và nhận biết khuôn mặt được được đề xuất và đạt độ chính xác vượt bậc trong năm 2015 – 2016.  31/12/2016-31/01/2017: Nghiên cứu, cài đặt lại và tinh chỉnh thuật toán SSD300 được đề xuất bởi Wei Liu et al năm 2016 để phát hiện khuôn mặt.  01/02/2017-15/03/2017: Nghiên cứu và phân tích cấu trúc mạng nơ-ron được đề xuất bởi Karen và Andrew năm 2015 để chỉnh sửa và huấn luyện lại trên tập dữ liệu FaceScrub đề xuất bởi Hong-Wei Ng và Stefan Winkler năm 2014.  16/03/2017-31/03/2017: Xây dựng server và hiện thực hóa thành Web API với hai chức năng phát hiện và nhận diện khuôn mặt.  01/04/2017-30/04/2017: Nghiên cứu và kỹ thuật tracking sử dụng DeepMatching đề xuất trong công trình của Philippe Weinzaepfel et al năm 2013.  01/05/2017-31/05/2017: Xây dựng ứng dụng phân tích, tóm lược và chỉ mục video dựa vào khuôn mặt của các nhân vật.  01/06/2017-30/06/2017: Thu thập các video mẫu, chạy thực nghiệm và xây dựng ứng dụng hỗ trợ người dùng xem video trong môi trường tương tác thông minh.  01/07/2017-15/07/2017: Tiến hành chạy thử nghiệm. | |
| **Xác nhận của GVHD**  **PGS.TS. Trần Minh Triết** | **Ngày tháng năm** 2017  **Nhóm SV Thực hiện**  **Nguyễn Thành An – Nguyễn Phát Tài** |

MỤC LỤC

[LỜI CÁM ƠN iii](#_Toc486693943)

[ĐỀ CƯƠNG CHI TIẾT iv](#_Toc486693944)

[MỤC LỤC vii](#_Toc486693945)

[DANH MỤC CÁC HÌNH x](#_Toc486693946)

[DANH MỤC CÁC BẢNG xii](#_Toc486693947)

[TÓM TẮT ĐỀ TÀI xiii](#_Toc486693948)

[Chương 1 Mở đầu 1](#_Toc486693949)

[1.1. Giới thiệu chung 1](#_Toc486693950)

[1.2. Hệ thống tương tác thông minh 3](#_Toc486693951)

[1.3. Lý do thực hiện đề tài 5](#_Toc486693952)

[1.4. Mục tiêu đề tài 6](#_Toc486693953)

[1.5. Nội dung đề tài 7](#_Toc486693954)

[Chương 2 Các công trình và tập dữ liệu liên quan 9](#_Toc486693955)

[2.1. Tổng quan 9](#_Toc486693956)

[2.2. Các công trình tiêu biểu về phát hiện và nhận biết mặt người 9](#_Toc486693957)

[2.2.1. Các công trình phát hiện mặt người 9](#_Toc486693958)

[2.2.2. Các công trình nhận biết mặt người 9](#_Toc486693959)

[2.3. Phát hiện vật thể bằng SSD300 14](#_Toc486693960)

[2.3.1. Cấu trúc SSD300 15](#_Toc486693961)

[2.3.2. Kết quả thực nghiệm được công bố 17](#_Toc486693962)

[2.4. Nhận biết mặt người bằng DNN – VGG16 17](#_Toc486693963)

[2.4.1. Cấu trúc VGG16 18](#_Toc486693964)

[2.4.2. Kết quả thực nghiệm được công bố 20](#_Toc486693965)

[2.5. Các tập dữ liệu liên quan 21](#_Toc486693966)

[2.5.1. Khảo sát các tập dữ liệu 22](#_Toc486693967)

[2.5.2. Phân tích tập dữ liệu FaceScrub 26](#_Toc486693968)

[2.6. Kết luận 30](#_Toc486693969)

[Chương 3 Huấn luyện mô hình phát hiện và nhận biết mặt người 31](#_Toc486693970)

[3.1. Mô hình phát hiện mặt người bằng SSD300 31](#_Toc486693971)

[3.1.1. Tinh chỉnh SSD300 31](#_Toc486693972)

[3.1.2. Xây dựng tập dữ liệu 32](#_Toc486693973)

[3.1.3. Huấn luyện và kết quả 32](#_Toc486693974)

[3.2. Mô hình nhận biết mặt người bằng VGG-16 Deep features 32](#_Toc486693975)

[3.2.1. Áp dụng kỹ thuật transfer learning 32](#_Toc486693976)

[3.2.2. Network định danh VGG16-Deep-Feature 34](#_Toc486693977)

[3.2.3. Huấn luyện và kết quả 34](#_Toc486693978)

[3.3. Kết luận 34](#_Toc486693979)

[Chương 4 Các phân hệ trong hệ thống tương tác thông minh 35](#_Toc486693980)

[4.1. Face Web APIs 35](#_Toc486693981)

[4.1.1. Kiến trúc hệ thống 36](#_Toc486693982)

[4.1.2. Đặc tả APIs 37](#_Toc486693983)

[4.1.3. Demo Website 40](#_Toc486693984)

[4.2. Thông tin tổng hợp từ video số dựa trên phát hiện và nhận biết mặt người 43](#_Toc486693985)

[4.2.1. Dữ liệu đầu ra dạng thô của hệ thống 43](#_Toc486693986)

[4.2.2. Ứng dụng Smart Video Editor – tổng hợp thông tin video 44](#_Toc486693987)

[4.3. Person-based news highlight 47](#_Toc486693988)

[4.3.1. Ngữ cảnh sử dụng 47](#_Toc486693989)

[4.3.2. Giao diện và hệ thống chức năng 48](#_Toc486693990)

[4.4. Character-based movie synopsis 49](#_Toc486693991)

[4.4.1. Ngữ cảnh sử dụng 50](#_Toc486693992)

[4.4.2. Hệ thống chức năng 50](#_Toc486693993)

[4.5. Character-based filter 50](#_Toc486693994)

[4.5.1. Ngữ cảnh sử dụng 50](#_Toc486693995)

[4.5.2. Kiến trúc hệ thống 50](#_Toc486693996)

[4.5.3. Hệ thống chức năng 50](#_Toc486693997)

[4.6. Kết luận 50](#_Toc486693998)

[Chương 5 Kết luận 51](#_Toc486693999)

[5.1. Các kết quả đạt được 51](#_Toc486694000)

[5.2. Hướng phát triển của đề tài 51](#_Toc486694001)

[Tài liệu tham khảo 53](#_Toc486694002)

DANH MỤC CÁC HÌNH

[Hình 1.1. Một số ví dụ về hệ thống tương tác thông minh [5] 4](#_Toc486694003)

[Hình 2.1. Kiến trúc DeepID3 net1 và net2 [4]. 10](#_Toc486694004)

[Hình 2.2. Mô hình cấu trúc (a) và hoạt động (b) của FaceNet [9] 11](#_Toc486694005)

[Hình 2.3. Các khuôn mặt ở nhiều độ tuổi được xử lý bởi LF-CNNs [14]. 13](#_Toc486694006)

[Hình 2.4. Minh hoạt cho mô hình hoạt động của [16]. 14](#_Toc486694007)

[Hình 2.5. Cấu trúc SSD300 [19]. 15](#_Toc486694008)

[Hình 2.6. Cơ chế tính toán ở tầng phân lọai (https://deepsystems.ai) 16](#_Toc486694009)

[Hình 2.7. Cách tạo ra các vùng mặc định (https://deepsystems.ai) 17](#_Toc486694010)

[Hình 2.8. Một số ảnh mẫu trong tập FaceScrub 22](#_Toc486694011)

[Hình 2.9. Tập ảnh ví dụ cho một người trong SCFace 23](#_Toc486694012)

[Hình 2.10. Vài mẫu trong tập dữ liệu MUCT Landmarked 24](#_Toc486694013)

[Hình 2.11. Các mẫu trong tập Bosphorus 25](#_Toc486694014)

[Hình 2.12. Biểu đồ phân bố dữ liệu trong tập FaceScrub 27](#_Toc486694015)

[Hình 2.13. Ví dụ về điều kiện chiếu sáng tự do trong FaceScrub 28](#_Toc486694016)

[Hình 2.14. Ví dụ về lão hóa trong tập FaceScrub 28](#_Toc486694017)

[Hình 2.15. Ví dụ về sự khác biệt khi dùng các camera quá khác nhau. 29](#_Toc486694018)

[Hình 2.16. Ví dụ về hóa trang và trang điểm trong tập FaceScrub. 29](#_Toc486694019)

[Hình 3.1. Vai trò của các lớp trong VGG16 network 33](#_Toc486694020)

[Hình 3.2. Qui trình nhận biết mặt người tổng quát. 33](#_Toc486694021)

[Hình 4.1. Mô hình hoạt động của Face Web APIs 36](#_Toc486694022)

[Hình 4.2. Diễn viên Daniel Radcliffe vai Harry Potter trong series phim cùng tên. 39](#_Toc486694023)

[Hình 4.3. Cấu trúc Demo Website 41](#_Toc486694024)

[Hình 4.4. Ví dụ về giao diện kết quả của Detection Site (a) và Recognition Site (b) 42](#_Toc486694025)

[Hình 4.5. Một tập tin đầu ra dạng JSON. 44](#_Toc486694026)

[Hình 4.6. Ví dụ tập tin đầu ra của Smart Video Editor (khung đỏ là các thuộc tính ứng với từng diễn viên) 45](#_Toc486694027)

[Hình 4.7. Ví dụ tập “info.json” 46](#_Toc486694028)

[Hình 4.8. Giao diện hoạt động của ứng dụng Smart Video Editor. 47](#_Toc486694029)

[Hình 4.9. Màn hình Home sau khi load dữ liệu của Person-based News Highlight. 48](#_Toc486694030)

[Hình 4.10. Giao diện xem video theo các phân đoạn của diễn viên 49](#_Toc486694031)

DANH MỤC CÁC BẢNG

[Bảng 1‑1. Các lĩnh vực ứng dụng phát hiện và nhận biết mặt người [1] 2](#_Toc486694032)

[Bảng 2‑1. Cấu trúc chi tiết network VGG16 [2] 18](#_Toc486694033)

[Bảng 2‑2. So sánh kết quả các mô hình bằng LFW unrestricted setting [2] 20](#_Toc486694034)

[Bảng 2‑3. So sánh kết quả các mô hình bằng Youtube Face unrestricted setting [2]. K là số lượng người dung để nhận biết trong các video. 20](#_Toc486694035)

[Bảng 2‑4. Một số tập dữ liệu dùng cho nhận biết mặt người 22](#_Toc486694036)

[Bảng 2‑5. Bảng so sánh kích thước tập CASIA-WebFace và một số tập khác [26] 24](#_Toc486694037)

[Bảng 3‑1. Cấu trúc network đề xuất để phân lớp VGG16-Deep-Feature 34](#_Toc486694038)

[Bảng 4‑1. Bảng phân loại Face Web APIs đã phát triển. 35](#_Toc486694039)

[Bảng 4‑2. Nghi thức hoạt động của Face Web APIs. 36](#_Toc486694040)

[Bảng 4‑3. Ưu và khuyết điểm của kiến trúc hệ thống. 37](#_Toc486694041)

[Bảng 4‑4. Địa chỉ IP của các server cung cấp Face Web APIs. 38](#_Toc486694042)

[Bảng 4‑5. Kết quả trả về của các APIs dưới dạng JSON. 38](#_Toc486694043)

[Bảng 4‑6. Ví dụ kết quả phát hiện và nhận biết mặt diễn viên Daniel Radcliffe. 40](#_Toc486694044)

[Bảng 4‑7. Cấu trúc thông tin kết quả của một frame hình. 43](#_Toc486694045)

[Bảng 4‑8. Cấu trúc tập tin đầu ra dạng JSON của Smart Video Editor 44](#_Toc486694046)

[Bảng 4‑9. Cấu trúc tập tin “info.json” 46](#_Toc486694047)

TÓM TẮT ĐỀ TÀI

Hiện nay, xem tivi, phim ảnh và các video tin tức là một hình thức giải trí phổ biến trên toàn cầu. Thế nhưng vẫn còn tồn động một số khó khăn cho người xem trong việc theo dõi, nắm bắt nhanh thông tin của video mà họ quan tâm trong điều kiện làm việc công nghiệp hạn hẹp về thời gian. Để giải quyết vấn đề đó, chúng em đã nghiên cứu bài toán phát hiện và nhân diện mặt người – một mảng lớn trong lĩnh vực thị giác máy tính – để mang đến trãi nghiệm mới, một hình thức tương tác thông minh cho khán giả qua những chức năng: khái quát nội dung video theo tỷ lệ xuất hiện của các nhân vật quan trọng dựa trên nhận diện khuôn mặt, phân tích và tóm tắt các đoạn theo sự xuất hiện của diễn viên từ đó cho phép người xem truy cập nhanh đến những cảnh mà họ quan tâm và sau cùng là khả năng truy vấn các đoạn, các video liên quan đến từng nhân vật mà khán giả muốn tìm kiếm dựa trên việc nhận diện khuôn mặt.

Hệ thống ứng dụng này được đề xuất dựa trên khả năng phát hiện khuôn mặt trong các frame ảnh của video, đồng thời định danh chính xác một số lượng lớn các nhân vật, diễn viên, chính khách,… nổi tiếng trong thời gian gần đây. Do đó, chúng em đã nghiên cứu những công trình khoa học được công báo gần đây về bài toán phát hiện và nhận diện khuôn mặt để ứng dụng và tinh chỉnh cho phù hợp với mục tiêu và chức năng đề ra.

Nội dung của đề tài này tập trung vào việc *Hệ thống dịch vụ tổng hợp và tìm kiếm trên video dựa vào phát hiện và nhận biết mặt người*. Ngoài việc nghiên cứu và xây dựng hệ thống xem video thông minh này, chúng em còn mở rộng các chức năng phát hiện và nhận diện thành API dạng web để mở rộng khả năng ứng dụng về sau.

Nội dung đề tài bao gồm 5 chương:

**Chương 1**: Mở đầu

**Chương 2**: Các công trình và tập dữ liệu liên quan

**Chương 3**: Huấn luyện mô hình phát hiện và nhận biết mặt người

**Chương 4**: Các phân hệ trong hệ thống tương tác thông minh

**Chương 5**: Kết luận

# Mở đầu

* Nội dung Chương 1 trình bày tiềm năng ứng dụng của bài toán phát hiện và nhận biết mặt người, đặc biệt trong lĩnh vực giải trí. Đồng thời, nêu lên những khó khăn khi khán giả muốn nắm bắt nội dung kênh tin tức, phim ảnh mà điều kiện thời gian hạn hẹp, trong đó hướng tới sử dụng môi trường tương tác thông minh như một giải pháp. Chương 1 cũng nêu lên mục tiêu, nội dung và ý nghĩa của đề tài.

## Giới thiệu chung

Phát hiện và nhận biết mặt người là các bài toán nổi tiếng trong ngành khoa học máy tính. Đây là một chủ đề trong lĩnh vực nhận dạng mẫu (pattern recognition), liên quan mật thiết với thị giác máy tính và xử lý ảnh. Vấn đề chính của phát hiện khuôn mặt là tìm ra tọa độ và kích thước của khuôn mặt trong một bức ảnh hay một frame trong video còn nhận biết mặt người thì tập trung vào phân loại một khuôn mặt mới chưa có trong cơ sở dữ liệu vào một lớp khuôn mặt đã biết hay nói cách khác đó là bài toán định danh khuôn mặt của một người.

Đầu vào của bài toán phát hiện là ảnh tĩnh hoặc một frame của video, nó đã tìm ra tọa độ và kích thước của từng khuôn mặt. Sau đó, các khuôn mặt này lại tiếp tục được dùng làm đầu vào cho bài toán nhận biết để tìm ra định danh của chúng. Vì lẽ đó, hai bài toán này có mối liên hệ chặt chẽ với nhau trong một qui trình.

Trước năm 1990, phát hiện và nhận biết mặt người chưa được ứng dụng rộng rãi vào đời sống vì khối lượng tính toán xử lý khi vận hành các mô hình này rất lớn và sức mạnh của phần cứng máy tính lúc bấy giờ chưa cho phép thực thi trong thời gian thực. Từ sau những năm 1990, với sự phát triển vượt trội của phần cứng phát hiện và nhận biết mặt người đã được đưa vào các hệ thống để phục vụ con người và từ đó đóng một vai trò không nhỏ. Bảng 1‑1 thể hiện các lĩnh vực nổi bật nhất ứng dụng phát hiện và nhận biết mặt người.

Bảng 1‑1. Các lĩnh vực ứng dụng phát hiện và nhận biết mặt người [1]

|  |  |
| --- | --- |
| **LĨNH VỰC** | **ỨNG DỤNG** |
| An ninh (Security) | Kiểm soát ra vào tòa nhà, hệ thống boarding chuyến bay, xác thực email trên mutimedia workstation, vào ra văn phòng. |
| Hệ thống tư pháp hình sự (Criminal Justice System) | Pháp y và phân tích hiện trường. |
| Điều tra cơ sở dữ liệu hình ảnh (Image Database Investigation) | Chứng minh thư quốc gia, đăng ký phúc lợi, cơ sở dữ liệu giấy phép lái xe cho phép tìm kiếm bằng hình ảnh, benefit recipient. |
| Giám sát (Surveillance) | Giám sát và truy nã người sử dụng chất gây nghiện, kiểm soát CCTV, giám sát lưới điện, kiểm tra thông tin. |
| Các ứng dụng thẻ thông minh (Smart Card Applications) | Face prints có thể được lưu trữ trong thẻ thông minh, mã vạch, băng từ và được xác thực bằng cách so sánh ảnh thực với các mẫu trong cơ sở dữ liệu. |
| Chỉ mục video (Video Indexing) | Đánh nhãn các khuôn mặt trong video. |
| Ứng dụng cá nhân (Civilian Applications) | Sách điện tử và thương mại điện tử. |
| Tương tác người-máy (Human Computer Interactions) | Game tương tác và máy tình chủ động (proactive computer). |
| Môi trường đa phương tiện với tương tác người-máy thích nghi (Multimedia Environment with Adaptive Human Computer Interface) | Một bộ phân của các hệ thống nhận biết ngữ cảnh hoặc phổ biến (ubiquitous), nhận diện khách hàng và gợi ý sản phẩm. |

Ngày nay, xem phim ảnh, các video tin tức hay kênh truyền hình là một hình thức giải trí phổ biến trong xã hội. Thế nhưng với nhịp sống công nghiệp và quỹ thời gian vô cùng hạn hẹp thì để theo dõi xuyên suốt những thông tin, diễn viên hay nhân vật mình yêu thích là điều rất khó khăn. Lấy một ví dụ cụ thể, đối với một tập phim, nếu khán giả có thể biết được nhân vật mà họ hâm mộ có xuất hiện thường xuyên không và xuất hiện ở những phân cảnh nào thì sẽ thật tiện lợi để tập trung vào các đoạn này và lướt nhanh hoặc bỏ qua các đoạn khác, từ đó tiết kiệm được rất nhiều thời gian. Với cách thức xem phim ảnh hiên nay thì không giải quyết được vấn đề này. Chính vì thế, để mang đến những trải nghiệm tốt hơn, cần phải xây dựng một hệ thống tương tác thông minh giữa khán giả và các thiết bị trình chiếu (tivi, máy vi tính, smart phone,…), trong đó cần đảm bảo các chức năng sau. Thứ nhất, duy trì khả năng xem phim ảnh và video vốn có. Thứ hai ngoài các thông tin cơ bản đó, cần cung cấp thêm các nội dung quan trọng có liên quan như các nhân vật, diễn viên xuất hiện, tỷ lệ thời gian của từng người cũng như thời điểm phân cảnh mà họ xuất hiện để khán giả có thể chuyển nhanh đến phân cảnh được quan tâm. Thứ ba cho phép tìm kiếm các phân đoạn liên quan đến một nhân vật hay diễn viên được yêu thích trong cơ sở dữ liệu để khán giả có thể xem nhanh các phân đoạn cùng nói về một chủ đề hay con người cụ thể mà không phải xem qua tất cả nội dung video.

Để làm được điều đó thì cần giải quyết thật tốt hai bài toán phát hiện và nhận biết mặt người. Trước đây hai bài toán này được giải quyết chủ yếu dựa trên hướng tiếp cận sử dụng hand-designed features (các đặc trưng được con người tạo ra, ví dụ SIFT, SURF, HAAR, …). Trong những năm gần đây sự phát triển nhanh chóng của lĩnh vực Deep Learning mang đến một hướng tiếp cận mới và đạt những thành công vượt trội cho phát hiện và nhận biết mặt người đó là feature-learning. Trong đó, Convolutional Neural Network là một công cụ đắc lực và hiệu quả mang lại độ chính xác vượt khả năng của con người trong nhận diện. Một số công trình tiêu biểu có thể kể đến là [2], [3], [4].

## Hệ thống tương tác thông minh

Việc nghiên cứu phát triễn những hệ thống và môi trường tương tác thông minh hiện nay đang được chú trọng đầu tư phát triển. Mục đích chính của xu hướng công nghệ này là giúp cho người dùng có thể giao tiếp, điều khiển và tương tác một cách tự nhiên nhất với ứng dụng hay thiết bị đang sử dụng. Ý tưởng này đã xuất hiện từ rất lâu và dễ dàng được bắt gặp trong các phim về khoa học viễn tưởng. Trong đó, những vật dụng bình thường như bàn làm việc, cửa sổ, … trở thành những công cụ trình chiếu và cho phép người sử dụng tương tác dễ dàng như chạm hay điều khiển bằng giọng nói và những tháo tác kéo thả trên mặt kính hay những màn hình lơ lửng giữa không gian trở thành nét đặc trưng của thể loại phim này. Với xu hướng công nghệ hiện tại thì những tính đó đang dần trở thành hiện thực. Với cùng một vật dụng là cửa sổ, Nokia đưa ra ý tưởng cho việc thể hiện nội dung tin nhắn (Hình 1.1a) và Microsoft sử dụng nó để trình chiếu các thông tin kinh doanh, báo cáo doanh số Hình 1.1b).

|  |  |
| --- | --- |
| F:\NokiaVision.png | F:\MicrosoftVision.png |
| (a) Ý tưởng hiển thị tin nhắn trên kính cửa sổ của Nokia | (b) Ý tưởng hiển thị thông tin kinh doanh trên kính cửa số của Microsoft |
| http://www.intomobile.com/wp-content/uploads/2011/01/IMG_8107.jpg | D:\HOC_KY_2_NAM_4_2015-2016\1-Thesis\0-BaoCaoLuanVan\1-Báo Cáo Luận Văn\1-Chương 1\Image\MicrosoftPixelSense.jpg |
| (c) Liên kết giữa điện thoại và kiosk thông minh của Alcatel-Lucent | (d) Bàn cảm ứng Microsoft PixelSense của Microsoft |

Hình 1.1. Một số ví dụ về hệ thống tương tác thông minh [5]

Kiosk thông minh (Hình 1.1c) tích hợp công nghệ ng Connect và 4G LTE của Alcatel-Lucent giới thiệu tại CES 2011 giúp cho một chiếc điện thoại thông minh có thể kết nối và hiển thị thông tin mà người dùng quan tâm trên màn hình của kiosk, đồng thời có thể xem và mua hàng tại kiosk. Microsoft PixelSense (còn gọi là Microsoft Surface) của hãng Microsoft, cho phép đồng thời nhiều người tương tác bằng cách chạm hay đặt các vật thể lên trên màn hình và chia sẻ các nội dung số với nhiều thiết bị di động cùng lúc (Hình 1.1d).

Với ý tưởng thay đổi thói quen và cách thức xem tin tức, phim ảnh của người dùng, trong đề tài này chúng em hướng tới xây dựng một hệ thống tương tác thông minh hỗ trợ chức năng xem video, đồng thời cung cấp thêm các thông tin liên quan đến nội dung bằng cách tổng hợp dựa trên phát hiện và nhận biết mặt người. Hệ thống cho phép người dùng tương tác một cách tự nhiên để xem các thông tin liên quan một cách phù hợp với ngữ cảnh.

## Lý do thực hiện đề tài

Các công trình nghiên cứu về phát hiện và nhận biết mặt người hiện nay đã đạt đến độ chính xác rất cao, vượt qua khả năng của con người. Trong xu thế phát triển của các hệ thống và môi trường tương tác thông minh, thì vai trò của các công trình trên càng nêu cao trong rất nhiều các lĩnh vực liên quan: kinh doanh, giải trí, an ninh,… Với mục tiêu nghiên cứu tìm hiểu cấu trúc và mô hình hoạt động của các thuật toán phát hiện và nhận biết mặt người tiêu biểu hiện nay, nhóm sinh viên tập trung vào phân tích, cài đặt và vận hành các mô hình để hiểu rõ hơn nguyên lý, đồng thời tinh chỉnh cho phù hợp với nhu cầu ứng dụng trong các nội dung liên quan đến đề tài tổng hợp thông tin các video số.

Sau khi hiểu rõ được cấu trúc cài đặt và vận hành của các mô hình phát hiện và nhận biết mặt người, chúng em tập trung vào xây dựng hệ thống ứng dụng, trong đó đưa ra một hướng tiếp cận thông minh cho việc cập nhật tin tức và xem phim ảnh giải trí thông qua video số. Bài toán đặt ra cho hệ thống ứng dụng này là giúp người dùng có thể xem thêm các thông tin liên quan đến nhân vật, diễn viên xuất hiện trong các đoạn video thông qua cách tương tác tự nhiên và hợp ngữ cảnh nhất; bên cạnh đó là giúp họ xem nhanh được những phần nội dung quan trọng, được quan tâm từ đó nắm bắt các chi tiết chính yếu, tiết kiệm thời gian trong thời buổi công nghiệp và bận rộn.

Bên cạnh đó, chúng em cũng xây dựng một hệ thống API cho hai chức năng chính là phát hiện mặt người trong ảnh tĩnh và định danh khuôn mặt của một số lượng lớn các diễn viên, nghệ sĩ nổi tiếng quốc tế.

## Mục tiêu đề tài

Mục tiêu của đề tài nhầm nghiên cứu để xây dựng hệ thống dịch vụ tổng hợp và tìm kiếm trên video dựa vào phát hiện và nhận biết mặt người. Trong đó, chú trọng phân tích và tinh chỉnh các mô hình có đạt chính xác vượt bậc trong thời gian gần đây, mà kiến trúc chủ yếu dựa trên convolutional neural network nói chung và neural network nói riêng. Các mô hình phát hiện và nhận biết mặt người này được huấn luyện trên các tập dữ liệu tiêu biểu thế nhưng để đưa vào ứng dụng trong một môi trường cụ thể thì đòi hỏi có sự điều chỉnh cho phù hợp. Với mục tiêu nghiên cứu nắm vững lý thuyết kiến trúc và phát triển hệ thống ứng dụng trực quan, nhóm sinh viên thực hiện đề tài mang đến sự đóng góp trong việc hiện thực hóa và sử dụng các mô hình như giải pháp thuật toán.

Trong quá trình nghiên cứu để nắm rõ lý thuyết, nhóm sinh viên đã trình bày lại chi tiết các thông số của từng mô hình cụ thể được sử dụng cho việc phát hiện và nhận biết mặt người trong nội dung Chương 2, khảo sát các công trình và tập dữ liệu liên quan đến đề tài và báo cáo lại như một tư liệu tham khảo cho các công trình về sau. Bên cạnh đó, việc thay đổi, tạo lập tập dữ liệu mới và cách huấn luyện trong đó sử dụng kỹ thuật transfer learning để kế thừa thông tin từ mô hình nguyên bản cũng được mô tả chi tiết trong Chương 3 như một hướng tiếp cận để tham khảo.

Về mặt hiện thực hóa, nhóm sinh viên đã xây dựng một hệ thống Web APIs hoàn chỉnh cho việc phát hiện và nhận biết mặt người, trong đó tập trung vào một lượng lớn các diễn viên và nghệ sĩ nổi tiếng của thế giới. Hệ thống này được cung cấp để phục vụ cho các ứng dụng trong nội bộ đề tài cũng như định hướng cho sự phát triển về sau hoặc hỗ trợ cho các đề tài nghiên cứu, ứng dụng khác. Thông tin chi tiết và các sử dụng được mô tả chi tiết trong phần 4.1 của luận văn.

Bên cạnh đó, đề tài còn xây dựng một hệ thống tương tác thông minh và tìm kiếm video dựa trên lý thuyết phát hiện và nhận biết mặt người đã nghiên cứu (phần 4.2, 4.4, 4.5). Hệ thống này cung cấp ba phân hệ: thứ nhất là tóm tắt nội dung video, các phân đoạn của từng nhân vật chính dựa trên tỷ lệ xuất hiện, thứ hai là sự mô tả và chỉ mục chi tiết sự xuất hiện của các nhân vật trong video như một biểu đồ trực quan để người dùng có thể nắm bắt cốt truyện dựa trên luồng xuất hiện đó, thứ ba là tìm kiếm các phân đoạn liên quan đến nhân vật được quan tâm. Các phân hệ này mang đến cho người dùng một trãi nghiệm thú vị và khả năng tương tác thông minh khi xem video giải trí. Hơn thế nữa là sự tiết kiệm thời gian trong thời đại công nghiệp bận rộn.

**Nội dung thực hiện chi tiết của khóa luận bao gồm:**

• Nghiên cứu các công trình về phát hiện và nhận biết khuôn mặt được đề xuất và đạt độ chính xác vượt bậc trong năm 2015 – 2016.

• Nghiên cứu, cài đặt lại và tinh chỉnh thuật toán SSD300 được đề xuất bởi Wei Liu et al năm 2016 để phát hiện khuôn mặt.

• Nghiên cứu và phân tích cấu trúc mạng nơ-ron được đề xuất bởi Karen và Andrew năm 2015 để chỉnh sửa và huấn luyện lại trên tập dữ liệu FaceScrub đề xuất bởi Hong-Wei Ng và Stefan Winkler năm 2014.

• Xây dựng server và hiện thực hóa thành Web API với hai chức năng phát hiện và nhận diện khuôn mặt.

• Nghiên cứu và kỹ thuật tracking sử dụng DeepMatching đề xuất trong công trình của Philippe Weinzaepfel et al năm 2013.

• Xây dựng ứng dụng phân tích, tóm lược và chỉ mục video dựa vào khuôn mặt của các nhân vật.

• Thu thập các video mẫu, chạy thực nghiệm và xây dựng ứng dụng hỗ trợ người dùng xem video trong môi trường tương tác thông minh.

## Nội dung đề tài

Nội dung đề tài bao gồm 5 chương và nội dung chính từng chương như sau:

**Chương 1**: Mở đầu

Trình bày tiềm năng ứng dụng của bài toán phát hiện và nhận biết mặt người, đặc biệt trong lĩnh vực giải trí. Đồng thời, nêu lên những khó khăn khi khán giả muốn nắm bắt nội dung kênh tin tức, phim ảnh mà điều kiện thời gian hạn hẹp, trong đó hướng tới sử dụng môi trường tương tác thông minh như một giải pháp. Chương 1 cũng nêu lên mục tiêu, nội dung và ý nghĩa của đề tài.

**Chương 2**: Các công trình và tập dữ liệu liên quan

Giới thiệu một số công trình phát hiện và nhận biết mặt người tiêu biểu trong các năm gần đây. Trong đó tập trung vào hai mô hình: SSD300 dùng cho phát hiện và VGG16 dùng cho nhận biết, vì đây là hai mô hình chính được kế thừa để thực hiện đề tài. Ngoài ra chương này còn khảo sát một số tập dữ liệu liên quan và phân tích tập dữ liệu FaceScrub – được dùng để thí nghiệm với mô hình nhận biết mặt người của đề tài.

**Chương 3**: Huấn luyện mô hình phát hiện và nhận biết mặt người

Trình bày các hướng tiếp cận và giải pháp mà nhóm thực hiện đề tài triển khai trên mô hình phát hiện và nhận biết mặt người được lựa chọn (SSD300 và VGG16). Các tinh chỉnh đối với mô hình cũ, đề xuất cấu trúc mới được trình bày cụ thể theo sau sự phân tích các mô hình này. Bên cạnh đó, quá trình và chiến lược huấn luyện cũng như việc xây dựng mới một số tập dữ liệu phù hợp được mô tả chi tiết để chứng minh tính hiệu quả của hướng tiếp cận và giải pháp mà nhóm đề ra.

**Chương 4**: Các phân hệ trong hệ thống tương tác thông minh

Trình bày bốn phân hệ trong hệ thống tương tác thông minh dựa trên tổng hợp thông tin bằng phát hiện và nhận biết mặt người, bao gồm: các Face Web APIs, Person-based news highlight, Character-based movie synopsis và Character-based filter. Trong mỗi phần tương ứng, nhóm thực hiện đề tài trình bày chi tiết kiến trúc hệ thống, ngữ cảnh sử dụng và các chức năng được cung cấp cũng như hướng dẫn sử dụng và demo tương ứng.

**Chương 5**: Kết luận

Trình bày các kết quả đạt được và hướng phát triển của đề tài.

# Các công trình và tập dữ liệu liên quan

* Nội dung Chương 2 giới thiệu một số công trình phát hiện và nhận biết mặt người tiêu biểu trong các năm gần đây. Trong đó tập trung vào hai mô hình: SSD300 dùng cho phát hiện và VGG16 dùng cho nhận biết, vì đây là hai mô hình chính được kế thừa để thực hiện đề tài. Ngoài ra chương này còn khảo sát một số tập dữ liệu liên quan và phân tích tập dữ liệu FaceScrub – được dùng để thí nghiệm với mô hình nhận biết mặt người của đề tài.

## Tổng quan

Trong những năm gần đây, lĩnh vực phát hiện và nhận biết mặt người có nhiều bước phát triển vượt bậc, trong đó tiêu biểu là các mô hình đạt độ chính xác cao, vượt qua cả giới hạn của con người. Những thành tựu này có được là sự đóng góp rất lớn từ các đề tài nghiên cứu trong lĩnh vực Deep Learning mà nổi bật là các mạng neural network nói chung và convolutional neural network nói riêng. Bên cạnh đó, sự phát triển của hệ thống thiết bị ghi hình và công nghệ thu thập lưu trữ dữ liệu cũng đóng góp một phần quan trọng trong việc cung cấp các tập mẫu lớn cho việc huấn luyện các mô hình trên. Phần tiếp theo trình bày một số công trình nổi bật trong phát hiện và nhận biết mặt người (do độ dài giới hạn nên luận văn chỉ tập trung giới thiệu một số công trình tiêu biểu trong các năm gần đây).

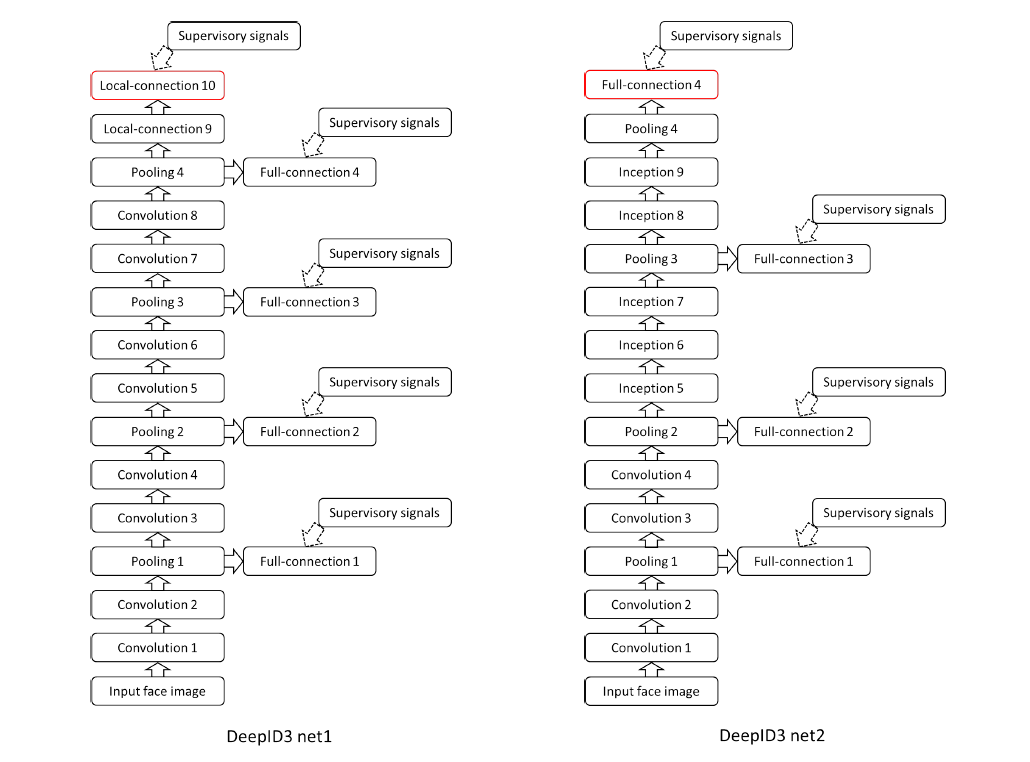
## Các công trình tiêu biểu về phát hiện và nhận biết mặt người

### Các công trình phát hiện mặt người

<Xải>Liệt kê 5 – 8 công trình tiêu biểu nhất 2015 – 2016

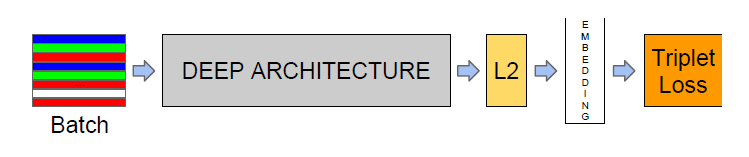
### Các công trình nhận biết mặt người

Đầu tiên phải kể đến công trình [4] với kiến trúc network được gọi là DeepID3. Với mục địch khảo sát sự hiệu quả của các neural network sâu (very deep neural network) trong việc nhận biết mặt người, Yi Sun et al đã tạo ra hai kiến trúc mới bằng các xây dựng lại các lớp convolution và inception kết chồng lên nhau được đề xuất trong VGG net [6] và GoogLeNet [7]. Điểm đặc biệt trong mô hình này là các tín hiệu kết hợp nhận biết và xác minh khuôn mặt có giám sát (joint face identification-verification supervisory signal) được thêm vào ngay sau các lớp rút trích đặc trung cuối cùng trong suốt quá trình huấn luyện mô hình. Hình 2.1 minh họa hai cấu trúc DeepID3 net1 và net2, trong đó mũi tên liền nét thể hiện hướng forward-propagation, các mũi tên nghiêng chỉ ra các lớp mà tại đó tín hiệu nhận biết và xác nhận khuôn mặt có giám sát được thêm vào, lớp rút trích đặc trưng cuối cùng trong khung màu đỏ phục vụ cho việc nhận biết khuôn mặt. Với sự cải tiến này DeepID3 đạt độ chính xác cao trên tập dữ liệu LFW [8] với 99.53% cho xác nhận khuôn mặt và 96% cho rank-1 face recognition.

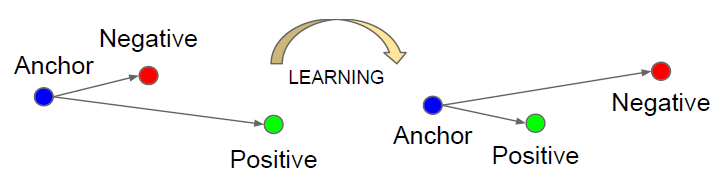


Hình 2.1. Kiến trúc DeepID3 net1 và net2 [4].

Nhóm tác giả [9] đưa ra một hệ thống gọi là FaceNet để học một ánh xạ từ các ảnh khuôn mặt vào một không gian Euclidean chặt chẽ mà ở đó khoảng cách tượng trưng trực tiếp cho sự tương đồng giữa các khuôn mặt. Minh họa cho cấu trúc và hoạt động của mô hình được thể thiện trong Hình 2.2. Với hướng tiếp cận này, các bài toán về nhận biết, xác minh và gom nhóm có thể được cài đặt dễ dàng với FaceNet embeddings (như là các vector đặc trưng). Phương pháp này sử dụng deep convolutional neural network được huấn luyện rồi để tự tối ưu ánh xạ hơn là tầng thắt cổ chai trung gian (intermediate bottleneck layer) như trong các hướng tiếp cận trước đây. Điểm nổi bật của mô hình là tính hiệu quả lớn trong việc thể hiện khuôn mặt : nhóm tác giả đạt hiệu suất vượt trội mà chỉ sử dụng 128 bytes cho một khuôn mặt. Kết quả thực nghiệm trên tập LFW [8] là 99.63% và YTF [10] là 95.12%, trong đó giảm đi 30% tỷ lệ lỗi trên cả hai tập dữ liệu so với kết quả tốt nhất được công bố trong [6].



* Mô hình cấu trúc của FaceNet



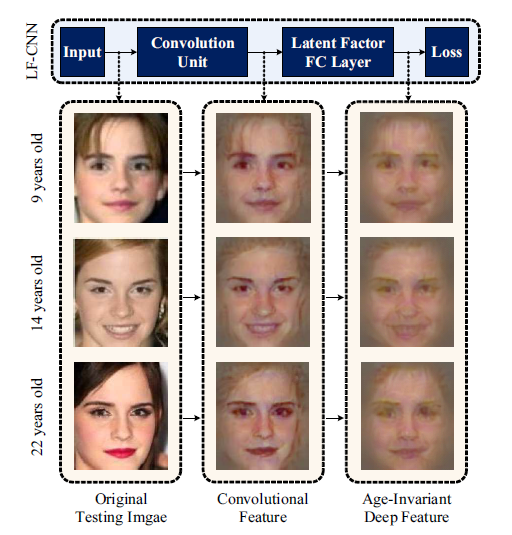
* Minh họa hoạt động của FaceNet

Hình 2.2. Mô hình cấu trúc (a) và hoạt động (b) của FaceNet [9]

Yi Sun et al đề xuất một hướng tiếp cận mới cho việc huấn luyện dữ liệu trong công trình [11]. Kiến trúc được sử dụng trong công trình này là convoluational neural network kế thừa từ baseline high-performance VGG-like deep neural network [6]. Điểm đặc trưng của mô hình mạng này là dữ liệu được huấn luyện theo chu kỳ. Mỗi lần một lớp bổ sung sẽ được sparsify và toàn bộ mô hình sẽ được huấn luyện lại với các tham số đã có từ chu kỳ trước. Độ chính xác của mô hình gốc đạt được trên tập LFW [8] là 98.95%. Với kiến trúc mới, Sparsifying Neural Network tăng độ chính xác trên cùng tập dữ liệu lên 99.30% và giảm error rate 33% trong khi chỉ giữ lại 12% tham số từ mô hình gốc. Điều này tăng tính khả dụng trên các thiết bị cấu hình thấp như mobile.

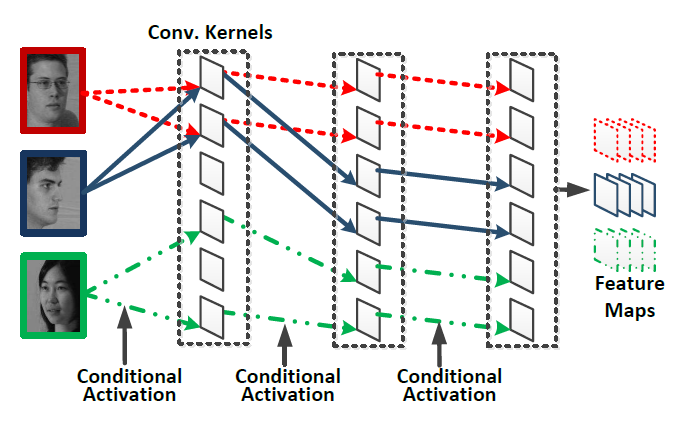
Trong [12], Iacopo Masi et al sử dụng deep convolutional neural networks để giải quyết vấn đề sự sai khác tư thế của khuôn mặt bằng cách sử dụng các mô hình xác định đa tư thế (multiple pose-specific models) và các ảnh khuôn mặt được phát sinh (render). Ý tưởng chính được sử dụng là huấn luyện theo các mô hình riêng ứng với từng tư thế của khuôn mặt, sau đó tổng hợp (fusion) các kết quả lại. Mô hình này được sử dụng cho cả xác nhận và định danh khuôn mặt. Trên tập [13], thực nghiệm đạt độ chính xác 0.895 ± 0.006 (FAR = 0.01) cho xác nhận khuôn mặt và 0.862 ± 0.0009 (Rank-1) cho nhận biết mặt người.

Hầu hết các phương pháp nhận diện đều dựa trên những đặc trưng cố định tại một độ tuổi của khuôn mặt dù là dùng hand-designed feature hay feature-learning. Các vấn đề về lão hóa và biến đổi khuôn mặt là một thách thức lớn. [14] là một công trình sử dụng neural netrok để giải quyết bài toán này. Trong đó, nhóm tác giả phát triển một latent variable model gọi là latent identity analysis (LIA) kết hợp với CNN để tìm ra các đặc trưng bất biến theo quá trình lão hóa bằng cách huấn luyện theo cặp các tham số của CNNs và LIA. Hình 2.3 minh một hoạt một ảnh ví dụ được xử lý qua các giai đoạn trong network. Giải pháp này đạt độ chính xác đáng ghi nhận: 97.51% trên tập MORPH Album2 [15] và 99.50% trên tập LFW [8].



Hình 2.3. Các khuôn mặt ở nhiều độ tuổi được xử lý bởi LF-CNNs [14].

Với cùng tư tưởng sử dụng convolutional neural network nhưng mỗi nhóm tác giả lại có cách cải biến và thiết kế cho phù hợp với từng bài toán cụ thể. Một ví dụ khác là [16]. Để xử lý cho các vấn đề của ảnh khuôn mặt như tư thế khuôn mặt, điều kiện chiếu sáng và che khuất, các tác giả đề xuất một cấu trúc convolutional neural network, trong đó sử dụng nhiều bộ kernel khác nhau và sẽ được kích hoạt tùy vào những điều kiện nhất định phụ thuộc vào ảnh đầu vào. Nói cách khác, các mẫu được xử lý bằng các kernel được kích hoạt động tùy thuộc vào dữ liệu. Tập hợp các kernel được kích hoạt xuyên suốt các lớp định hướng luồng tính toán theo từng mẫu. Mô hình hoạt động của [16] được mô tả trong Hình 2.4. Phương pháp này đạt độ chính xác 73.54% trên tập Multi-PIE [17].



Hình 2.4. Minh hoạt cho mô hình hoạt động của [16].

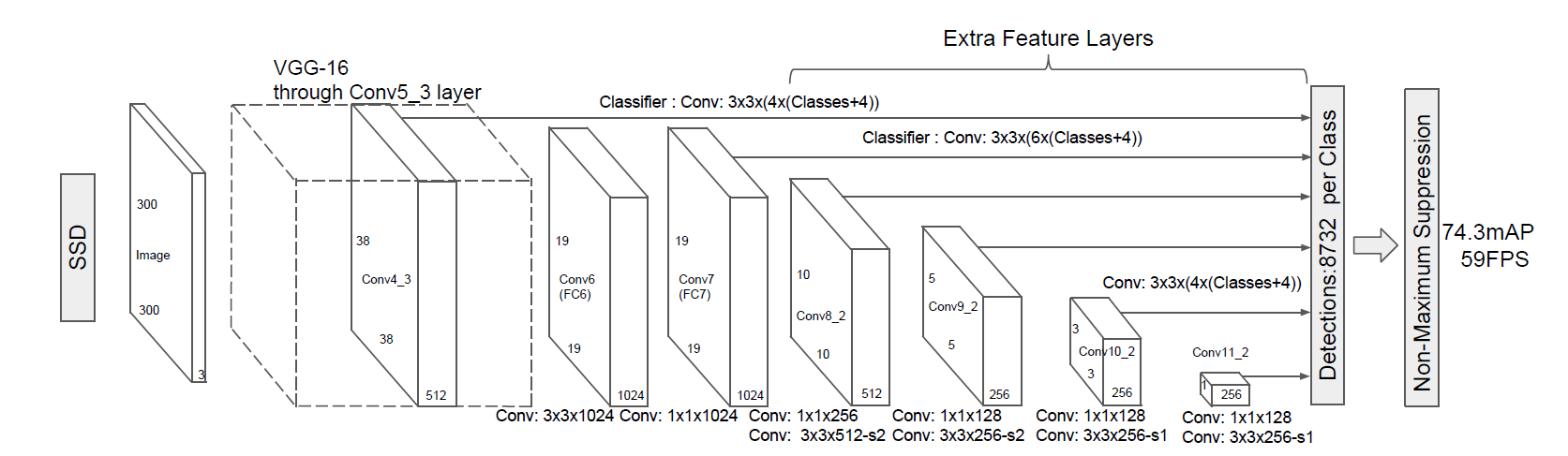
Với mục đích tạo ra các đặc trưng bất biến với các phép biến đổi phức mà có thể mô hình hóa cục bộ thành đơn nhất phục vụ cho bài toán nhận biết mặt người và ước tính tư thế khuôn mặt, nhóm tác giả [18] đề xuất một phương pháp gọi là “bells and whistles free”. Bằng cách sử dụng một phương pháp đơn giản hoạt động trên các điểm ảnh thô, [18] đạt kết quả vượt trội trên Multi-PIE database protocol [17] (75.75%), LFW [8] unsupervised protocol (91.54%) và LFW [8] image-restricted, label-free outside data protocol (88.67%). Trong đó, đề tài mang đến ba đóng góp quan trọng nhất. Thứ nhất đề xuất một hướng tiếp cận đơn giản để học các đặc trưng phi tuyến phân biệt bất biến với các phép biến đổi đơn nhất, mở rộng phạm vi lý thuyết gần đây về bất biến với đặc trưng phân biệt và kernelized. Thứ hai là đưa ra một hướng tiệp cận đơn giản dense-landmark-free tạo ra một framework có khả năng nhận biết mặt người open-set pose-invariant và ước tính tư thế khuôn mặt đồng thời. Thứ ba là đề xuất một hướng tiệp cận nối tiếp để tạo ra bất biến các đa biến đổi nội nhóm (multi sub-groups transformations) từ đó có được một framework landmark-free hoàn chỉnh cho nhận biết khuôn mặt và ước tính tư thế bất biến với các phép biến đổi.

## Phát hiện vật thể bằng SSD300

Hiện nay, các hệ thống phát hiện vật thể đều tiếp cận theo hướng chọn ra các bao chữ nhật; tính toán, rút trích đặc trưng rồi phân loại các bao đó để cho ra kết quả cuối cùng. Các phương pháp này đều dựa trên kiến trúc mạng Fast R-CNN và hiện là hướng tiếp cận tối ưu dựa vào kết quả chạy thí nghiệm trên các tập dữ liệu nổi tiếng trong lĩnh vực này như PASCAL VOC, COCO, ILSVRC. Mặc dù có độ chính xác cao nhưng các hệ thống này chưa thể cho ra kết quả theo thời gian thực. Ngay cả Fast R-CNN nhanh nhất cũng chỉ đạt ở ngưỡng 7 fps. Một số khác có thời gian chạy tốt, đổi lại độ chính xác sẽ giảm đáng kể. SSD (Single Shot multibox Detector), được đề xuất vào tháng 12 năm 2015, để giải quyết bài vần đề độ chính xác và tốc độ cho bài toán phát hiện vật thể. Cụ thể bằng thí nghiệm, kết quả cho thấy có sự cải tiến vượt bậc về tốc độ nhưng vẫn đảm bảo độ chính xác cao (59 fps với mAP 74.3% trên tập giám định VOC2007, so với Faster R-CNN 7 fps với mAP 73.2% hay YOLO 45 fps với mAP 63.4%) [19].

### Cấu trúc SSD300

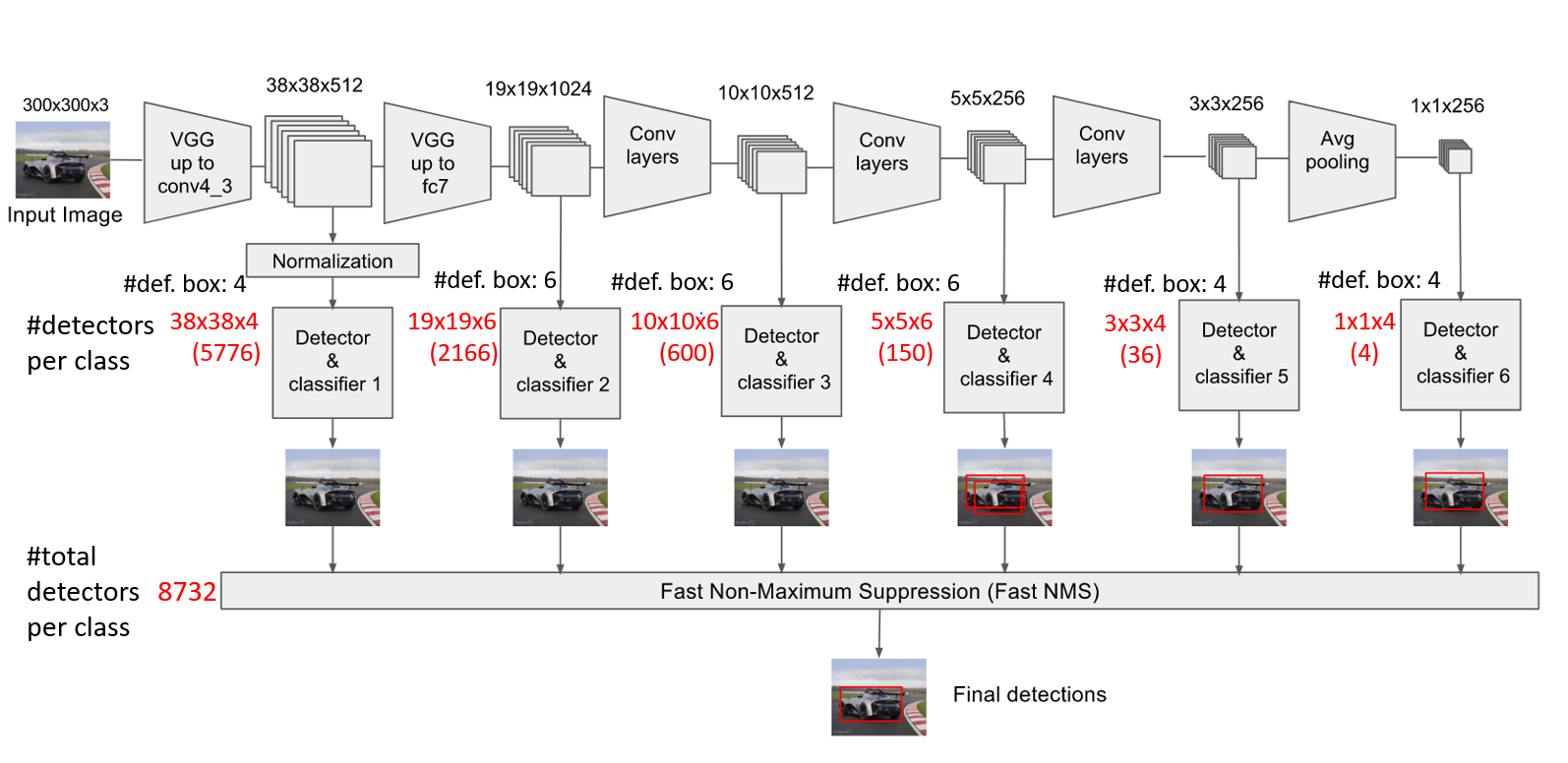
SSD được xây dựng dựa trên mô hình feed-forward convolutional network. Có cấu trúc gồm tầng đầu là mạng cơ sở, là toàn bộ mạng VGG-16 từ đầu đến lớp “fc7 – fully-connected” (trước lớp phân loại). Sau khi hình ảnh đi qua mạng cơ sở, đầu ra được phân loại bằng tầng thứ hai, trả về tập các vùng chữ nhật và trọng số cho từng vật thể trong vùng đó. Sau đó tập này được xử lý cho ra kết quả tìm thấy cuối cùng bằng thuật toán “none-maximun suppression” [19] (Xem Hình 2.5, Hình 2.6).



Hình 2.5. Cấu trúc SSD300 [19].

Các tính chất của cấu trúc mạng SSD [19]:

* Dự đoán theo vùng (Convolutional predictors for detection): Mỗi một bộ phân loại không chỉ tính toán dựa trên một ô riêng biệt mà dữa trên kết quả tổng hợp với các ô xung quanh sử dụng bộ lọc xoắn (convolutional filter). Vì thế mỗi lớp đặc trưng tạo số lượng cố định bộ phân loại. Với lớp có kích thước sau khi đi qua bộ lọc sẽ cho bộ phân loại.
* Bản đăc trưng đa tỉ lệ (Multi-scale feature maps for detection): Bộ phân loại được thêm vào dần dần từ lớp đặc trưng cục bộ “Conv5\_5” đến “Conv11\_2” (Hình 2.5), với kích thước của các lớp giảm dần. Điều này làm cho bản đặc trưng đa dạng về mối quan giữa các vùng lân cận trong ảnh.

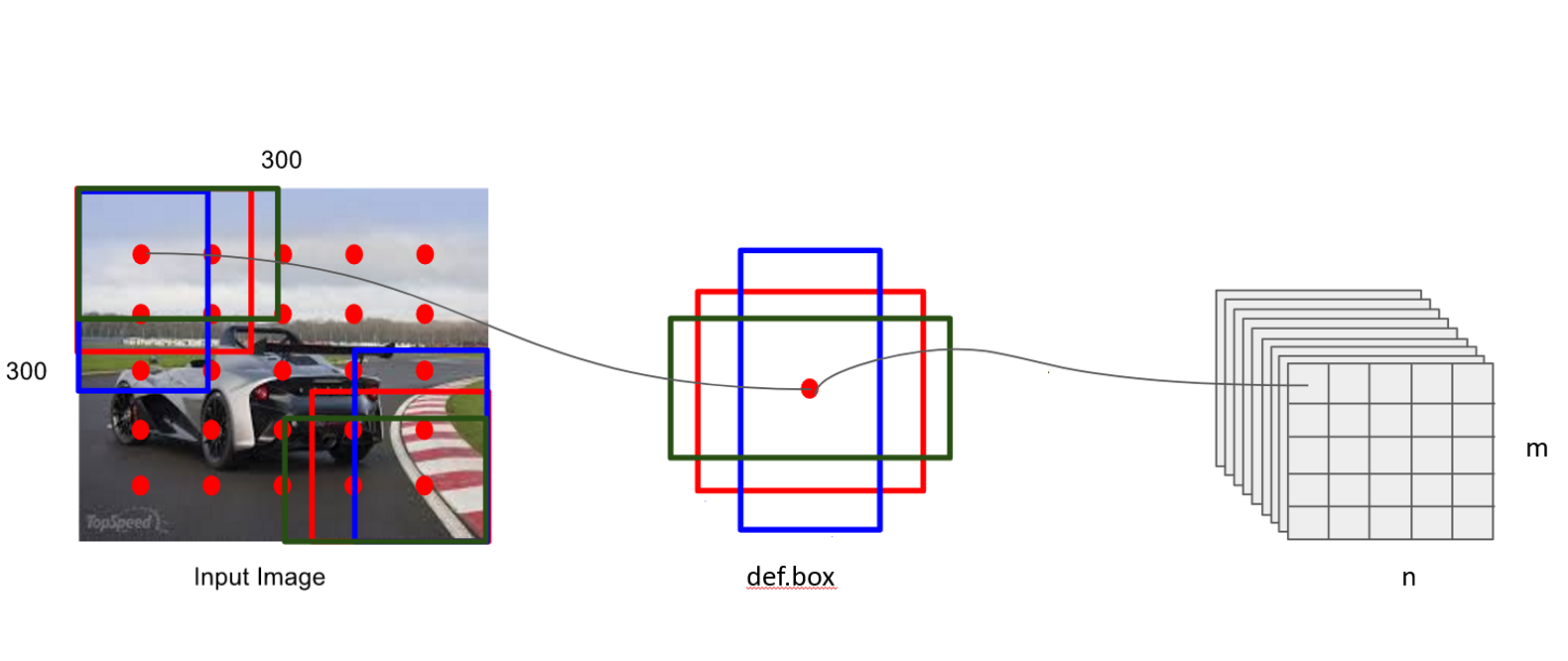


Hình 2.6. Cơ chế tính toán ở tầng phân lọai (<https://deepsystems.ai>)

* Vùng mặc định (default box - ) và tỉ lệ (aspect ratios - : Với bản đặt trưng, có kích thước , được chia thành ô, mỗi ô chứa vùng mặc định có kích thước chuẩn là (Hình 2.7). Ta được:

Mỗi vùng mặc định ứng với một bộ phân loại, cho ra kết quả dự đoán bao gồm hình dạng của vùng bao vật thể cùng xác suất mà vật thể trong vùng này có thể là một trong các loại khảo sát (gồm vật thể) có kích thước .

* Ví dụ, “classifier 4” có bộ phân loại và kích thước là (Hình 2.6).



Hình 2.7. Cách tạo ra các vùng mặc định (<https://deepsystems.ai>)

Sau khi tổng hợp được 8732 bộ phân loại cho mỗi loại, kết quả này được tính toán bằng thuật toán “none-maximun suppression” để đưa ra kết quả cuối cùng là tập các vùng chưa vật thể đã được gán nhãn.

### Kết quả thực nghiệm được công bố

<Xải>

## Nhận biết mặt người bằng DNN – VGG16

Các năm gần đây, sự phát triển của mạng neural network đã mang lại những bước tiến vượt bậc trong lĩnh vực thị giác máy tính, đặc biệt là các bài toán về detection, segmention, classification,… Bài toán nhận biết mặt người cũng đạt được độ chính xác cao, vượt qua khả năng nhận biết của người. Năm 2015, Omkar M. Parkhi, Andrea Vedaldi và Andrew Zisserman thực hiện công trình nghiên cứu [2], trong đó đề xuất mô hình nhận biết mặt người sử dụng Deep Convolutional Neural Network. Mô hình này đã trở thành baseline và kiến trúc tiêu biểu được tham chiếu bởi gần 400 (Google Shoclar) công trình nghiên cứu khác năm 2015.

### Cấu trúc VGG16

Bảng 2‑1. Cấu trúc chi tiết network VGG16 [2]

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Layer** | **Type** | **Name** | **Support** | **Filt dim** | **Num filts** | **Stride** | **Pad** |
| 0 | input | - | - | - | - | - | - |
| 1 | conv | conv1-1 | 3 | 3 | 64 | 1 | 1 |
| 2 | relu | relu1-1 | 1 | - | - | 1 | 0 |
| 3 | conv | conv1-2 | 3 | 64 | 64 | 1 | 1 |
| 4 | relu | relu1-2 | 1 | - | - | 1 | 0 |
| 5 | mpool | pool1 | 2 | - | - | 2 | 0 |
| 6 | conv | conv2-1 | 3 | 64 | 128 | 1 | 1 |
| 7 | relu | relu2-1 | 1 | - | - | 1 | 0 |
| 8 | conv | conv2-2 | 3 | 128 | 128 | 1 | 1 |
| 9 | relu | relu2-2 | 1 | - | - | 1 | 0 |
| 10 | mpool | pool2 | 2 | - | - | 2 | 0 |
| 11 | conv | conv3-1 | 3 | 128 | 256 | 1 | 1 |
| 12 | relu | relu3-1 | 1 | - | - | 1 | 0 |
| 13 | conv | conv3-2 | 3 | 256 | 256 | 1 | 1 |
| 14 | relu | relu3-2 | 1 | - | - | 1 | 0 |
| 15 | conv | conv3-3 | 3 | 256 | 256 | 1 | 1 |
| 16 | relu | relu3-3 | 1 | - | - | 1 | 0 |
| 17 | mpool | pool3 | 2 | - | - | 2 | 0 |
| 18 | conv | conv4-1 | 3 | 256 | 512 | 1 | 1 |
| 19 | relu | relu4-1 | 1 | - | - | 1 | 0 |
| 20 | conv | conv4-2 | 3 | 512 | 512 | 1 | 1 |
| 21 | relu | relu4-2 | 1 | - | - | 1 | 0 |
| 22 | conv | conv4-3 | 3 | 512 | 512 | 1 | 1 |
| 23 | relu | relu4-3 | 1 | - | - | 1 | 0 |
| 24 | mpool | pool4 | 2 | - | - | 2 | 0 |
| 25 | conv | conv5-1 | 3 | 512 | 512 | 1 | 1 |
| 26 | relu | relu5-1 | 1 | - | - | 1 | 0 |
| 27 | conv | conv5-2 | 3 | 512 | 512 | 1 | 1 |
| 28 | relu | relu5-2 | 1 | - | - | 1 | 0 |
| 29 | conv | conv5-3 | 3 | 512 | 512 | 1 | 1 |
| 30 | relu | relu5-3 | 1 | - | - | 1 | 0 |
| 31 | mpool | pool5 | 2 | - | - | 2 | 0 |
| 32 | conv | fc6 | 7 | 512 | 4096 | 1 | 0 |
| 33 | relu | relu6 | 1 | - | - | 1 | 0 |
| 34 | conv | fc7 | 1 | 4096 | 4096 | 1 | 0 |
| 35 | relu | relu7 | 1 | - | - | 1 | 0 |
| 36 | conv | fc8 | 1 | 4096 | 2622 | 1 | 0 |
| 37 | softmx | prob | 1 | - | - | 1 | 0 |

Chi tiết về cấu trúc của Deep Convolutional Neural Network – VGG16 được thể hiện trong Bảng 2‑1. Network bao gồm 11 khối, trong đó mỗi khối bao gồm một toán tử tuyến tính (linear operator) và theo sau đó là một hay nhiều toán tử phi tuyến (non-linearity) như ReLU hay max pooling. Tám khối đầu tiên là convolutional, sử dụng bank of filters. Ba khối cuối được gọi là Fully Connected, về bản chất thì các khối này tương tự như các lớp convolutional nhưng kích thước filter thì bằng với kích thước của dữ liệu đầu vào để mà từ đó các filter “quan sát” được toàn bộ ảnh.

Theo sau tất cả các lớp convolutional là lớp rectification (ReLU) tương tự như trong [20]. Tuy nhiên, không hoàn toàn giống như [20] mà tương tự [21], VGG16 không sử dụng Local Response Normalisation operator. Trong các lớp fully connected sau cùng thì hai lớp đầu tiên có 4096 chiều và lớp cuối có N = 2622 hoặc L = 1024 chiều tùy thuộc vào hàm lỗi (loss functions) được sử dụng để tối ưu hoặc mục đích dự đoán cho N classes (N-way class prediction). Vector kết quả sau đó được đưa vào lớp softmax để tính toán xác suất hậu nghiệm (posterior probabilities).

Ảnh đầu vào của network này có kích thước 224 x 224 và đã trừ đi ảnh khuôn mặt trung bình (tính toán từ tập dữ liệu).

### Kết quả thực nghiệm được công bố

#### Các tập dữ liệu và cách đánh giá

* Tập dữ liệu Labled Faces in the Wild (LFW) [8]

Tập dữ liệu Labled Faces in the Wild (được mô tả chi tiết ở 2.5.1) bao gồm 13,233 ảnh của 5,749 người và đây là một tập dữ liệu chuẩn để huấn luyện và đánh giá các thuật toán về xác nhận khuôn mặt (face verification). Nhóm tác giả [1] dùng hình thức đánh giá tiêu chuẩn “unrestricted setting” sử dụng thêm dữ liệu ngoài để huấn luyện và chọn Equal Error Rate (EER) để làm độ đo. Độ đo này được định nghĩa như là độ lỗi tại điểm trên đường cong ROC (Receiver operating characteristic) mà tại đó tỷ lệ true positive và false negative là bằng nhau.

* Tập dữ liệu YouTube Faces (YTF) [10]

Tập dữ liệu Youtube Faces (mô tả chi tiết ở 2.5.1) bao gồm 3,425 video của 1,595 người thu thập từ Youtube, trong đó mỗi người có trung bình 2 video. Đây được xem là một tập dữ liệu tiêu chuẩn cho xác nhận mặt người (face verification) trong video. Độ lỗi EER cũng được sử dụng để đánh giá trên Youtube Faces tương tự như 2.4.2.1a.

#### Kết quả

Bảng 2‑2. So sánh kết quả các mô hình bằng LFW unrestricted setting [2]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **Phương pháp** | **Số ảnh** | **Số network** | **Độ chính xác** |
| 1 | Fisher Vector Faces [22] | - | - | 93.10 |
| 2 | DeepFace [3] | 4M | 3 | 97.35 |
| 3 | Fusion [23] | 500M | 5 | 98.37 |
| 4 | DeepID-2,3 |  | 200 | 99.47 |
| 5 | FaceNet [9] | 200M | 1 | 98.87 |
| 6 | FaceNet [9] + Alignment | 200M | 1 | 99.63 |
| 7 | VGG16 [2] | 2.6M | 1 | 98.95 |

Bảng 2‑3. So sánh kết quả các mô hình bằng Youtube Face unrestricted setting [2]. K là số lượng người dung để nhận biết trong các video.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **Phương pháp** | **Số ảnh** | **Số network** | **100% - EER** | **Độ chính xác** |
| 1 | Video Fisher Vector Faces | - | - | 87.7 | 83.8 |
| 2 | DeepFace [3] | 4M | 1 | 91.4 | 91.4 |
| 3 | DeepID-2,2+,3 |  | 200 | - | 93.2 |
| 4 | FaceNet [9] + Alignment | 200M | 1 | - | 95.1 |
| 5 | VGG16 [2] (K=100) | 2.6M | 1 | 92.8 | 91.6 |
| 6 | VGG16 [2] (K=100) + Embedding Learning | 2.6M | 1 | 97.4 | 97.3 |

Bảng 2‑2 và Bảng 2‑3 so sánh độ chính xác của VGG16 với các mô hình đạt kết quả cao nhất trên tập dữ liệu [8] và trên [10]. Điểm nổi bật network trong [2] là đạt được độ chính xác xấp xỉ các mô hình hàng đầu (mặc dù chưa vượt qua được) nhưng số lượng dữ liệu cần sử dụng ít hơn một cách đáng kể và cấu trúc network cũng đơn giản hơn rất nhiều (chỉ sử dụng 1 network).

## Các tập dữ liệu liên quan

Trong hai thập kỷ qua, cùng với sự phát triển của lĩnh vực Machine Learning nói chung và Face Recognition nói riêng, có rất nhiều tập dữ liệu mới được tạo ra nhằm mục đích phục vụ cho khoa học và thương mại. Các tập dữ liệu ban đầu có kích thước nhỏ, với các định dạng đơn giản như grayscale hay RGB có độ phân giải chưa cao (còn bị ảnh hưởng nhiều bởi các yếu tố về nhiễu, ánh sang, tương phản,…). Trong các năm gần đây, với sự phát triển cao của các thiết bị thu nhận hình ảnh cùng với nhu cầu từ các đề tài nghiên cứu cũng như ứng dụng từ giới công nghiệp, rất nhiều tập dữ liệu “không lồ” được tạo ra. Trong đó, số lượng ảnh thu thập tăng lên đáng kể từ hàng trăm ngàn đến hàng triệu. Chất lượng ảnh cũng được cải thiện rõ rệt nhờ cấu tạo tiên tiến của hệ thống camera, có nhiều định dạng ảnh mới ra đời như RGB-D, các mô hình ba chiều,…. Phần tiếp theo tập trung giới thiệu các tập dữ liệu điển hình cho bài toán nhận biết mặt người và giới thiệu tập dữ liệu được chọn để thực hiện đề tài. Bảng 2‑4 liệt kê một số tập dữ liệu tiêu biểu dùng cho nhận biết mặt người.

Bảng 2‑4. Một số tập dữ liệu dùng cho nhận biết mặt người

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên** | **Số người** | **Số mẫu** | **Năm** |
| 1 | FaceScrub [24] | 530 | 106,863 | 2014 |
| 2 | SCFace [25] | 130 | 4160 | 2011 |
| 3 | YouTube Faces [10] | 1,595 | 3,425 | 2011 |
| 4 | CASIA-WebFace [26] | 10,575 | 494,414 | 2011 |
| 5 | The MUCT Landmarked [27] | 276 | 3755 | 2010 |
| 6 | Bosphorus [28] | 105 | 4666 | 2008 |
| 7 | CMU Multi-PIE Face [17] | 337 | 750,000 | 2008 |
| 8 | Labeled Faces in the Wild [8] | 1680 | 13,000 | 2007 |

### Khảo sát các tập dữ liệu

#### FaceScrub

FaceScrub được tạo ra với mục đích cung cấp một tập dữ liệu lớn phục vụ cho nghiên cứu nhận biết mặt người. Nhóm tác giả [24] phát triển một hệ thống phát hiện các khuôn mặt trong ảnh trả về từ việc tìm kiếm ảnh các nhân vật nổi tiếng trên internet, trong đó hệ thống tự động lọc bỏ các ảnh không thuộc về người đang được tìm kiếm.

Tập dữ liệu bao gồm 106,863 ảnh màu của 530 diễn viên nổi tiếng, trong đó mỗi người có trung bình 200 mẫu. Mỗi ảnh được chụp trong môi trường tự nhiên và không có điều kiện ràng buộc nào. Thông tin về tên và giới tính của 265 nam và 265 nữ diễn viên được cung cấp đầy đủ.

Hình 2.8 thể hiện ví dụ về một số khuôn mặt trong tập FaceScrub với sự đa dạng về điều kiện chụp, góc nhìn, độ sáng và rất nhiều thông số camera.



Hình 2.8. Một số ảnh mẫu trong tập FaceScrub

*(http://vintage.winklerbros.net/facescrub.html*)

#### SCFace

SCFace là tập dữ liệu chứa các ảnh tinh được thu thập trong môi trường tự do từ năm camera giám sát. Tập dữ liệu gồm 4160 ảnh của 130 người. Ảnh từ các camera với điều kiện khác nhau mô phỏng theo điều kiện thế giới thực cho rất có ích cho các nghiên cứu, kiểm tra hiệu năng của các thuật toán nhận biết mặt người.

Các ảnh ví dụ về ảnh được chụp từ các camera khác nhau được thể hiện trong Hình 2.9.



Hình 2.9. Tập ảnh ví dụ cho một người trong SCFace

*(http://www.scface.org/)*

#### Youtube Faces

Youtube faces dataset được thiết kế nhằm phục vụ cho cho việc nghiên cứu các vấn đề nhận biết mặt người trong video tự do. Tập dữ liệu chứa 3,425 của 1,595 người và được download từ Youtube. Trung bình mỗi người có 2.15 video. Độ dài các video trung bình là 181.3 frame, video ngắn nhất có 48 frame và dài nhất là 6,070 ms.

Nhóm tác giả [10] cung cấp các bộ test chuẩn để đánh giá hiệu suất của các kỹ thuật video pair-matching. Bảng mô tả (descriptor) cho sự xuất hiện của các khuôn mặt được cung cấp sử dụng các bảng mô tả chuẩn.

#### CASIA-WebFace

Tập dữ liệu đóng một vai trò quan trọng trong các nghiên cứu về nhận biết mặt người, vì thế nhóm tác giả [26] đề xuất một phương pháp bán tự động để thu thập ảnh từ internet và thành lập một tập dữ liệu mới. CASIA bao gồm 494,414 ảnh tĩnh của 10,575 người và trở thành tập dữ liệu cao thứ hai, chỉ nhỏ hơn tập không được công khai của Facebook lúc công bố.

Bảng 2‑5. Bảng so sánh kích thước tập CASIA-WebFace và một số tập khác [26]

|  |  |  |  |
| --- | --- | --- | --- |
| **Tập dữ liệu** | **Số người** | **Số ảnh** | **Công khai** |
| LFW [8] | 5,749 | 13,233 | + |
| WDRef [29] | 2,995 | 99,773 | + (chỉ đặc trưng) |
| CelebFaces [30] | 10,177 | 202,599 | + |
| SFC [3] | 4,030 | 4,400,000 | - |
| CACD [31] | 2,000 | 163,446 | + (một phần chú thích) |
| CASIA-WebFace [26] | 10,575 | 494,414 | + |

Bảng 2‑5 so sánh kích thước tập CASIA-WebFace (số lượng người và số mẫu) với một số tập dữ liệu lớn.

#### The MUCT Landmarked

Tập dữ liệu MUCT được tạo ra để cung cấp các mẫu mặt người đa dạng về độ sáng, tuổi và dân tộc với 3755 khuôn mặt của 76 người. Một vài ví dụ mẫu được thể hiện trong Hình 2.10.



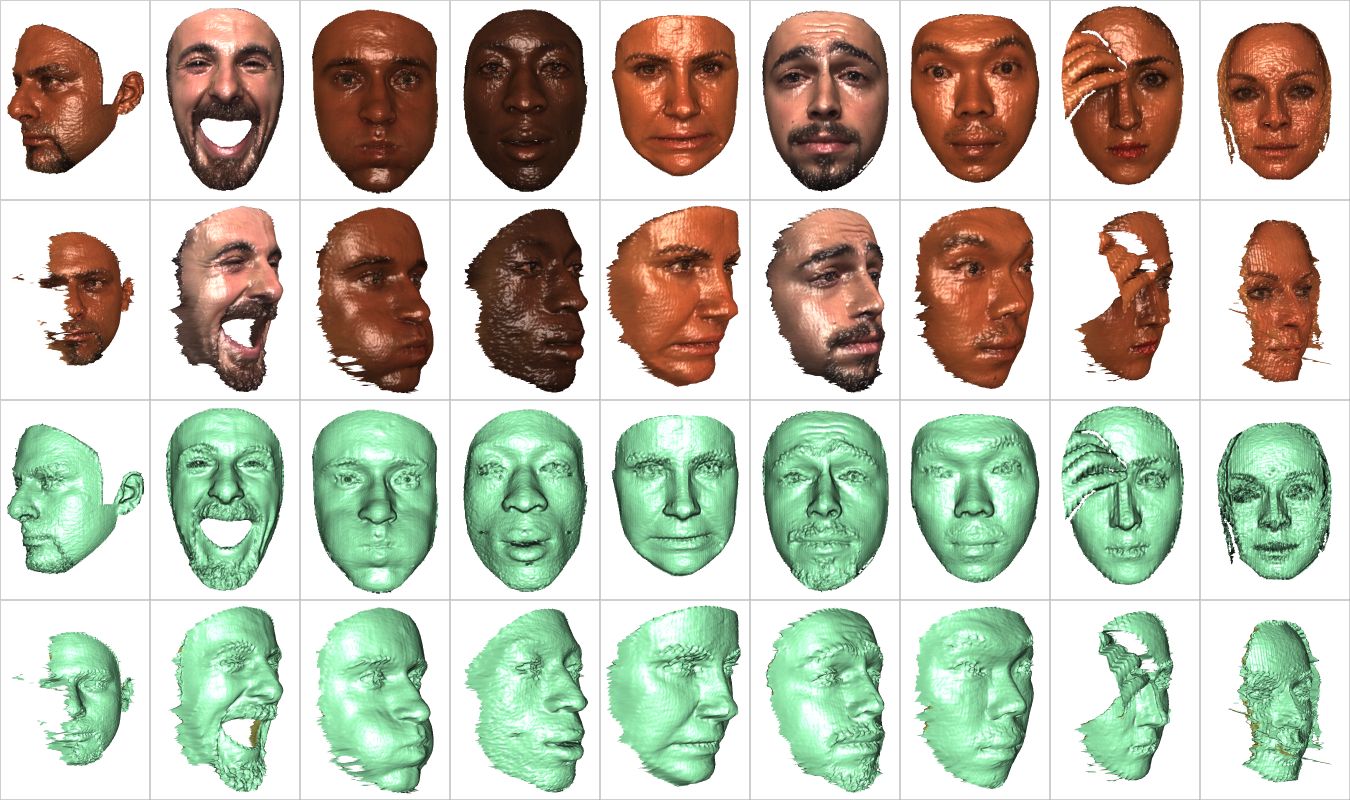
Hình 2.10. Vài mẫu trong tập dữ liệu MUCT Landmarked

*(http://www.milbo.org/muct/)*

#### Bosphorus

Bosphorus là tập dữ liệu phụ vụ cho nghiên cứu về các bài toán xử lý mặt người 2D và 3D, bao gồm: nhận biết cảm xúc, phát hiện cử chỉ trên mặt (facial action unit detection), ước lượng cường độ đơn vị cử chỉ trên mặt (facial action unit intensity estimation), nhận biết mặt người trong điều kiện bất lợi (face recognition under adverse conditions), mô hình hóa khuôn mặt biến dạng được, tái cấu trúc khuôn mặt ba chiều.

Có tất cả 4666 khuôn mặt của 105 người trong tập dữ liệu với ba đặc điểm chính: đa dạng về biểu cảm (có tới 35 trạng thái cho mỗi người, FACS scoring – bao gồm cường độ và mã bất đối xứng cho mỗi AU, một phần ba tập dữ liệu là các diễn viên chuyên nghiệp), tư thế khuôn mặt có hệ thống (bao gồm 13 kiểu nghiêng và xoay), có rất nhiều loại che khuất (râu, tóc, tay, mắt kính) (Hình 2.11).



Hình 2.11. Các mẫu trong tập Bosphorus

*(http://bosphorus.ee.boun.edu.tr/default.aspx)*

#### CMU Multi-PIE Face

Năm 2000, tập dữ liệu PIE database [32] được thu thập để phục vụ nghiên cứu về nhận biết khuôn mặt, trong đó tập trung vào hai yếu tố cản trở việc nhận biết là tư thế và điều kiện chiếu sáng. Tuy đóng vai trò hiệu quả cho công đề tài nhưng tập PIE vẫn còn hạn chế ở một số mặt như sau: số lượng cá thể ít, thu thập tại cùng một lần (a single recording session), không đa dạng về biểu cảm. Vì thế, tập Multi-PIE được ra đời, phát triển từ tập dữ liệu cũ để giải quyết các vấn đề hiện hữu.

Tập CMU Multi-PIE Face chứa hơn 750,000 ảnh thu thập từ 337 người với 15 góc nhìn, 19 điều kiện chiếu sáng trong bốn phiên khác nhau. Điều này làm nên sự đa dạng lớn cho tập dữ liệu.

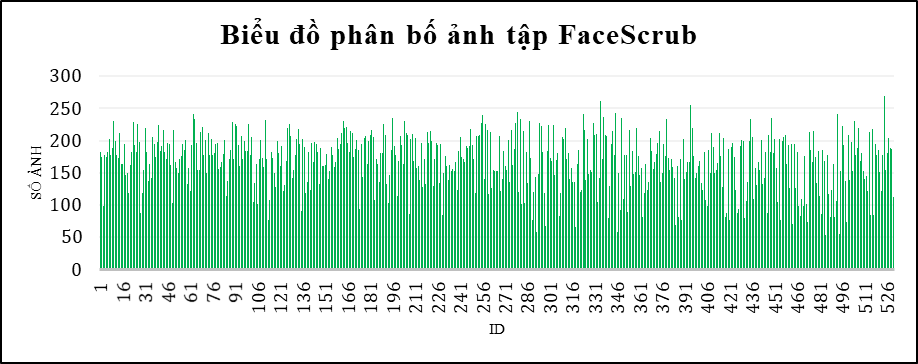
#### Labeled Faces in the Wild

Labeled Faces in the Wild là tập dữ liệu được thiết kế cho bài toán nhận biết mặt người trong điều kiện tự do, bao gồm 13,233 ảnh thu tập từ internet của 5,749 người. Trong đó 1,680 người có nhiều hơn một ảnh. Mỗi ảnh được đặt tên theo người trong hình và điểm chung của các khuôn mặt trong tập dữ liệu này đó là đều được phát hiện bởi Viola-Jones face detector.

Đến nay, có tất cả bốn tập LFW bao gồm một tập gốc và ba tập được căn chỉnh. Ba tập dữ liệu mới là “funneled images” (ICCV 2007), LFW-a (được căn chỉnh bằng một thuật toán chưa công bố) và “deep funneled images” (NIPS 2012). Trong số đó, LFW-a và “deep funneled images” được dùng hiệu quả hơn cho các thuật toán xác nhận mặt người (face verification) hơn các tập còn lại (ICCV 2007).

### Phân tích tập dữ liệu FaceScrub

Tập dữ liệu FaceScrub nguyên bản gồm 106,863 ảnh màu của 530 diễn viên nổi tiếng và được tác giả cung cấp theo dạng các URL để tải ảnh về từ internet. Nhóm thực hiện đề tài đã xây dựng crawler tự động để thu thập. Do một số URL không hợp lệ hoặc dữ liệu không còn tồn tại tại thời điểm tải nên nhóm thực hiện đề tài chỉ thu được tổng cộng 89295 mẫu, với phân bố được thống kê trong Hình 2.12.



Hình 2.12. Biểu đồ phân bố dữ liệu trong tập FaceScrub

Trong đó, người có số lượng ảnh cao nhất là 269, thấp nhất là 53 và trung bình là 168 ảnh/người.

Đây là tập dữ liệu được tìm kiếm từ internet theo tên các diễn viên trong đó, do vậy tồn tại một số vấn đề về chất lượng dữ liệu và điều đó gây ảnh hưởng khá lớn đến độ chính xác của mô hình nhận biết mặt người.

1. Số lượng ảnh của mỗi người

Số lượng ảnh giữa các lớp có sự khác biệt khá lớn (± 42 ảnh). Điều này làm cho những người có số lượng ảnh lớn dễ lấn át kết quả phân lớp những người có ít ảnh hơn trong tập.

1. Điều kiện chiếu sáng

Điều kiện chiếu sáng tự do và đang dạng trong các ảnh mẫu làm cho rất nhiều khuôn mặt bị che khuất và do đó mất đi rất nhiều đặc điểm về màu sắc, hình dáng, đường nét,… giúp ích cho việc phát hiện và nhận biết khuôn mặt (Hình 2.13).



Hình 2.13. Ví dụ về điều kiện chiếu sáng tự do trong FaceScrub

1. Sự lão hóa khuôn mặt

Các ảnh được thu thập từ internet và không có ràng buộc nào đảm bảo các ảnh của một người sẽ cùng thuộc một giai đoạn tuổi tác của họ. Chính vì thế, sự lão hóa và biến dạng khuôn mặt làm thay đổi đặc điểm nhận dạng rất nhiều. Điều này là mốt vấn đề lớn và đang được rất nhiều đề tài nghiên cứu quan tâm trong lĩnh vực nhận biết mặt người (Hình 2.14).



Hình 2.14. Ví dụ về lão hóa trong tập FaceScrub

1. Thông số camera

Thông số camera là một điều đặc biệt quan trọng trong việc quyết định chất lượng hình ảnh. Các thông số tiêu biểu có thể kể đến là góc chụp, hệ màu, độ phân giải,… Sự khác biệt giữa các ảnh do sử dụng các camera quá khác nhau về chất lương và thông số gây ra những biến thể khó cho việc định danh nhân vật (Hình 2.15).



Hình 2.15. Ví dụ về sự khác biệt khi dùng các camera quá khác nhau.

1. Trang điểm và hóa trang

Do đặc thù tập dữ liệu FaceScrub bao gồm các diễn viên và nghệ sĩ nên hóa trang, trang điểm là việc hết sức đa dạng. Tùy thuộc vào phim, vào nhân vật trong truyện hay môi trường mà các diễn viên này có sự thay đổi khá lớn, đơn cử là các trường hợp biến hóa thành quái vật, siêu nhân,… Lúc này ảnh khuôn mặt thay đổi lớn thách thức cả bài toán phát hiện khuôn mặt (Hình 2.16).



Hình 2.16. Ví dụ về hóa trang và trang điểm trong tập FaceScrub.

## Kết luận

Chương này, chúng em trình bày một số khả sát về các công trình tiên tiến trong lĩnh vực phát hiện và nhận biết mặt người. Trong đó tập trung vào hai mô hình tiêu biểu là [19]cho phát hiện và [2] cho nhận biết mặt người. Đồng thời giới thiệu các tập dữ liệu liên quan, phân tích tập dữ liệu FaceScrub [24] – tập dữ liệu tiêu biểu được chọn để thực hiện các thí nghiệm về sau. Đây là cơ sở cho việc tinh chỉnh cũng như huấn luyện các mô hình ở Chương 2.

# Huấn luyện mô hình phát hiện và nhận biết mặt người

* Nội dung Chương 3 trình bày các hướng tiếp cận và giải pháp mà nhóm thực hiện đề tài triển khai trên mô hình phát hiện và nhận biết mặt người được lựa chọn (SSD300 và VGG16). Các tinh chỉnh đối với mô hình cũ, đề xuất cấu trúc mới được trình bày cụ thể theo sau sự phân tích các mô hình này. Bên cạnh đó, quá trình và chiến lược huấn luyện cũng như việc xây dựng mới một số tập dữ liệu phù hợp được mô tả chi tiết để chứng minh tính hiệu quả của hướng tiếp cận và giải pháp mà nhóm đề ra.

## Mô hình phát hiện mặt người bằng SSD300

SSD không sử dụng chọn lại các đặc trung giống Fast R-CNN nên không thể phát hiện được các vật thể có kích thước nhỏ mà mặt người xuất hiện trong ảnh thực thì thường chiếm diện tích nhỏ. Ngoài ra, dữ liệu trong tập huấn luyện thì diện tích mặt người lớn và kích thước của vật thể cũng là một giá trị quyết định việc có tìm ra được của vật thể đó trong bức ảnh hay không. Vì thế khó có thể huấn luyện với gương mặt lớn và đòi hỏi phải tìm được gương mặt nhỏ hơn nhiều lần, vì dữ liệu huấn luyện có sẵn chỉ chứa một người trên một ảnh và dữ liệu thực thế thì không như thế. Để giải quyết vấn đề trên, nhóm em đề xuất một số thay đổi để áp dụng mô hình SSD vào bài toán phát hiện mặt người.

### Tinh chỉnh SSD300

Thay vì tìm khuôn mặt trên cả ảnh, nhóm chỉ quan tâm đến những vùng « khả thi », vùng chứa người. Cụ thể, nhóm dùng mô hình SSD300 [[1]](#footnote-1) với bộ tham số được huấn luyện trên tập COCO + VOC2007+VOC2012, phương pháp cân bằng giữa độ chính xác và tốc độ [19]. Sau khi được các vùng « khả thi », nhóm lại áp dụng SDD300, được huấn luyện lại với tập dữ liệu khuôn mặt (Xem 3.1.2), tạo ra danh sách các ứng cử có thể là mặt người. Danh sách này được tính toán dựa trên tỉ lệ giao, trùng lấp lên nhau và trọng số để quyết định kết quả cuối cùng. Tuy nhiên nếu ở bước tìm ra các vùng chứa người thực hiện ra kết quả không tốt sẽ ảnh hưởng đến bước tìm khuôn mặt. Các trường hợp ảnh có quá nhiều người sẽ làm cho thuật toán chạy chậm; hoặc kích thước người nhỏ làm cho SSD khó có thể phát hiện ra.

### Xây dựng tập dữ liệu

<Xải>

### Huấn luyện và kết quả

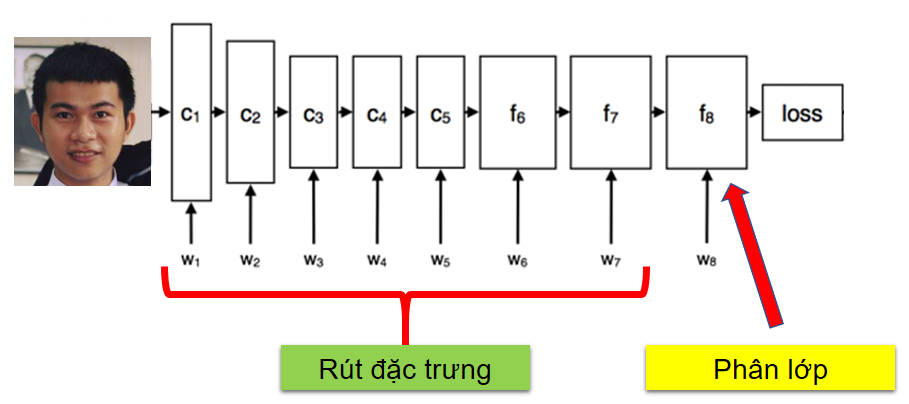
<Xải>

## Mô hình nhận biết mặt người bằng VGG-16 Deep features

### Áp dụng kỹ thuật transfer learning

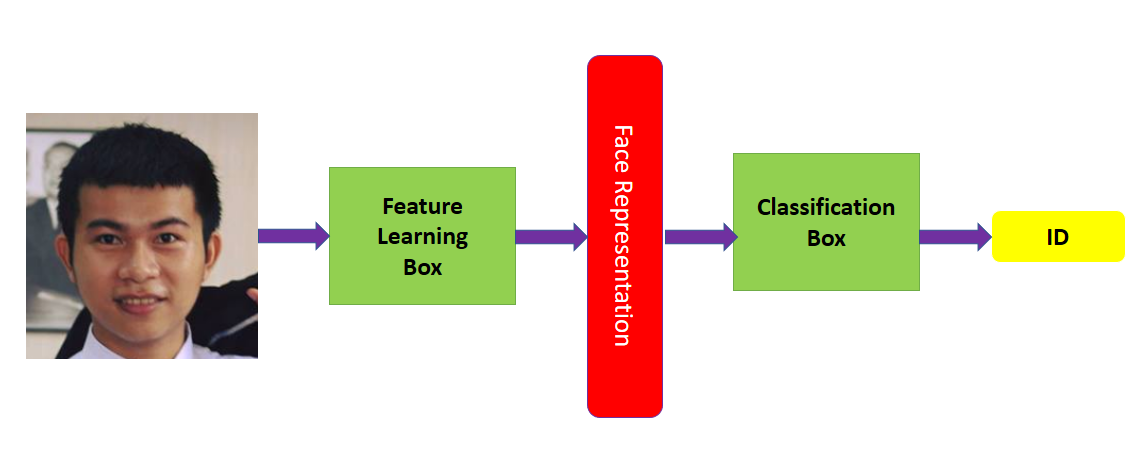
[2] đã đạt kết quả vượt bậc trên hai tập dữ liệu là LFW [8] và YTF [10] và điều đó chứng minh kiến trúc network này rất phù hợp cho bài toán nhận biết mặt người. Thế nhưng để sử dụng lại cấu trúc này trên một tập dữ liệu hoàn toàn mới thì đòi hỏi một quá trình huấn luyện lâu và tốn nhiều chi phí. Chính vì thế nhóm thực hiện đề tài sử dụng kỹ thuật transfer learning để huấn luyện và đáp ứng yêu cầu trên tập dữ liệu FaceScrub [24].

Nếu phân tích cấu trúc network của VGG16 [2] thì có thể chia các lớp thành hai nhóm chủ yếu là các lớp học đặc trưng và các lớp định danh/ phân loại. Nhóm học đặc trưng (feature learning) bao gồm các lớp convolution và hai lớp Fully Connected đầu tiên. Tầng Fully Connected cuối cùng đảm nhiệm vai trò định danh cho các đặc trưng ảnh được rút ra. Hình 3.1 trình bày vai trò các các lớp trong cấu trúc VGG16 network: các khối C***x*** tượng trưng cho các lớp convolution theo sau là ReLU nằm giữa các tầng max-pooling, các khối f***x*** tương trưng cho các tầng Fully Connected.



Hình 3.1. Vai trò của các lớp trong VGG16 network

Nhìn một các tổng quát thì bài toán nhận biết khuôn mặt theo hướng tiếp cận neural network có thể được chia làm các giai đoạn như sau: từ ảnh đầu vào 🡪 hộp đen học đặc trưng 🡪 các biểu diễn ảnh theo một chiều không gian khác 🡪 hộp đen phân lớp 🡪 định danh (Hình 3.2). VGG16 Net [2] đã được huấn luyện để có bộ trọng số tốt cho việc biến đổi ảnh đầu vào thành một cách biễu diễn khác chặt chẽ và cô đọng hơn rất nhiều trong đó làm nổi bật các đặc trưng của ảnh, chính vì thế nhóm thực hiện đề tài đã đùng khối feature learning này rút ra các deep feature trên tập dữ liệu FaceScrub [24] rồi tiến hành phân lớp định danh lại bằng một cấu trúc network khác được trình bày trong phần 3.2.2. Deep feature được rút ra sau tầng Fully Connected thứ hai (f7).



Hình 3.2. Qui trình nhận biết mặt người tổng quát.

### Network định danh VGG16-Deep-Feature

Sau khi rút các deep feature theo cách được trình bày ở phần 3.2.1, nhóm thực hiện đề tài thiết kế một Deep Neural Network mới để phân lớp cho các đặc trưng này. Network mới có cấu trúc đơn giản chỉ bao gồm các lớp Fully Connected phù hợp với việc định danh cho khuôn mặt – tương ứng với vector đặc trưng đầu vào. Cấu trúc network được mô tả chi tiết trong Bảng 3‑1. Kích thước deep feature đưa vào là 25,088 – được giữ nguyên so với VGG16, kích thước đầu ra là 530 tương ứng với số lượng diễn viên trong tập FaceScrub [24].

Bảng 3‑1. Cấu trúc network đề xuất để phân lớp VGG16-Deep-Feature

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Layer** | **Type** | **Name** | **#. nodes** | **Input Size** |
| 0 | input | - | - | - |
| 1 | Fully Connected | fc1 | 4096 | 25088 |
| 2 | Fully Connected | fc2 | 2048 | 4096 |
| 3 | Fully Connected | Prob | 530 | 2048 |

### Huấn luyện và kết quả

## Kết luận

Nội dung Chương 3 trình bày giải pháp mà nhóm thực hiện đề tài đã làm để tinh chỉnh thuật toán SSD300 [19] cho phù hợp với yêu cầu phát hiện mặt người (mục 3.1.1), xây dựng tập dữ liệu mới dựa trên sự kế thừa các tập dữ liệu đã có (mục 3.1.2), đề xuất hướng sử dụng deep feature từ VGG16 network (mục 3.2.1) cũng như thiết kế một cấu trúc mới và huấn luyện cho việc định danh các đặc trưng này (mục 3.2.2). Bên cạnh đó, quá trình huấn luyện và các kết quả đạt cũng được trình bày chi tiết và thống kê cụ thể sau mỗi phần tương ứng nhằm chứng minh tính hiệu quả của các hướng tiếp cận và giải pháp mà nhóm thực hiện đề tài đưa ra.

# Các phân hệ trong hệ thống tương tác thông minh

* Nội dung Chương 4 trình bày bốn phân hệ trong hệ thống tương tác thông minh dựa trên tổng hợp thông tin bằng phát hiện và nhận biết mặt người, bao gồm: các Face Web APIs, Person-based news highlight, Character-based movie synopsis và Character-based filter. Trong mỗi phần tương ứng, nhóm thực hiện đề tài trình bày chi tiết kiến trúc hệ thống, ngữ cảnh sử dụng và các chức năng được cung cấp cũng như hướng dẫn sử dụng và demo tương ứng.

## Face Web APIs

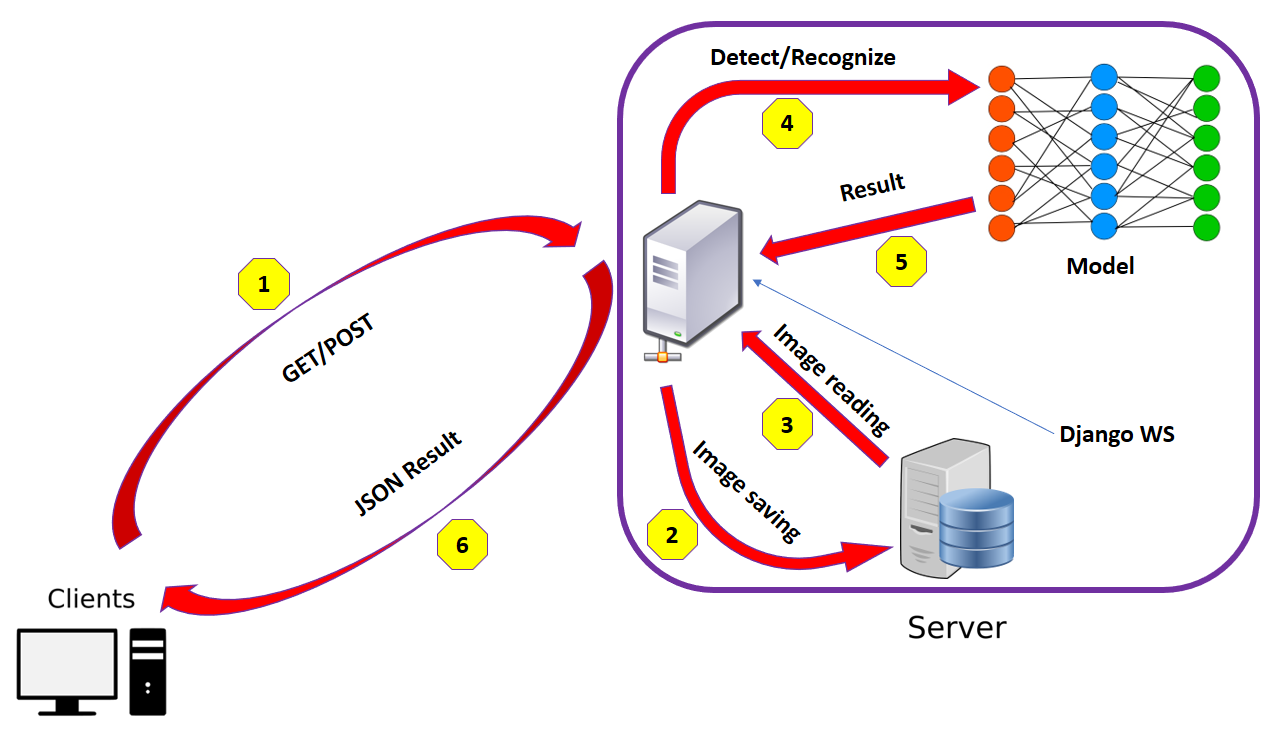
Sau khi huấn luyện hoàn chỉnh các mô hình phát hiện và nhận biết mặt người (trình bày ở Chương 3), nhóm thực hiện đề tài phát triển thành Web APIs với mục đích phục vụ cho các ứng dụng về sau và demo trực quan cho khả năng và tính hiệu quả của các network này. Các APIs được chia thành hai nhóm là phát hiện và nhận diện và được trình bày trong Bảng 4‑1. Với nhóm phát hiện khuôn mặt, nhóm phát triển hai APIs sử dụng hai phương pháp khác nhau là SSD300 [19] và OpenCV sử dụng đặc trưng Frontal Haar Cascade. Nhóm nhận diện sử dụng phương pháp duy nhất là VGG16+NN (VGG16-Deep-Feature và phân lớp bằng neural network tự đề xuất). Tuy nhiên, các APIs nhận diện bao gồm hai bước là phát hiện và định danh, do đó, nhóm phát triển hai APIs nhận diện với giai đoạn phát hiện sử dụng hai giải pháp khác nhau đã đề cập.

Bảng 4‑1. Bảng phân loại Face Web APIs đã phát triển.

|  |  |  |
| --- | --- | --- |
| **Số thứ tự** | **API** | **Công dụng** |
| 1 | SSD300 [19] **(\*)** | Phát hiện |
| 2 | OpenCV + Frontal Haar Cascade **(\*\*)** | Phát hiện |
| 3 | VGG16+NN phát hiện bằng **(\*)** | Nhận diện |
| 4 | VGG16+NN phát hiện bằng **(\*\*)** | Nhận diện |

### Kiến trúc hệ thống

Các chứng năng phát hiện và nhận biết mặt người được nhóm sinh viên thực hiện đề tài phát triển dưới dạng Web APIs hoạt động theo mô hình client-server như sau :



Hình 4.1. Mô hình hoạt động của Face Web APIs

Nghi thức hoạt động của hệ thống Face Web APIs được trình bày chi tiết trong Bảng 4‑2.

Bảng 4‑2. Nghi thức hoạt động của Face Web APIs.

|  |  |  |
| --- | --- | --- |
| **Giai đoạn** | **Client Side** | **Server Side** |
| 1 | Client request lên server bằng giao thức GET/POST trong đó kèm theo ảnh cần xử lý (dạng URL/stream). | *Waiting…* |
| 2 | *Waiting…* | Server nhận request download ảnh từ URL/stream. |
| 3 | *Waiting…* | Server đọc ảnh lên và chuẩn bị các dữ liệu khác tương ứng. |
| 4 | *Waiting…* | Server gọi ứng dụng bên thứ ba đảm trách vô trò vận hành model và thực hiện detect/recognize. |
| 5 | *Waiting…* | Server nhận kết quả trả về, chuẩn hóa và định dạng kết quả. |
| 6 | *Waiting…* | Server response client với kết quả xử lý dưới dạng JSON. |

Triển khai hệ thống theo kiến trúc như trên mang lại một số ưu điểm và khuyết điểm đi kèm so với khi triển khai trên nội bộ máy tính (Bảng 4‑3).

Bảng 4‑3. Ưu và khuyết điểm của kiến trúc hệ thống.

|  |  |  |
| --- | --- | --- |
| **STT** | **Khuyết điểm** | **Ưu điểm** |
| 1 | Đòi hỏi kết nối internet để truy cập server. | Không giới hạn nền tảng phát triển ứng dụng phía client. |
| 2 | Thời gian xử lý chậm hơn do phải truyền dữ liệu lên xuống. | Server có khả nặng xử lý mạnh hơn máy cá nhân, đặc biệt hiệu quả trong vận hành các model lớn. |
| 3 | Khó triển khai cho các hệ thống realtime. | Dễ dàng nâng cấp và cải tiến hệ thống. |

### Đặc tả APIs

Các APIs được phát triển trên framework Django [33] sử dụng ngôn ngữ Python 2 và được sử dụng để phát hiện hay nhận biết các khuôn mặt trong một ảnh tĩnh. Trong đó, nhóm hỗ trợ hai phương thức **GET** và **POST** với định dạng request URL sau:

* Phương thức GET :

http://**<IP server>**:8000/**<loại API>**?url=**<Image URL>**

* Phương thức POST :

http://**<IP server>:**8000/recognise

Trong đó:

- **<IP server>** : địa chỉ server cung cấp API tương ứng (Bảng 4‑4).

- **<Image URL>** : URL đến ảnh mong muốn cần xử lý.

- **<loại API>** : “**detect**” hoặc “**recognise**”.

- Phương thức POST gửi ảnh lên với tên: “**image**”.

Bảng 4‑4. Địa chỉ IP của các server cung cấp Face Web APIs.

|  |  |  |
| --- | --- | --- |
| **Số thứ tự** | **API** | **IP** |
| 1 | SSD300 [19] **(\*)** | **128.199.90.168** |
| 2 | OpenCV + Frontal Haar Cascade **(\*\*)** | **139.59.252.244** |
| 3 | VGG16+NN phát hiện bằng **(\*)** | **128.199.205.131** |
| 4 | VGG16+NN phát hiện bằng **(\*\*)** | **139.59.252.244** |

Kết quả được trả về theo định dạng JSON và mô tả chi tiết trong Bảng 4‑5.

Bảng 4‑5. Kết quả trả về của các APIs dưới dạng JSON.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Result**  **Code** | **Nội dung** | **JSON** | **Loại API** |
| 1 | **-1** | Request bằng phương thức không phải GET/POST | {"code":-1} | - |
| 2 | **-2** | Image URL lỗi | {"code":-2} | - |
| 3 | **-3** | Lỗi hệ thống | {"code":-3} | - |
| 4 | **0** | **<số lượng>**: số lượng khuôn mặt được phát hiện.  **<x, y, width, height>**: tọa độ theo định dạng của bounding box (top-left).  <**URL**>: đường dẫn đến ảnh kết quả được visualized. **(\*)** | {  "code":0,  "num":**<số lượng>**,  "coordinates":[  "**<x,y,wh>**",  **...**  ],  "url":<**URL**>  } | GET |
| 5 | **0** | **<số lượng>**: số lượng khuôn mặt được phát hiện.  **<x, y, width, height>**: tọa độ theo định dạng của bounding box (top-left).  <**tên**>: tên nhân vật. **(\*\*)**  <**URL**>: đường dẫn đến ảnh kết quả được visualized. **(\*)** | {  "code":0,  "num":**<số lượng>**,  "names":[  "**<tên>**",  **...**  ]  "coordinates":[  "**<x,y,wh>**",  **...**  ],  "url":<**URL**>  } | POST |

**(\*)**: truy cập theo đường dẫn với định dạng: <**IP server**>:8000<**URL**>.

**(\*\*)**: thứ tự tên trong “**names**” tương ứng với thứ tự tọa độ trong “**coordinates**”.

Bảng 4‑6 trình bày ví dụ về kết quả xử lý ảnh của diễn viên Daniel Radcliffe vai Harry Potter trong series phim cùng tên (Hình 4.2), trong đó sử dụng thuật toán SSD300 [19] cho phát hiện và VGG16+NN-SSD300 [2] [19] cho nhận diện.



Hình 4.2. Diễn viên Daniel Radcliffe vai Harry Potter trong series phim cùng tên.

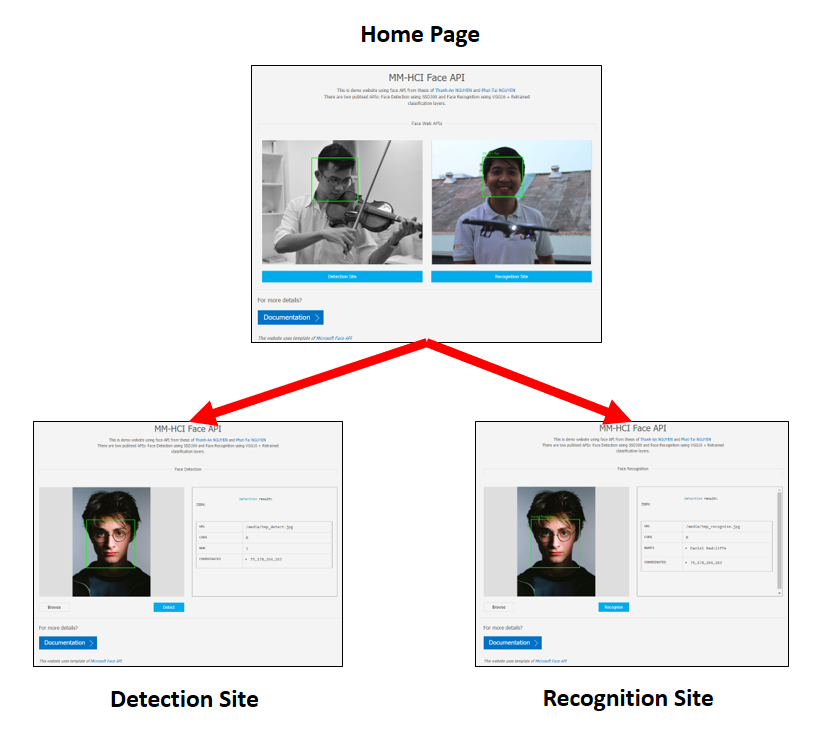
Bảng 4‑6. Ví dụ kết quả phát hiện và nhận biết mặt diễn viên Daniel Radcliffe.

|  |  |  |
| --- | --- | --- |
| **API** | **Kết quả** | **Ảnh kết quả** |
| Phát hiện | {  "url":"\media\tmp\_detect.jpg",  "code": 0,  "num": 1,  "coordinates": [  "65,53,76,76"  ]  } | http://128.199.90.168:8000/media/tmp_detect.jpg |
| Nhận diện | {  "url":"\media\tmp\_recognise.jpg",  "code": 0,  "names": [  "Daniel Radcliffe"  ],  "coordinates": [  "65,53,76,76"  ]  } | http://128.199.205.131:8000/media/tmp_recognise.jpg |

### Demo Website

Trang chủ của website có địa chỉ <http://128.199.70.20:8000/> bao gồm hai trang con tương ứng với hai loại API chính của đề tài là phát hiện và nhận biết khuôn mặt. Trang chủ có nhiệm vụ giới thiệu và cung cấp thông tin. Hai trang con cung cấp giao diện cho phép người dùng upload ảnh và thực hiện detect/recognise các khuôn mặt trong ảnh đó. Kết quả trả về được thể hiện bằng hai cách: xuất thông tin dưới dạng bảng và vẽ trên ảnh để thể hiện một các trực quan nhất. Hình 4.3 minh họa cấu trúc cũng như giao diện các trang trong ứng dụng Demo Website.

Thuật toán được sử dụng trong Detection Site là SSD300 [19] và trong Recognistion Site là VGG16+NN\_SSD300 [2], [19]. Điểm đặc biệt trong ứng dụng này là webserver năm hoàn toàn độc lập với các server cung cấp Face APIs. Mọi thao tác xử lý tính toán đều thông qua sự liên kết và trao đổi dữ liệu giữa các server. Điều này cho phép nâng cấp và bảo trì hệ thống một cách dễ dàng.

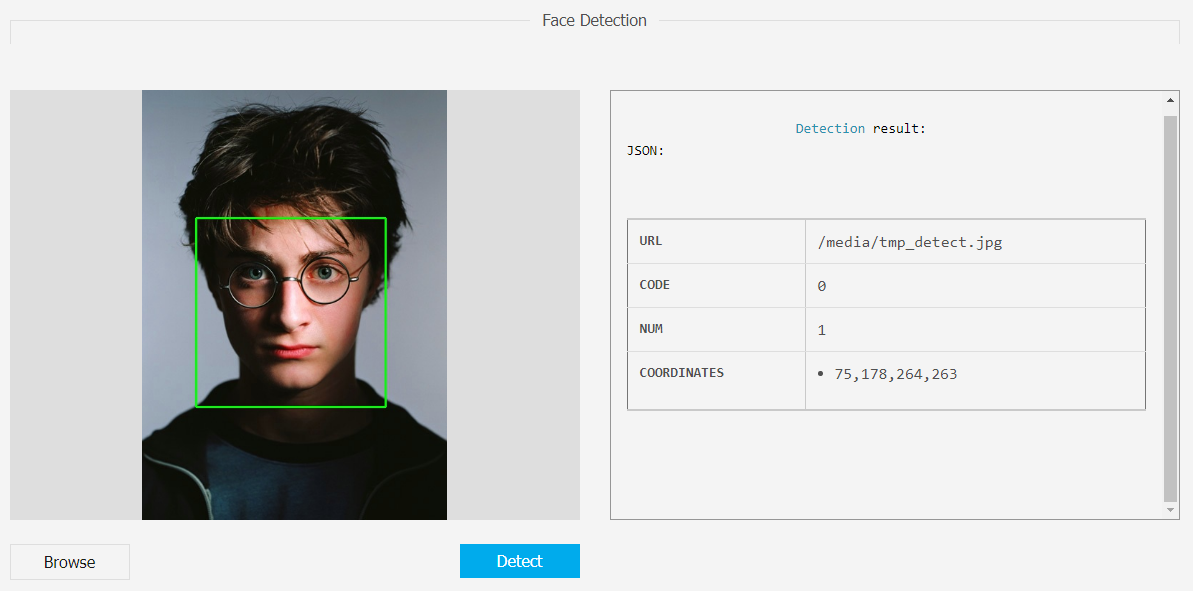


Hình 4.3. Cấu trúc Demo Website

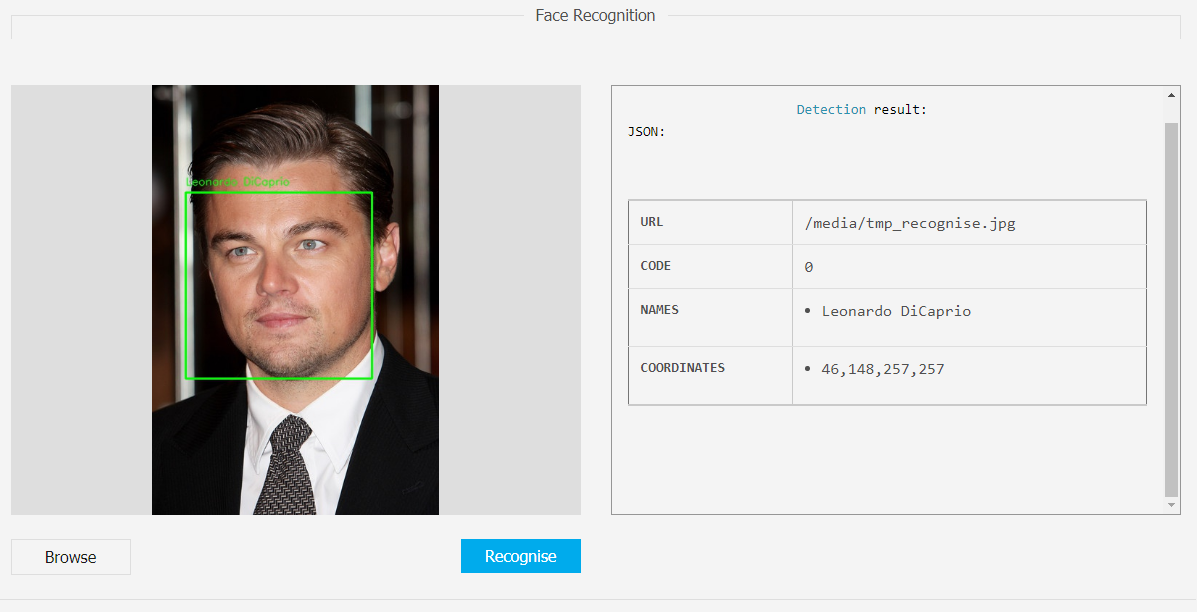
Qui trình hoạt động chi tiết của Detection Site và Recognition Site bao gồm các bước :

1. Người dùng duyệt ảnh bằng nút “Browse” và thực hiện xử lý bằng nút “Detect” hoặc “Recognise”: ảnh sẽ được upload lên server của Demo Website.
2. Server lưu trữ ảnh upload lên và tạo request đến server chuyên trách Detect hoặc Recognise bằng phương thức POST với dữ liệu gửi đi là ảnh vừa nhận
3. Server nhận kết quả trả về dạng JSON, download ảnh được visualized từ server chuyên trách, sau đó render các thông tin này để trả về cho người dùng.

Hình 4.4 trình bày hai ví dụ về giao diện kết quả của Detection Site (a) và Recognition Site (b). Hình ảnh thể hiện được vẽ khung cho từng khuôn mặt (có thêm tên ở góc trái trên nếu là nhận diện) và bên phải là bảng các thông số trả về (xem Bảng 4‑5).



***(a) Detection Site***



***(b) Recognition Site***

Hình 4.4. Ví dụ về giao diện kết quả của Detection Site (a) và Recognition Site (b)

## Thông tin tổng hợp từ video số dựa trên phát hiện và nhận biết mặt người

### Dữ liệu đầu ra dạng thô của hệ thống

Đề tài của nhóm sinh viên hướng tới xử lý các video số như phim, tin tức, kênh giải trí,… có thời lượng lớn và đó là một cản trở lớn cho người xem khi không đủ thời gian.

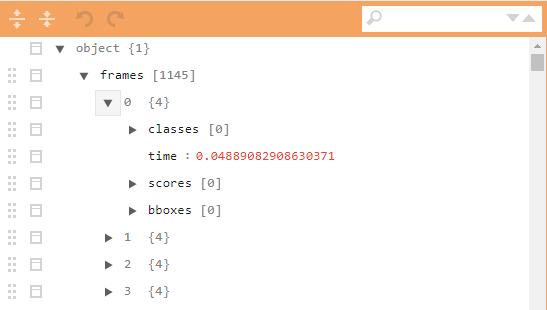
Các video này sẽ được sử dụng làm đầu vào, hệ thống sẽ phân tích và tổng hợp thông tin dựa trên phát hiện và nhận biết mặt người. Sau đó, đầu ra là thông tin chi tiết về nội dung từng frame ảnh (xem Bảng 4‑7).

Bảng 4‑7. Cấu trúc thông tin kết quả của một frame hình.

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Nội dung** |
| 1 | classes | Mảng các bộ gồm N số. N là số lượng diễn viên được nhận diện trong video và mỗi phần tử của bộ cho biết tỷ lệ mà người đó được dự đoán. |
| 2 | scores | Mảng các số, mỗi số ứng với tỷ lệ vùng bounding box là được dự đoán là mặt người. |
| 3 | time | Tổng thời gian xử lý theo giây. |
| 4 | bboxes | Mảng các bộ gồm 4 số. Mỗi bộ là một bounding box với các phần tử lần lược là tọa độ y\_min, x\_min, y\_max, x\_max. |

Để dễ dàng xử lý và tránh quá tải bộ nhớ, các video được cắt thành từng đoạn nhỏ hơn rồi mới đưa vào hệ thống. Mỗi video nhỏ hơn đưa vào hệ thống sẽ cho ra một tập tin định dạng JSON chứa kết quả xử lý của từng frame và giữ đúng thứ tự của chúng. Hình 4.5. Một tập tin đầu ra dạng JSON.Hình 4.5 là ví dụ cho một tập tin đầu ra ứng với một video nhỏ đưa vào.

Với mục đích tổng hợp thông tin và thể hiện dưới dạng cô động, dễ dàng sử dụng nhất có thể, chúng em đã xây dựng một ứng dụng xử lý các tập tin đầu ra dạng thô và trình bày chi tiết trong phần tiếp theo.



Hình 4.5. Một tập tin đầu ra dạng JSON.

*(Công cụ hiển thị sử dụng:* <http://jsoneditoronline.org/>*)*

### Ứng dụng Smart Video Editor – tổng hợp thông tin video

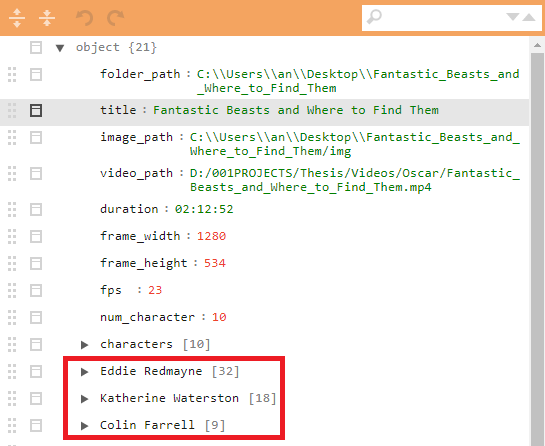
Ứng dụng này được phát triển bằng ngôn ngữ C# trên nền tảng .NET framework 4.5.2 với mục địch tổng hợp thông tin từng frame rời rạc dạng thô (4.2.1) thành dạng cô đọng và dễ dàng sử dụng hơn.

Smart Video Editor được thiết kế dành cho nhà biên tập với đầu vào là dữ liệu nội dung thô của từng frame và đầu ra là thông tin chi tiết của video bao gồm các nhân vật chính, mỗi nhân vật xuất hiện ở những phân đoạn nào, … (xem Bảng 4‑8).

Bảng 4‑8. Cấu trúc tập tin đầu ra dạng JSON của Smart Video Editor

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Nội dung** |
| 1 | folder\_path | Đường dẫn đến thư mục input |
| 2 | title | Tên video |
| 3 | image\_path | Đường dẫn đến thư mục ảnh các diễn viên. Mỗi người có một ảnh và đặt tên theo định dạng **<tên>.jpg** |
| 4 | video\_path | Đường dẫn đến video |
| 5 | duration | Tổng thời lượng video |
| 6 | frame\_width | Chiều rộng frame ảnh |
| 7 | frame\_height | Chiều cao frame ảnh |
| 8 | fps | Tốc độ video |
| 9 | num\_character | Số lượng diễn viên |
| 10 | characters | Danh sách tên diễn viên |
| 11 | **<Tên diễn viên>** | Mảng chứa các phân đoạn. Mỗi phần tử chứa thông tin: tên diễn viên, frame bắt đầu, frame kết thuc, thời điểm bắt đầu, thời điểm kết thúc. |

Hình 4.6 là một ví dụ về tập tin đầu ra của ứng dụng Smart Video Editor, trong đó phần khung màu đỏ là các thuộc tính ứng với tên diễn viên chứa các phân đoạn mà họ xuất hiện.



Hình 4.6. Ví dụ tập tin đầu ra của Smart Video Editor (khung đỏ là các thuộc tính ứng với từng diễn viên)

*(Công cụ hiển thị sử dụng:* <http://jsoneditoronline.org/>*)*

Đầu vào của ứng dụng là một thư mục chứa dữ liệu thô xuất ra từ quá trình phát hiện và nhận biết khuôn mặt có cấu trúc như sau:

* Thư mục “img”: chứa ảnh của các diễn viên. Tên tập tin đặt theo cú pháp

<tên diễn viên>.jpg

* Thư mục “meta”: chứa các tập tin json nội dung các frame. Mỗi tập tin tương ứng với một video nhỏ được cắt ra khi xử lý và đặt tên theo cú pháp:

**<số thứ tự (3 chữ số)>.json**

* Tập tin “info.json”: gồm các thông tin được mô tả trong Bảng 4‑9 (xem ví dụ Hình 4.7)

Bảng 4‑9. Cấu trúc tập tin “info.json”

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Nội dung** |
| 1 | title | Tên video |
| 2 | path | Đường dẫn đến video |
| 3 | duration | Tổng thời lượng video |
| 4 | frame-width | Chiều rộng frame ảnh |
| 5 | frame-height | Chiều cao frame ảnh |
| 6 | fps | Tốc độ video |
| 7 | num-meta | Số lượng file json (số lượng video được cắt nhỏ) |
| 8 | num-character | Số lượng diễn viên |
| 9 | characters | Danh sách tên diễn viên |

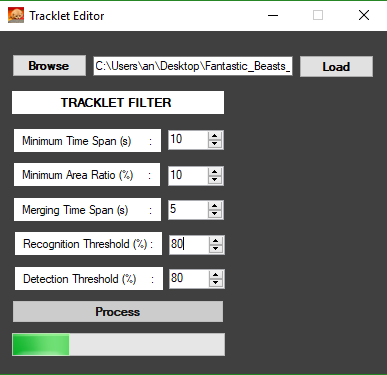


Hình 4.7. Ví dụ tập “info.json”

*(Công cụ hiển thị sử dụng:* <http://jsoneditoronline.org/>*)*

Như đã trình bày ở trên, Smart Video Editor là ứng dụng danh cho biên tập viên để tổng hợp thông tin là các phân đoạn mà các diễn viên xuất hiện. Trong quá trình tổng hợp đó, nhà biên tập có thể lựa chọn và thay đổi các thông số đặc trưng cho phù hợp với từng video cụ thể, bao gồm (xem thêm Hình 4.8):

* Thời lượng tối thiểu cho mỗi phân đoạn được phát hiện.
* Tỷ lệ diện tích tối thiểu để một khuôn mặt được tính là có xuất hiện trong frame.
* Khoảng thời gian giữa hai phân đoạn của cùng một diễn viên để gộp thành một.
* Độ chính xác tối thiểu cho thuật toán phát hiện khuôn mặt.
* Độ chính xác cho thuật toán nhận biết mặt người.



Hình 4.8. Giao diện hoạt động của ứng dụng Smart Video Editor.

## Person-based news highlight

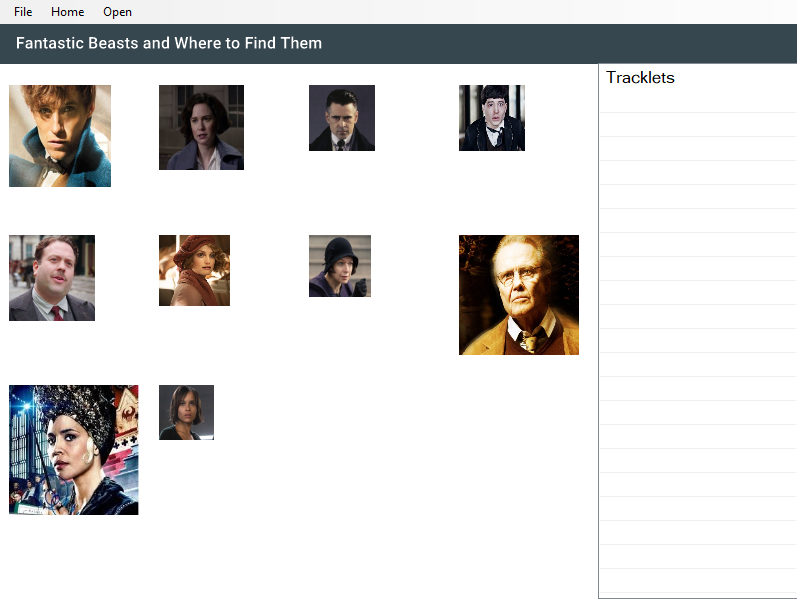
### Ngữ cảnh sử dụng

Ứng dụng Person-based news highlight được xây dựng bằng ngôn ngữ C# trên nền tảng .NET Framwork 4.5.2 với mục đích mang đến một khả nặng tương tác thông minh và trãi nghiệm thú vị cho người xem. Bên cạnh đó ứng dụng này đặc biệt hữu ích cho những người có cuộc sống, làm việc bận rộn nhưng vẫn muốn theo dõi nội dung phim, tin tức và đặc biệt là nhân vật, diễn viên mà họ yêu thích.

Person-based news highlight sử dụng dữ liệu được output từ Smart Video Editor (4.2.2) để thể hiện trực quan một bộ phim thành dạng poster. Trong đó bao gồm khuôn mặt với kích thước khác nhau cũng các diễn viên tùy thuộc vào độ quan trọng của họ. Độ quan trọng này được xác định dựa trên tổng thời lượng mà họ xuất hiện. Ứng với mỗi khuôn mặt là một dãy các phân đoạn mà họ xuất hiện giúp người xem có thể lướt nhanh qua nội dung và xem những phần chính họ mong muốn. Bằng cách tiếp cận này, ứng dụng đã tiết kiệm được rất nhiều thời gian cho người dùng và mang đến một trãi nghiệm mới.

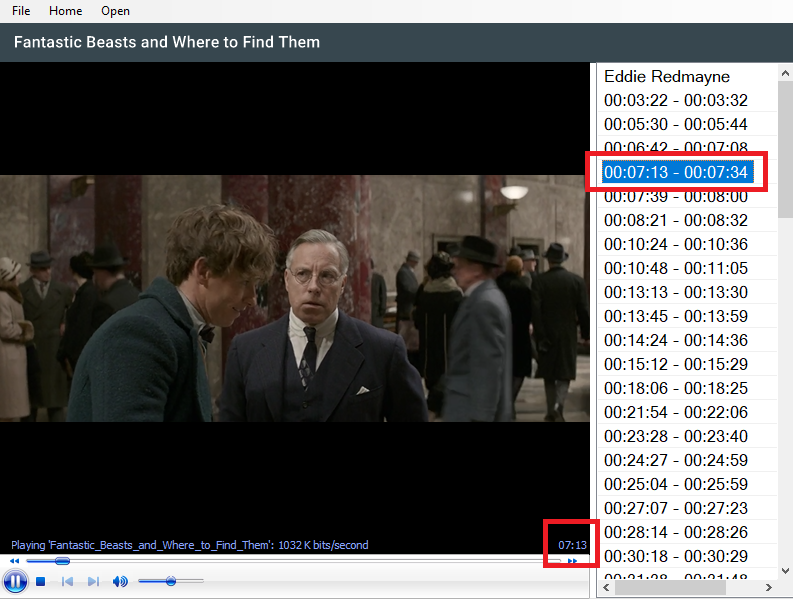
### Giao diện và hệ thống chức năng

Phần 4.3.2 tập trung giới thiệu giao diện trực quan và hệ thống chức năng hỗ trợ của ứng dụng Person-based news highlight. Hình 4.9 là giao diện màn hình **Home** sau khi tải dữ liệu bằng menu **Open**.



Hình 4.9. Màn hình Home sau khi load dữ liệu của Person-based News Highlight.

Khi nhấn vào ảnh của diễn viên quan tâm thì cửa sổ xem video cùng danh sách các phân đoạn tương ứng hiện lên cho người dùng lựa chọn (Hình 4.10). Về tổng thể video vẫn chạy tuần tự và chỉ khi người dùng chọn phân đoạn nào đó thì video mới chuyển đến phần tương ứng. Ngoài ra, có thể nhấn **Home** để quay về giao diện chính và lựa chọn nhân vật khác.



Hình 4.10. Giao diện xem video theo các phân đoạn của diễn viên

Eddie Redmayne.

## Character-based movie synopsis

Giới thiệu khái quát chức năng, input, output

### Ngữ cảnh sử dụng

### Hệ thống chức năng

## Character-based filter

Giới thiệu khái quát chức năng, input, output

### Ngữ cảnh sử dụng

### Kiến trúc hệ thống

### Hệ thống chức năng

## Kết luận

Trong chương này, chúng em đã trình bày về các thành phần chứng thực người dùng của hệ thống tương tác thông minh, bao gồm phân hệ chứng thực người dùng để đăng nhập vào hệ điều hành Windows và chứng thực với các dịch vụ trực tuyến và các chức năng cũng như quy trình hoạt động của từn phân hệ một cách tổng quan. Trong Chương 5, chúng em sẽ trình bày cụ thệ kiến trúc và quy trình của các phân hệ này trong hệ thống tương tác thông minh do chúng em đề xuất.

# Kết luận

* Nội dung của Chương 5 trình bày các kết quả đạt được và hướng phát triển của đề tài.

## Các kết quả đạt được

## Hướng phát triển của đề tài

Chúng em sẽ tiếp tục tìm hiểu sâu hơn về đề tài phát hiện và nhận biết mặt người, trong đó chú trọng nghiên cứu và phân tích các kiến trúc sử dụng convolutional neural network. Trên nền tảng đó đề xuất các kiến trúc mới để giải quyết các vấn đề còn tồn động như tư thế khuôn mặt (pose), điều kiện chiếu sáng (illumination), che khuất (occlustion), lão hóa (aging), …. Tham khảo các công trình tiên tiến và các tập dữ liệu tiêu chuẩn để có công bố một công trình khoa học hoàn chỉnh.

Về mặt ứng dụn và hiện thực hóa đề tài, nhóm sinh viên sẽ nghiên cứu thêm hai lĩnh vực là xác nhận khuôn mặt (face verification) và định danh (face identification). Sau đó, phát triển thành một hệ thống Face Web APIs hoàn chỉnh với bốn nhóm modules. Hiện tại với cùng ý tưởng này thì đã có hai hệ thống APIs nổi tiếng và hiệu quả đang hoạt động của Google và Microsoft Project Oxford. Chúng em sẽ tìm hiểu và phát triển một hệ thống tương tự để hỗ trợ cho các ứng dụng trong nước và các đề tài nghiên cứu liên quan. Chức năng mà nhóm đang hướng tới là cho phép người dùng tự tạo ra tập dữ liệu cá nhân và hệ thống sẽ tự động huấn luyện để có mô hình định danh cho những người trong tập dữ liệu ấy.

Bên cạnh đó, để phục vụ tốt hơn cho người dùng, nhóm tập trung vào tối ưu hiệu suất cho các mô hình tính toán. Với các cấu trúc nhiều lớp phức tạp thì khả năng đáp ứng cho các ứng dụng realtime là không khả thi khi giao tiếp theo kiến trúc client-server qua các Web APIs. Chúng em hướng tới các giải pháp tối ưu như sau: triển khai hệ thống APIs trên các server có cấu hình mạnh và đặc biệt là sử dụng GPU để tăng tốc độ xử lý, thu nhỏ kiến trúc mô hình để tạo thành các framework offline hỗ trợ cho các ứng dụng realtime trên máy cá nhân và thiết bị di động.

Đối với hệ thống tương tác thông minh hiện tại trong việc hỗ trợ người dùng xem các video, nhóm sinh viên tiếp tục phát triển khả nặng tương tác bằng cách tích hợp các thiệt bị và sensor hiện đại vào hệ thống. Định hướng của nhóm là cho phép người dùng có thể xem trên các mặt phẳng khác như cửa sổ, bàn, tường, … thay vì màn hình LCD như hiện tại. Bên cạnh đó, sử dụng thêm các camera đặc biệt như Intel Realsense, PrimeSense hay Kinect để hỗ trợ giao tiếp thông qua cử chỉ tay. Từ các cải tiến đó, nhóm hi vọng mang lại trãi nghiệm thú vị nhất cho người dùng.

# Tài liệu tham khảo

|  |  |
| --- | --- |
| [1] | N. H. Barnouti, S. S. M. Al-Dabbagh and W. E. Matti, "Face Recognition: A Literature Review," in *International Journal of Applied Information Systems 2016 (IJAIS 2016)*, New York, 2016. |
| [2] | O. M. Parkhi, A. Vedaldi and A. Zisserman, "Deep Face Recognition," in *British Machine Computer Vision 2015 (BMVC 2015)*, Swansea, 2015. |
| [3] | Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in *Computer Vision and Pattern Recognition 2014 (CVPR 2014)*, Columbus, 2014. |
| [4] | Y. Sun, D. Liang, X. Wang and X. Tang, "DeepID3: Face Recognition with Very Deep Neural Networks," in *CoRR, abs/1502.00873*, 2015. |
| [5] | T. PHAN-DUONG and M.-P. NGUYEN, "Chứng thực với thiết bị di động cho môi trường tương tác thông minh," Hochiminh, 2015. |
| [6] | K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *CoRR abs/1409.1556*, 2014. |
| [7] | C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going deeper with convolutions," arXiv:1409.4842, 2014. |
| [8] | G. B. Huang, M. Ramesh, T. Berg and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Technical Report 07-49, University of Massachusetts*, Amherst, 2007. |
| [9] | F. Schroff, D. Kalenichenko and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015. |
| [10] | L. Wolf, T. Hassner and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. CVPR*, 2011. |
| [11] | Y. Sun, X. Wang and X. Tang, "Sparsifying Neural Network Connections for Face Recognition," in *CVPR*, 2016. |
| [12] | I. Masi, S. Rawls, G. Medioni and P. Natarajan, "Pose-Aware Face Recognition in the Wild," in *CVPR*, 2016. |
| [13] | B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark-A," in *CVPR*, 2015. |
| [14] | Y. Wen, Z. Li and Y. Qiao, "Latent Factor Guided Convolutional Neural Networks for Age-Invariant Face Recognition," in *CVPR*, 2016. |
| [15] | K. R. Jr and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *FG*, 2006. |
| [16] | C. Xiong, X. Zhao, D. Tang, K. Jayashree, S. Yan and T.-K. Kim, "Conditional Convolutional Neural Network for Modality-aware Face Recognition," in *ICCV*, 2015. |
| [17] | R. Gross, I. Matthews, J. F. Cohn, T. Kanade and S. Baker, "Multi-PIE," in *Proceedings of the Eighth IEEE International Conference on Automatic Face and Gesture Recognition*, 2008. |
| [18] | D. K. Pal, F. Juefei-Xu and M. Savvides, "Discriminative Invariant Kernel Features: A Bells-and-Whistles-Free Approach to Unsupervised Face Recognition and Pose Estimation," in *CVPR*, 2016. |
| [19] | W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg, "SSD: Single Shot MultiBox Detector," in *ECCV*, 2016. |
| [20] | A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional," in *Conference on Neural Information Processing Systems NIPS 2012*, 2012. |
| [21] | K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations ICLR 2015*, 2015. |
| [22] | O. M. Parkhi, K. Simonyan, A. Vedaldi and A. Zisserman, "A compact and discriminative face track descriptor," in *Proc. CVPR*, 2014. |
| [23] | Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "Web-scale training for face identification," in *Proc. CVPR*, 2015. |
| [24] | H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *Proc. IEEE International Conference on Image Processing*, Paris, 2014. |
| [25] | M. Grgic, K. Delac and S. Grgic, "SCface - surveillance cameras face database," *Multimedia Tools and Applications Journal,* vol. 51, pp. 863-879, 2011. |
| [26] | D. Yi, Z. Lei, S. Liao and S. Z. Li, "Learning Face Representation from Scratch," in *arXiv preprint arXiv:1411.7923.*, 2014. |
| [27] | S. Milborrow, J. Morkel and F. Nicolls, "The MUCT Landmarked Face Database," in *Pattern Recognition Association of South Africa*, 2010. |
| [28] | A. Savran, N. Alyüz, H. Dibeklioglu, O. Çeliktutan, B. Gökberk, B. Sankur and L. Akarunh, "Bosphorus Database for 3D Face Analysis," in *Biomedical Innovation and Development Conference*, 2008. |
| [29] | D. Chen, X. Cao, L. Wang, F. Wen and J. Sun, "Bayesian face revisited: A joint formulatio," in *ECCV* , Springer, 2012. |
| [30] | Z. Liu, P. Luo, X. Wang and X. Tang, "Deep Learning Face Attributes in the Wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. |
| [31] | B.-C. Chen, C.-S. Chen and W. H. Hsu, "Face Recognition using Cross-Age Reference Coding with Cross-Age Celebrity Dataset," in *IEEE Transactions on Multimedia*, 2015. |
| [32] | T. Sim, S. Baker and M. Bsat, "The CMU Pose, Illumination, and Expression (PIE) Database," in *International Conference on Automatic Face and Gesture Recognition*, 2002. |
| [33] | F. Wiles, D. Procida, J. Bennett, R. Conley, K. Love and K. W. Alger, "Django," Django Software Foundation, [Online]. Available: https://www.djangoproject.com/. |

1. Mã nguồn được lấy từ trang: https://github.com/balancap/SSD-Tensorflow [↑](#footnote-ref-1)