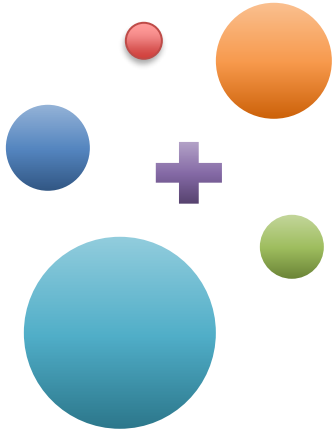
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HOC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Báo cáo đồ án

Face Recognition At a Distance (FRAD)



GVHD: Nguyễn Ngọc Thảo

Mục lục

THÔNG TIN NHÓM

Thành viên 1: Trần Quang Huy - 1512211

Thành viên 2: Phùng Tiến Hào - 1612174

Thành viên 3: Võ Quốc Huy – 1612269

1. Giới thiệu về Nhận dạng từ xa (FRAD)

- Nhận dạng khuôn mặt đã có những bước tiến lớn trong những năm qua. Tuy nhiên, đa số các phương pháp nhận dạng khuôn mặt đòi hỏi sự hợp tác của đối tượng trong phạm vi gần nên giới hạn trong việc áp dụng vào thực tế.
- Từ lý do đó mà nhận dạng từ xa (Face recognition at a distance – FRAD) ra đời. Với FRAD, người dùng có thể tự động nhận dạng khuôn mặt mà không cần sự hợp tác của đối tượng và trong phạm vi lớn.
- FRAD có rất nhiều áp dụng trong thực tế: Cửa tự động, Nhận dạng tội phạm, Tìm hiểu hành động của khách hàng,…

2. Các thách thức

Thách thức chính của FRAD là lấy được hình ảnh tốt nhất từ xa mà không có sự trợ giúp của đối tượng. Thách thức này có thể chia thành các chuyên mục sau:

2.1. Optics and Light Intensity

- Khi khoảng cách giữa camera và đối tượng tăng, tiêu cự và đường kính của ống nhìn phải được tăng lên để đảm bảo phạm vi quan sát và độ sáng.
- Tuy nhiên, sẽ có lúc tiêu cự và đường kính không thể tăng được nữa. Khi đó, ta cần có ống kính lớn hơn, việc đó sẽ tốn kém tiền bạc và sẽ làm camera nặng hơn.
- Ngoài ra, việc tăng đường kính của ống kính sẽ làm giảm độ sâu. Tuy nhiên, việc này thường không phải là vấn đề của FRAD.

2.2. Exposure Time and Blur

- Khi ống kính thích hợp được chọn, độ sáng của tấm ảnh chụp từ xa sẽ bằng với tấm ảnh chụp gần. Tất nhiên, điều này không lúc nào cũng khả thi. Khi đó, ta cần phải tăng độ sáng của bức ảnh. Nhưng điều này sẽ tăng độ nhiễu.
- Ngoài cách đó, ta cũng có thể tăng tốc độ lá chắn sáng. Mặt xấu là làm vậy sẽ gây motion blur.
- Độ rung của máy ảnh cũng có thể gây mờ.

2.3. Image Resolution

- Khi kích thước giữa camera và đối tượng càng tăng, độ phân giải của bức ảnh càng giảm. Đôi lúc, ta vẫn muốn lấy đối tượng nằm ngoài giới hạn khoảng cách của máy ảnh. Việc này đòi hỏi một hệ thống nhận dạng có thể xử lý hình ảnh với độ phân giải thấp

2.4. Pose, Illumination and Expression

- Khó có thể bắt được đối tượng trong tư thế thích hợp trong không gian rộng, có nhiều hướng đi. Để giải quyết vấn đề này, ta có thể tăng thời gian quan sát để có nhiều đối tượng với tư thế thích hợp hoặc camera ở vị trí mà chỉ cho phép một hướng di chuyển.
- Ngoài tư thế còn có vấn đề ánh sáng. Hệ thống nhận dạng thường đặt ở ngoài trời và dựa vào ánh sáng của mặt trời và các vật phát sáng cố định. Ánh sáng có thể bị ảnh hưởng bởi thời gian, khí hậu và các vật xung quanh.
- Còn vấn đề về biểu hiện thì thường không là vấn đề của FRAD vì các đối tượng không cố gắng né tránh hay noi chuyện thường có biểu hiện bình thường

3. Các hướng tiếp cận

3.1. High-Definition Stationary Camera

- Một hướng giải quyết vấn đề của FRAD là sử dụng camera cố định với độ phân giải cao. Nếu một hế thống phải hoạt động trong không gian rộng 20 m và đòi hỏi hình ảnh khuôn mặt có 100 pixel, ta cần một camera với độ phân giải ngang 15000 pixel. Nếu phạm vi quan sát tăng, độ phân giải của camera phải tăng.

3.2. Active-Vision Systems

- FRAD thường được giải quyết với hệ thống nhiều camera với một hay nhiều camera với phạm vi rộng (WFOV) với độ phân giải thấp để quan sát và xác định đối tượng và một hay nhiều camera với phạm vi hẹp (NFOV) có độ phân giải cao để bắt được hình ảnh mặt. WFOV và NFOV thường được gọi là master và slave camera.
- Có rất nhiều cách có thể áp dụng phương pháp này. Ta có thể dùng một camera duy nhất có thể chuyển đổi giữa WFOW và NFOV, hay một NFOV camera có thể quan sát một khu vực rộng.
- Khi có nhiều đối tượng, ta cần phải biết phân bố tài nguyên và thời gian để NFOV camera có thể bắt được hình ảnh, đặt biệt khi có nhiều NFOV camera. Có rất nhiều yếu tố ảnh hưởng đến điều này như hướng di chuyển và tốc độ của đối tượng, độ phân giải và chất lượng hình ảnh lấy được, số lần chụp mặt của từng đối tượng,... Cần có thuật toán dựa vào các yếu tố này để điều khiển NFOV camera.

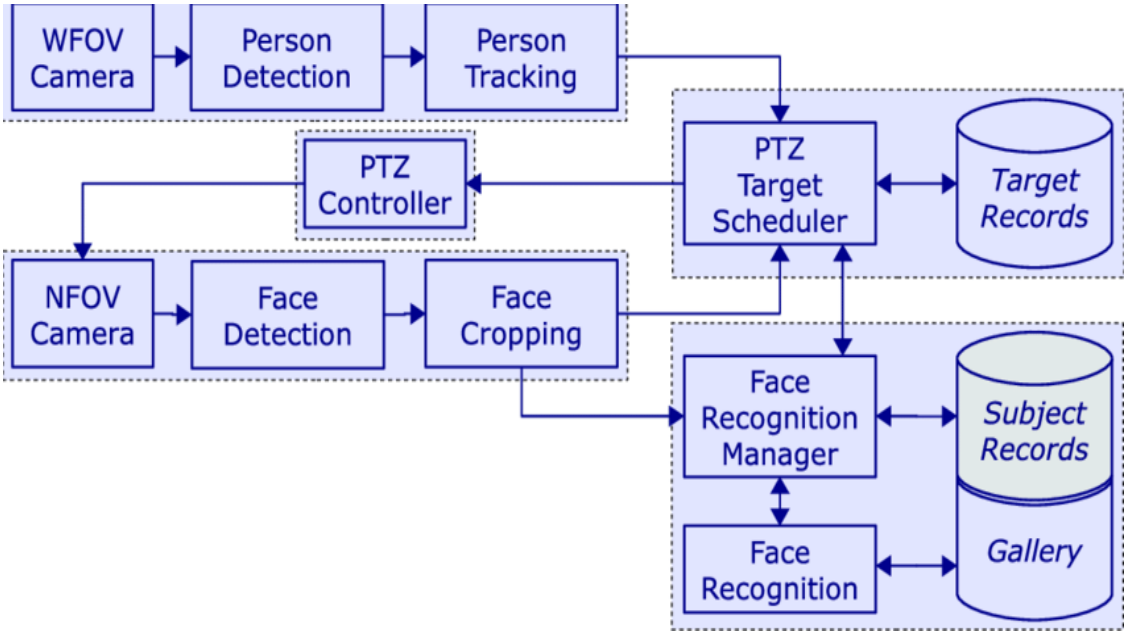


Figure 1: Biểu đồ các thành phần của hệ thống giám sát

4. Literature Review

4.1. Databases

Hầu hết các test databases cho hệ thống nhận dạng mặt người (face recognition) chứa ảnh hoặc video thu được ở pham vi gần với dối tượng có hợp tác. Chúng rất phù hợp để huấn luyện và xác thực nhận dạng cho các ứng dụng kiểm soát ra vào (Access control system). Tuy nhiên, thực tế là rất ít tập dữ liệu thích hợp cho hệ thống nhận dạng ở khoảng cách.

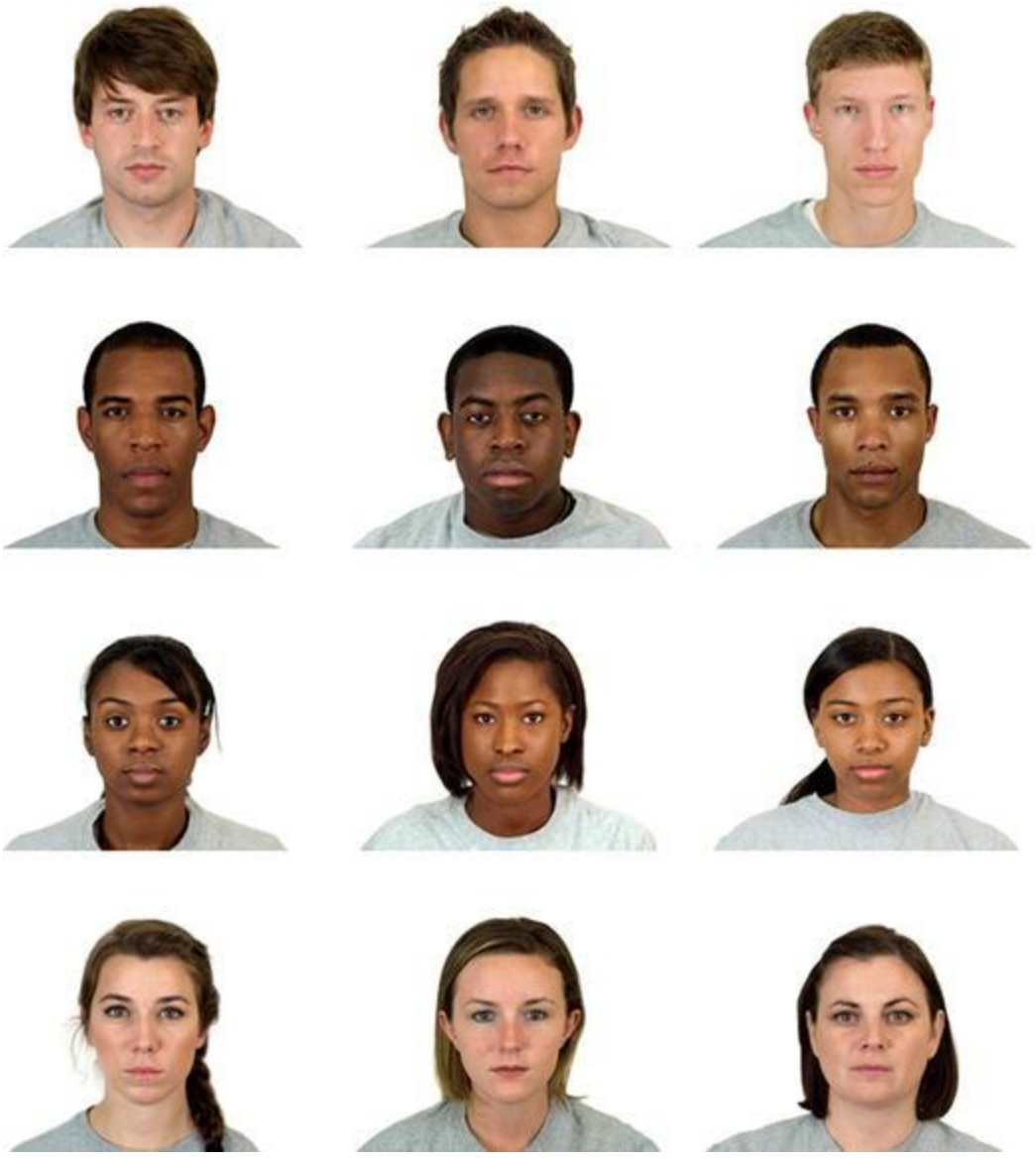


Figure 2:The Chicago face database

Figure 3: Access control system

Đại học Texas ở Dallas có thu thập database cho chương trình DARPA Human ID gồm các ảnh và video tĩnh của các đối tượng và đồng thời video của người đi bộ tiến gần đến camera từ khoảng cách đến 13.6 m và video của những người đang nói chuyện và thể hiện cử chỉ từ khoảng 8 m. Bộ sưu tập này được triển khai ở hệ thống kín nhưng ở không gian mở được làm từ các vách kính có thể xấp xỉ điều kiện sáng ngoài trời. Hệ số zoom cũng được tính đến trong bộ sưu tập này.

Yao của đại học Tennessee thu thập face database có tên Knoxville Long Range High Magnification (UTK-LRHM) của các đối tượng có hợp tác ở khoảng cách giữa 10 m và 20 m trong không gian kín và khoảng cách cực lớn giữa 50 m và 300 m ngoài trời. Hệ số zoom trong không gian kín là khoảng 3 đến 20 và ngoài trời là phạm vi đến 284. Hình ảnh ở các khoảng cách có thể biến dạng do nhiệt độ không khí và độ dốc áp suất, và hệ thống quang học làm mờ thêm ở độ phóng đại như vậy.

NIST Multiple Biometric Grand Challenge(MBGC) tập trung vào nhận diện khuôn mặt và mống mắt (face and iris recognition) bằng cả hình ảnh và video và tài trợ cho một loạt các thử thách đặt ra. Để hỗ trợ cho các thách thức nhận dạng khuôn mặt không bị giới hạn, chương trình 360 đã thu thập video ngoài trời độ nét cao và độ nét tiêu chuẩn của các đối tượng đi về phía máy ảnh và đứng ở phạm vi lên tới khoảng 10 m. Khi các đối tượng tiến gần về phía máy ảnh, bắt rõ khuôn mặt của họ. MBGC cũng đang sử dụng dữ liệu ID người DARPA được mô tả trên.

Điều quan trọng cần nhớ là khi khoảng cách tăng lên, khả năng nhận diện khuôn mặt gặp khó khăn hơn.

Mỗi cơ sở dữ liệu này ghi lại hình ảnh hoặc video bằng máy ảnh tĩnh. Cảm biến của FRAD nói chung là hệ thống camera được điều khiển chủ động theo thời gian thực. Như vậy hệ thống rất

khó kiểm tra. Đánh giá hệ thống camera hoạt động với bộ dữ liệu được chia sẻ chung hoặc tiêu chuẩn hóa là không khả thi vì sự tích hợp phần mềm và phần cứng thời gian thực không được mô hình hóa. Các thành phần của các hệ thống này, chẳng hạn như Phát hiện khuôn mặt, phát hiện người, theo dõi và nhận diện khuôn mặt có thể được kiểm tra bằng sự cách ly trên bộ dữ liệu chia sẻ chung thích hợp. Nhưng tương tác giữa phần mềm và các thành phần phần cứng chỉ có thể được kiểm tra đầy đủ trên các cảnh live-stream trực tiếp. Môi trường ảo cũng có thể được sử dụng để kiểm tra nhiều khía cạnh của hệ thống quan sát chủ động (Active-Vision Systems).

4.2. Active-Vision Systems

Đã có rất nhiều đổi mới và hệ thống được phát triển cho diện rộng phát hiện người và theo dõi dùng máy ảnh NFOV để chụp ảnh khuôn mặt tại một khoảng cách

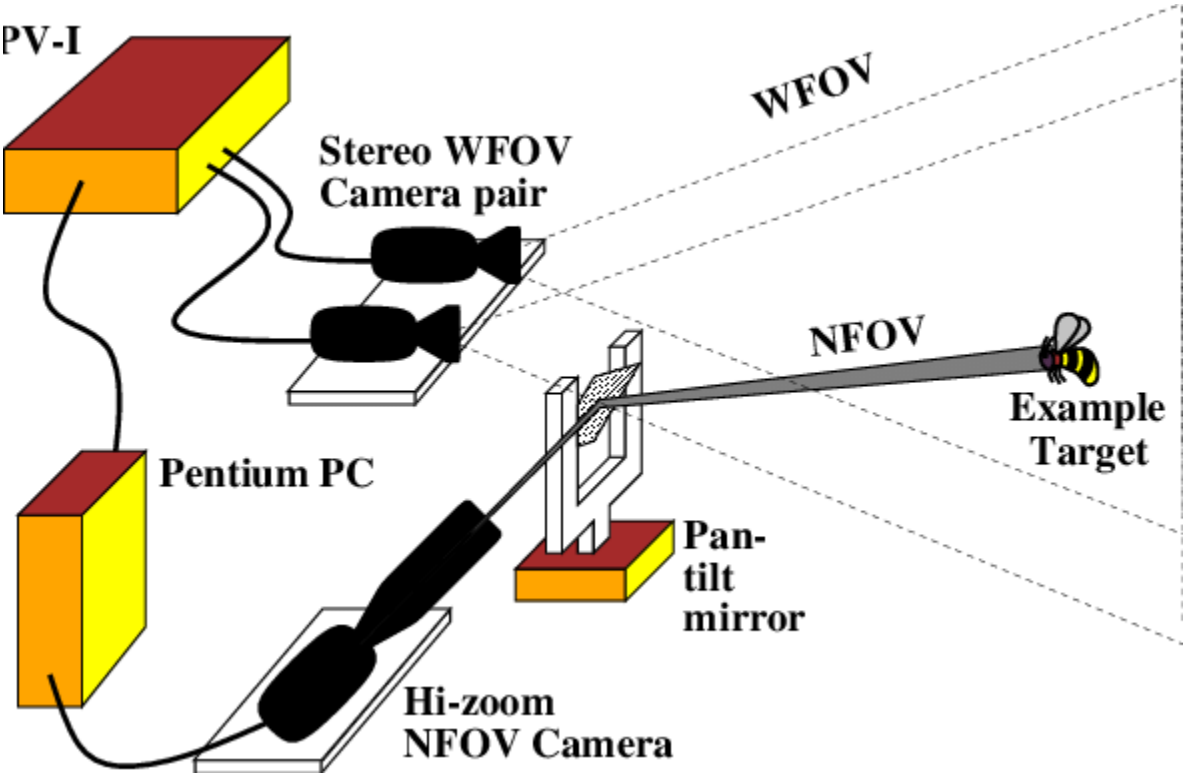
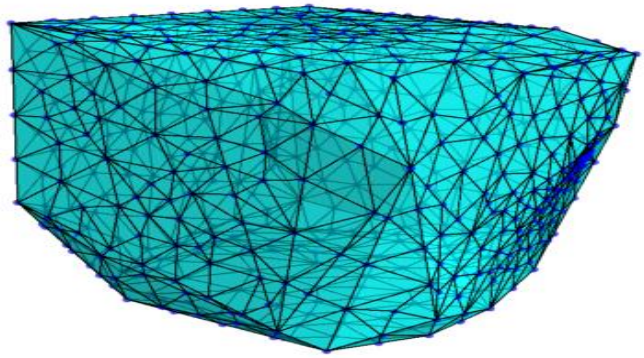


Figure 4: NFOV camera và WFOV camera

.

Dưới đây là một vài hệ thống điển hình đã ra đời theo thứ tự thời gian:

- Stillman và cộng sự đã phát triển một hệ thống camera chủ động để nhận dạng người bằng cách sử dụng hai camera WFOV và hai camera NFOV. Hệ thống này vẫn có những giới hạn về điều kiện trên phạm vi xa nhưng có khắc phục bằng cách nhận diện dựa vào màu da, vị trí 3D trangularization và hướng camera NFOV tới các mặt người. Điều này đã tạo động lực để làm ra hệ thống tính toán thông minh để phát hiện sự xuất hiện của người trong ảnh để cải thiện sự tương tác.



- Greiffenhagen mô tả một hệ thống chụp khuôn mặt camera kép trong đó camera WFOV là một camera đa hướng trên cao và NFOV có pan, tilt và zoom. Tác giả hiết kế hệ thống nhận diện theo thời gian thực và thống kê các đặc tính của các thành phần hệ thống sao cho độ đô không chắc chắn có thể được kiểm soát và xác suất chụp khuôn mặt là 0,99 vẫn được đảm bảo. Hệ thống này xử lý được nhiều đối tượng miễn sao xác suất bị che lấp là nhỏ.

Figure 5: Hệ thống dùng nhiều NFOV camera để giám sát và truy vết đối tượng.

- Zhou phát triển hệ thống Distant Human Identification (DHID) để thu thập thông tin sinh trắc học của con người tại khoảng cách cho nhận dạng mặt người và dáng (face recognition and gait). Một camera WFOV duy nhất có góc nhìn 60° và cho phép theo dõi những người ở khoảng cách 50 m. Phát hiện và theo dõi người bằng sự kết hợp của phép trừ nền, khác biệt thời gian, dòng quang và phát hiện blob dựa trên màu sắc. Trước tiên hệ thống phát hiện mặt, dáng đi bằng cách zoom-in với lượng nhỏ bằng camera WFOV sau đó camera NFOV theo vết đối tượng dựa vào đoạn video của WFOV.
- Marchesotti phát triển một máy ảnh chụp hai mặt (two-camera face capture) ở khoảng cách. Người bị phát hiện và theo dõi bằng cách sử dụng đặc trưng blob trong video WFOV, và một camera NFOV được dịch và nghiêng để thu được các đoạn video ngắn của khuôn mặt các đối tượng.

- Hampapu phát triển hệ thống phân hạng mục (cataloger system). Hệ thống dùng hai camera WFOV được phân tách rộng rãi với các góc nhìn chồng chéo nhau của không gian phòng thí nghiệm là 20 ft x 19 ft. Để phát hiện mặt người, thiết bị theo dõi 2D multi-blob dược áp đùng vào video của camera WFOV và kết quả này kết hợp với thiết bị theo dõi 3D multi-blob để xác định vị trí 3D của đầu người đó trong hệ tọa hiệu chỉnh độ chung. Đặc biệt, hệ thống này sẽ zoom lên khi mà đối tượng di chuyển chậm. Tác giả cũng cho thấy có sự đánh đổi giữa hệ số zoom của camera NFOV và xác suất thành công khi thu được mặt người. Khi camera NFOV zoom lên để tập trung vào vung điểm ảnh nhất định, khả năng thiếu sót đối tượng sẽ cao hơn. Senior kế thừa ý tưởng này đồng thời đơn giản hóa và cải tiến thủ tục để thực hiện ngoài trời.

- Bagdanov đã phát triển một phương pháp để chụp ảnh khuôn mặt trên khu vực rộng chỉ với một camera pan-tilt. Học tăng cường (Reinforcement learning) được sử dụng để điều khiển máy ảnh sao cho cực đại hóa khả năng thu được một ảnh mặt người mà có chuyển động của đôi tượng. Hệ thống sẽ tự động học như nếu, khi nào, ở đâu cần zoom lên để chụp. Cái lợi của phương pháp này là chỉ dùng có một camera và không có hiệu chỉnh gì cả.

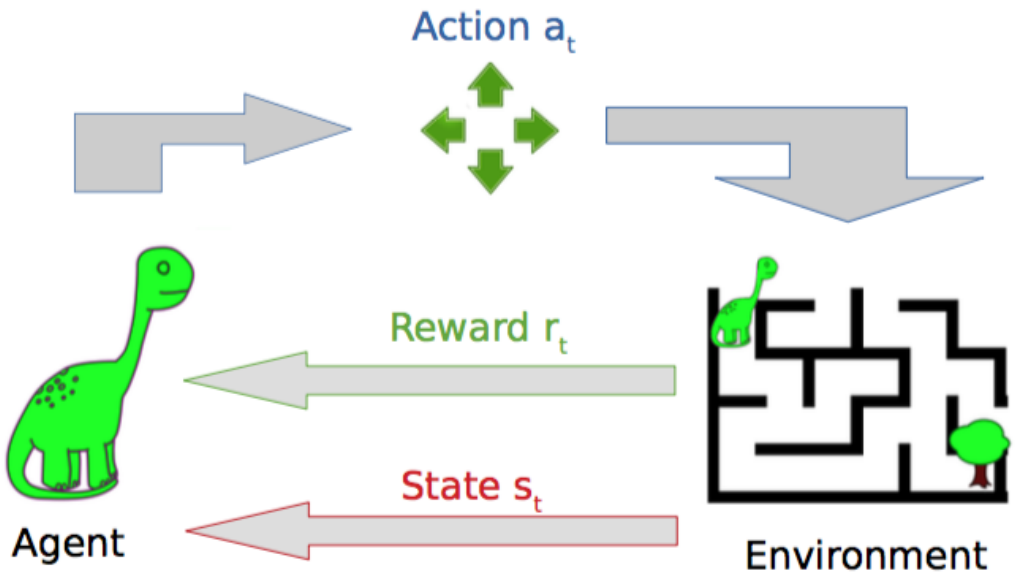


Figure 6: Reinforcement learning

- Prince, Elder đã phát triển hệ thống giải quyết các thách thức về sưu tập dữ liệu và tư thế. Họ sử dụng foveated sensor gồm một camera tĩnh với góc nhìn 135° và một foveated camera với góc nhìn 13°. Khuôn mặt được phát hiện qua camera cố định bằng cách sử dụng phát hiện chuyển động, mô hình nền và da. Sau đó, pan và tilt chỉ đạo foveal camera để phát hiện khuôn mặt.
- Davis phát triển phương pháp tự động quét một khu vực rộng và phát hiện người với một camera PTZ. phát hiện hành vi của con người và tìm hiểu tần suất con người xuất hiện trên toàn bộ vùng phủ sóng.

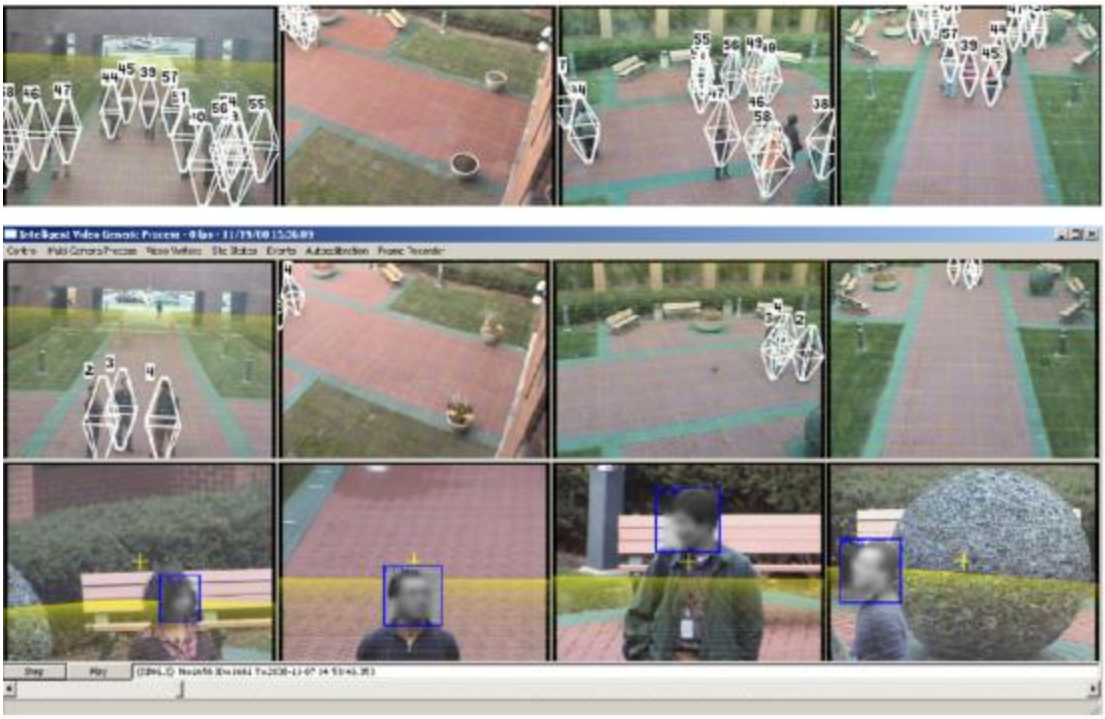


Figure 7: Hệ thống giám sát với nhiều camera để theo vết đối tượng từ xa, kể cả trong đám đông.

- Krahnstoever đã phát triển hệ thống chụp khuôn mặt ở một khoảng cách. Bốn camera cố định với các góc nhìn chồng chéo được sử dụng để theo dõi nhiều đối tượng trong một khu vực 10 m x 30 m. Theo dõi được thực hiện trong khung tọa độ trong thế giới thực, điều khiển việc nhắm mục tiêu và kiểm soát bốn camera PTZ riêng biệt bao quanh khu vực được giám sát. Các PTZ camera được kiểm soát để thua được ảnh có độ phân giải cao, bằng cách này sẽ giúp phát hiện dược nhiều khuôn mặt hơn trong một khung hình.
- Bellotto và cộng sự phát triển một kiến trúc cho nhiều camera quan set chủ động (active multi-camera surveillance) và chụp khuôn mặt trong đó các trình theo dõi được liên kết với mỗi camera và các thuật toán lý luận cấp cao (high-level reasoning algorithms) giao tiếp thông qua cơ sở dữ liệu SQL. Thông tin từ những người bị phát hiện bởi trình theo dõi WFOV được sử dụng để chỉ định các camera NFOV theo vết các đối tượng cụ thể với vận tốc có kiểm soát.
- Yu đã sử dụng hệ thống này để theo dõi các nhóm người theo thời gian, liên kết danh tính với từng người được theo dõi và ghi lại mức độ tương tác chặt chẽ giữa các cá nhân được xác định.

4.3. NFOV Resource Allocation

Vấn đề phân bổ tài nguyên luôn bị hạn chế trong thế giới thực khi triển khai mô hình dùng các camera chủ động. Với số lượng máy ảnh NFOV hạn chế và số lượng lớn các mục tiêu tiềm năng, việc dự đoán các khoảng thời gian khả thi trong tương lai là điều cần thiết. Trong đó một

mục tiêu có thể được chụp bởi máy ảnh NFOV ở độ phân giải và tư thế mong muốn, tiếp theo là lên lịch cho các camera NFOV dựa trên các cửa sổ tạm thời này.

Lim và cộng sự giải quyết vấn đề trước đây bằng cách xây dựng những gì đã biết như một "khoảng tầm nhìn nhiệm vụ" ("Task Visibility Interval"), đóng gói các thông tin yêu cầu.

Bimbo and Pernici giải quyết vấn đề lập lình chương trình NFOV cho chụp ảnh mặt người với một mạng camera chủ động (active camera network). Họ hình thành vấn đề như một bài toán nhân viên bán hàng đi du lịch Kinetic (Kinetic Traveling Salesman Problem - KTSP) để xác định làm thế nào để có được càng nhiều mục tiêu càng tốt.

Một loạt các chính sách lập lịch trình của NFOV đã được Costello và cộng sự phát triển và đánh giá.

Qureshi and Terzopoulos đã phát triển một trình mô phỏng môi trường ảo rộng lớn cho một nhà ga xe lửa lớn với tính tự trị thực tế theo hành vi người đi bộ di chuyển mà không va chạm và thực hiện hoạt động như chờ đợi xếp hàng, mua vé, mua đồ ăn thức uống, chờ tàu và tiến hành đến khu vực hòa nhạc.

Động cơ render video có thể xử lý các trường hợp như che khuất, mô hình máy ảnh jitter và phản ứng màu không hoàn hảo.

Mục đích là để để phát triển và thử nghiệm các hệ thống lập kế hoạch và kiểm soát máy ảnh với nhiều máy ảnh WFOV và nhiều máy ảnh NFOV ở quy mô lớn mà các thí nghiệm trong thế giới thực sẽ rất tốn kém. Tuy rằng hình ảnh mô phỏng sẽ không được chân thực nhưng hệ thống này cho phép thiết lập và đánh giá những chuyên viên làm nhiệm vu theo vết và các thuật toán lập lịch trình camera với hàng trăm đối tượng và máy ảnh trên một diện tích rất lớn.

4.4. Very Long Distances

Yao và cộng sự đã khám phá nhận dạng khuôn mặt ở khoảng cách đáng kể, sử dụng cơ sở dữ liệu khuôn mặt UTK-LRHM của họ. Đối với dữ liệu trong nhà, với một bộ sưu tập 55 người và một hệ thống nhận diện khuôn mặt thương mại, chúng cho thấy sự suy giảm tỷ lệ nhận dạng từ 65,5% đến 47,3% khi hệ số zoom tăng từ 1 đến 20 và khoảng cách chủ thể tăng để duy trì độ phân giải hình ảnh bằng mắt 60 pixel.

Nói thêm, tỷ lệ nhận dạng ở hệ số zoom là 20 có thể tăng lên tới 65,5% với khả năng làm mờ dựa trên bước sóng (wavelet-based deblurring)

Yao và cộng sự đã sử dụng phương pháp siêu phân giải(super-resolution approach) dựa trên tần suất đăng ký tên miền và phép nội suy cubic-spline trên hình ảnh khuôn mặt từ cơ sở dữ liệu khuôn mặt UTKLRHM. Siêu phân giải dường như hiệu quả nhất khi hình ảnh khuôn mặt bắt đầu ở độ phân giải thấp. Đối với hình ảnh khuôn mặt với khoảng 35 pixel mắt, độ phân giải siêu cao đã tăng tỷ lệ nhận dạng từ 10% lên 30% với tập dữ liệu 55 người. Độ phân giải siêu cao

cùng với mặt nạ unsharp để tăng nét trên ảnh nâng tỷ lệ nhận dạng lên 38% và mang lại hiệu suất đặc trưng tích lũy thích hợp (cumulative match characteristic performance) khá cao khi sử dụng zoom quang học để tăng gấp đôi độ phân giải của hình ảnh khuôn mặt.

4.5. 3D Imaging

Hầu hết hệ thống chụp mặt 3D dùng phương pháp stereo hoặc ánh sáng có cấu trúc (structured light).

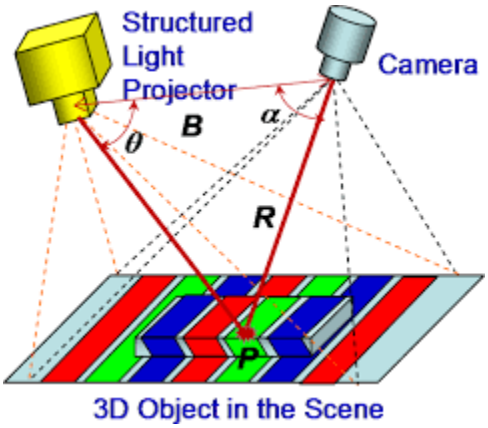


Figure 8: Phuong pháp structure light

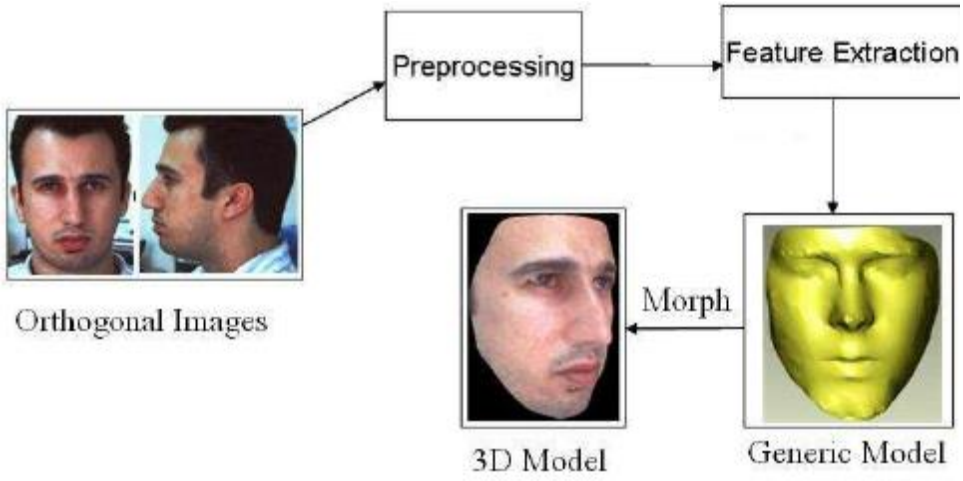


Figure 9: Phương pháp Stereo

Hệ thống chụp stereo sử dụng hai camera có mối quan hệ hình học biết trước. Khoảng cách đến các điểm đặc trưng (feature points) được phát hiện trong mỗi ảnh của camera sau đó được tìm thấy thông qua triangularization.

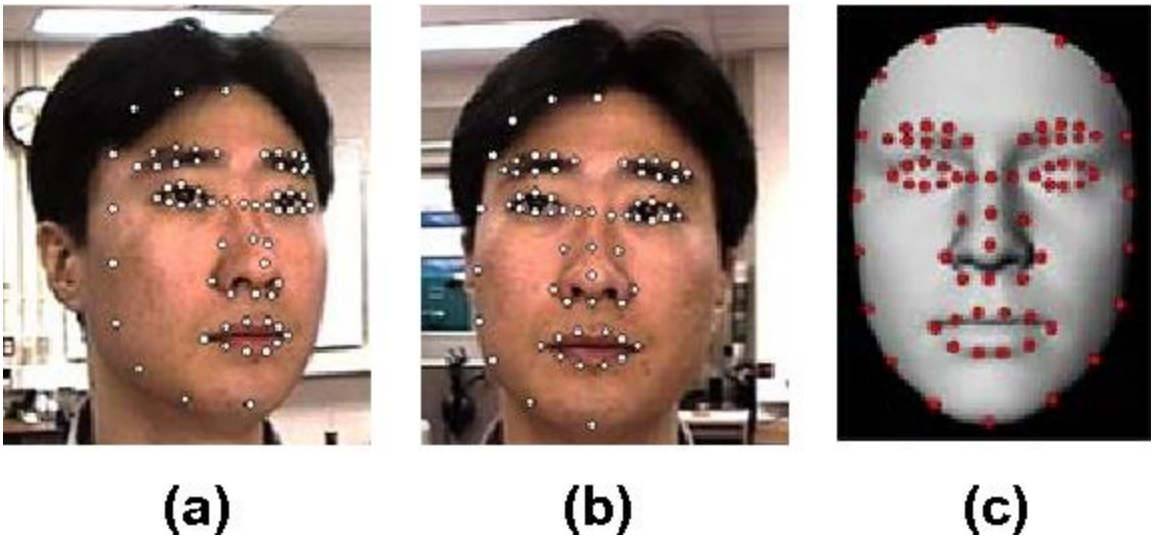


Figure 10: Feature points

Các hệ thống stuctured light sử dụng máy chiếu ánh sáng và máy ảnh, cũng có mối quan hệ hình học biết trước. Mẫu ánh sáng (light pattern) được phát hiện trong hình ảnh camera và các điểm 3D được xác định. Mỗi hệ thống được đặc trưng bởi một khoảng cách đường cơ bản (baseline distance), giữa các stereo camera hoặc giữa máy chiếu ánh sáng và camera.

Với 2 phương pháp trên, độ chính xác của triangulated 3D data suy giảm với khoảng cách chủ thể nếu khoảng cách baseline cố định. Để duy trì độ chính xác của tái tạo 3D khi khoảng cách chủ thể tăng, khoảng cách baseline phải tăng theo tỷ lệ. Điều này cấm một hệ thống nhỏ gọn về mặt vật lý và là một thách thức cơ bản đối với chụp khuôn mặt 3D ở khoảng cách xa với các phương pháp này.

Một số hệ thống mới dưới sự phát triển đang vượt qua thử thách cơ bản cho khuôn mặt 3D chụp ở khoảng cách xa.

Medioni và cộng sự giải quyết FRAD cho các cá nhân không hợp tác với một camera đơn và tái tạo khuôn mặt 3D. Họ đề xuất một hệ thống sử dụng máy ảnh có độ phân giải cực cao (ultra-high resolution) 3048 x 4560 pixel bằng cách chuyển đổi chế độ đọc. Có thể sử dụng các chỉ số tốc độ khung hình nhanh với độ phân giải thấp trên toàn khung hình để phát hiện, theo dõi và các chỉ số độ phân giải cao của một phần khung hình dùng để thu được một loạt các hình ảnh khuôn mặt của người bị phát hiện. Phát hiện người được thực hiện mà không cần mô hình nền, dựa trên đặc trưng cạnh (edgelet). Công trình này nhấn mạnh đến việc tái tạo khuôn mặt 3D với 100 pixel eye-to-eye sử dụng hình dạng từ chuyển động trên dữ liệu thu được với một prototye của hệ thống hình dung (envisioned system). Tái tạo 3D được thực hiện ở khoảng cách lên đến 9 m. Mặc dù các thử nghiệm hiện tại cho thấy nhận dạng khuôn mặt 2D vượt trội so với 3D, cả 2 có thể hợp nhất hoặc dữ liệu 3D có thể cho phép hiệu chỉnh tư thế.

Rara và cộng sự đạt được thông tin hình dạng khuôn mặt 3D ở khoảng cách lên tới 33 m bằng cặp stereo camera với đường cơ sở 1,76 m. Mô hình Active Appearance định vị các dặc trưng

khuôn mặt (facial landmark) từ từng góc nhìn và triangulation tạo ra các vị trí dặc trưng 3D (3D landmark positions). Các tác giả có thể đạt được tỷ lệ nhận dạng 100% ở 15 m, mặc dù kích cỡ bộ sưu tập là 30 đối tượng và dữ liệu thu thập có hợp tác và kiểm soát. Cần lưu ý rằng thông tin chiều sâu ở khoảng cách xa như vậy với baseline có thể bị nhiễu và không ảnh hưởng đáng kể vào độ chính xác nhận dạng.

Redman và các đồng nghiệp tại Lockheed Martin Coherent Technologies phát triển một hệ thống chụp ảnh khuôn mặt 3D cho sinh trắc học bằng cách sử dụng Fourier Transform Profilometry. Điều này liên quan đến việc chiếu một mẫu rìa có hình sin (sinusoidal fringe pattern) lên đối tượng khuôn mặt sử dụng các chùm tia laser bước sóng ngắn an toàn cho mắt và chụp ảnh đối tượng được chiếu sáng bằng một camera được đặt lệch về phía sau từ nguồn sáng. Phương pháp này được xếp vào giải pháp stuctured light nhưng với yêu cầu baseline nhỏ. Một thử nghiệm hệ thống hiện tại chụp ảnh khuôn mặt 3D ở khoảng cách chủ thể 20 m với phạm vi sai lệch chuẩn (standard deviation) khoảng 0,5 mm và khoảng cách baseline chỉ 1,1 m.

Andersen đã phát triển hệ thống radar 3D laser và ứng dụng đẻ bắt ảnh khuôn mặt 3D. Ưu thế của hệ thống này là nó chạy với thời gian thực. Một ảnh 3D chụp chỉ mất vài giây. Trong đó có 50-100 hệ số phản xạ của ảnh thu được để tạo ra ảnh 3D. Tuy nhiên, hệ thống này lại có những hạn chế như nhiễu loạn khí quyển, camera bị rung và lỗi hệ thống.

Phương pháp sử dụng Fourier Transform Profilometry và biến đổi hình học kỹ thuật số (Digital Holography) được dùng để giải quyết bài toán ở khoảng cách xa nhưng nó vẫn phải kết hợp với camera WFOV để có thể bắt được ảnh 3D mặt người trên khu có diện tích rộng.

4.5. Face and Gait Fusion

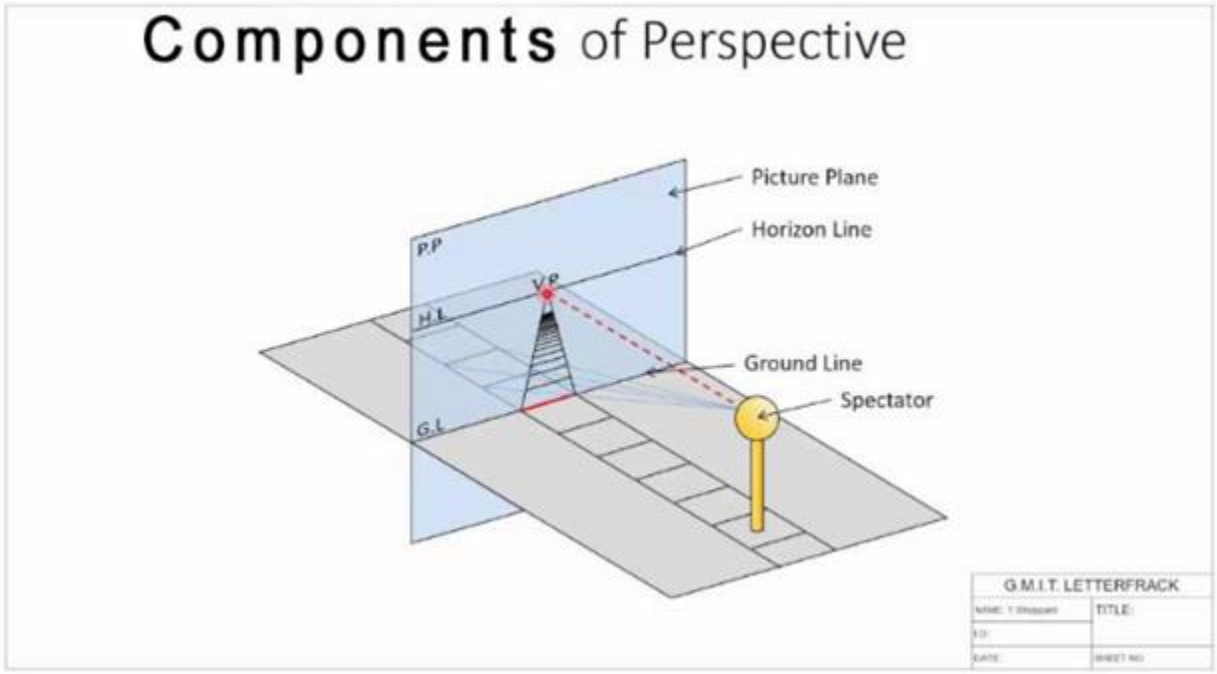
Liu đã kết hợp thuật toán nhận dạng bằng cách dùng phân tích mặt người và dáng để cải thiện hiệu quả nhận dạng. Với ảnh mặt người dược thu thập ở khoảng cách gần ngoài trời.
Zhou khá thành công trong bài toán nhận dạng kết hợp thông tin của mặt người và dáng người. Khởi thủy, bài toán nhận dạng mặt dựa vào các đặc trưng cong (curvature features) của khuôn mặt. Sau này thì còn sử dụng thêm cả phần bên hông khuôn mặt cũng như nâng độ phân giải của phần bên hông mặt bằng cách dùng nhiều khung hình có độ phân giải cao. Giúp thu được nhiều thông tin hơn.

Figure 11: Hệ thống giám sát sinh trắc học (Biometric Surveillance System). Hệ thống được đặt trên chiếc xe đẩy với bên trên có 2 camera góc cao bên trên và 1 camera góc gần.

5. Face Capture at a Distance

Tìm hiểu **Biometric Surveilance System**. Những đặc trưng chính của hệ thống này bao gồm:

- Ground-plane tracking of subjects: Trong phép chiếu perspective, ground plane là mặt phẳng gắn với bề mặt trái đất . Kết quả khi theo vết một đối tượng bằng phương pháp này rất đáng tin cậy.



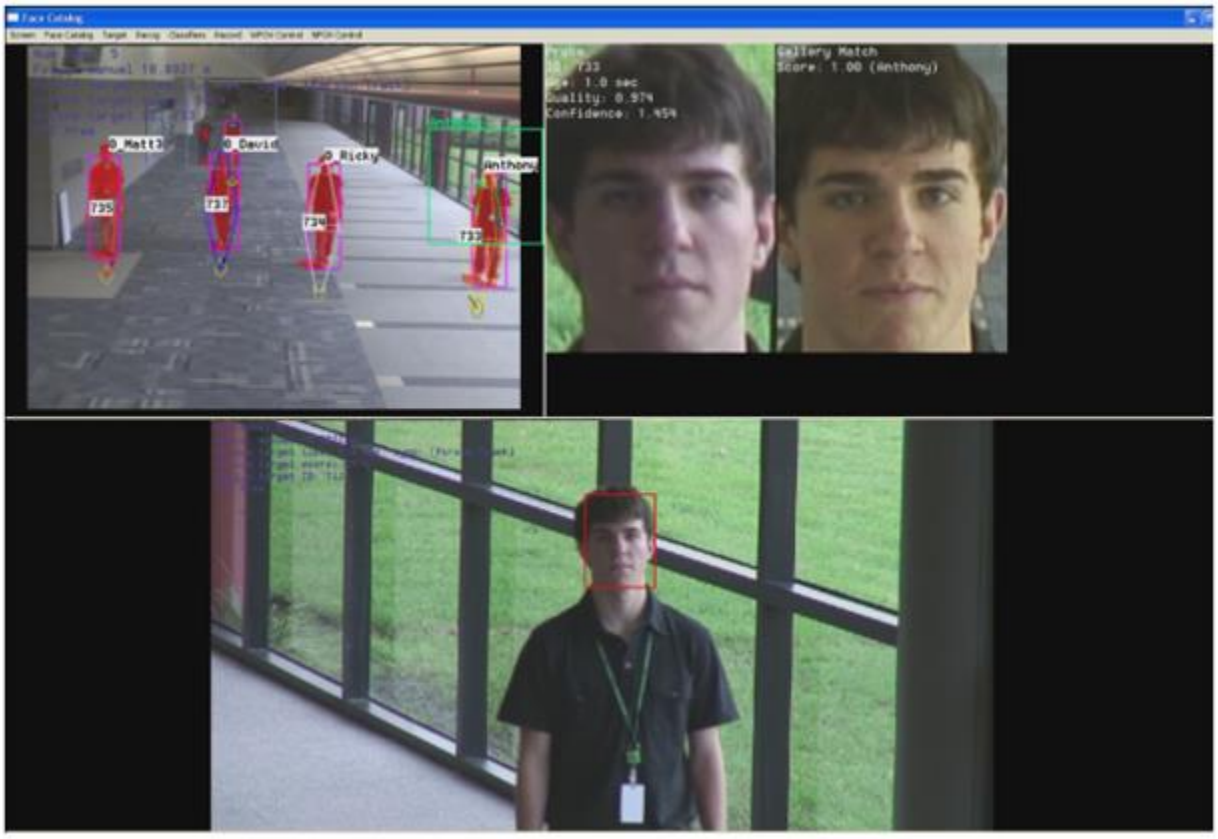
Hình. Ground plane là mặt phẳng màu xám

- Predictive targeting
- A target priority scoring system: Khi sử dụng WFOV camera thì sẽ detect và track được nhiều đối tượng, nhưng ảnh của các đối tượng này chỉ có độ phân giải thấp (low

resolution). Do đó, cần phải có một cơ chế tính độ ưu tiên để các NFOV PTZ camera (high-resolution camera) chọn được ảnh thích hợp trong số các ảnh ở trên. Độ ưu tiên của một đối tượng được tính dựa trên các hành động trong quá khứ và lẫn hiện tại của họ. Với mỗi đối tượng "tracked" được, sẽ có một *target record* (hồ sơ ghi chép) cho đối tượng đó. Target record bao gồm:…

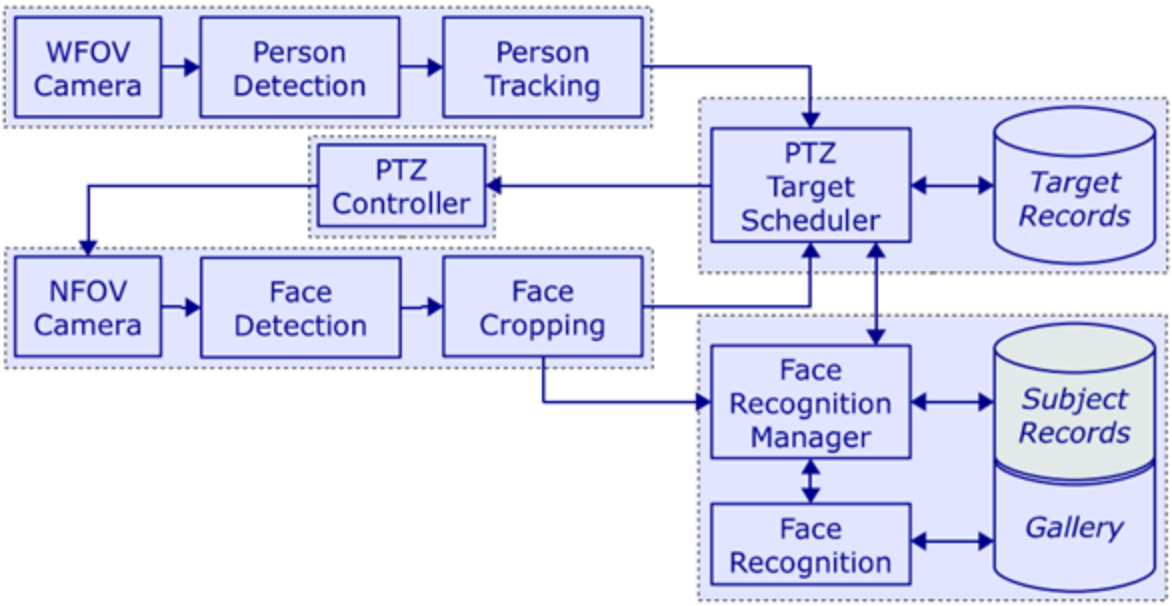
- Many configurable operating modes: hệ thống này có nhiều chế độ vận hành để người dùng có thể cấu hình được, bao gồm:
  + Cơ chế auto-enrollment:
  + Network-based sharing of auto-enrollment data for re-identification.

Các thông tin về tracking, target scoring, target selection, target status, các lần cố gắng nhận dạng nhưng không thành công (attempted recognition), các lần nhận dạng thành công (successful recognition) và các enrollment đều được hiển thị trên giao diện người dùng.



Hình.Các thông tin hiển thị trên giao diện (ở góc trên trái)

Quy trình của Biometric Surveilance System:



Hình. Lược đồ thể hiện các thành phần tính toán chính của Biometirc Surveilance System.

5.1. Target selection (Target Scheduler)

Khi có nhiều người hiện diện, hệ thống cần phải xác định xem người nào được chọn để đưa đến bước high-resolution face capture. Từ ảnh có được từ WFOV camera, có thể dễ dàng xác định 3 đại lượng: khoảng cách từ camera đến đối tượng, góc hợp bởi hướng chiếu của camera và hướng nhìn của đối tượng, tốc độ di chuyển của đối tượng.

Một bản ghi chép (record) sẽ được lưu cho mỗi đối tượng “tracked” được. Mỗi bản ghi chép (cho mỗi đối tượng) sẽ lưu trữ 3 thông tin: số lượng lần “target” đối tượng, số lượng lần lấy thành công ảnh khuôn mặt, số lượng lần nhận dạng thành công. Các thông tin này sẽ được sử dụng trong phần target selection.

Giữa các đối tượng “tracked” được, cần có một cơ chế tính điểm ưu tiên (priority scoring mechanism) để chỉ chọn ra một số các đối tượng đưa vào giai đoạn high-resolution facial recognition (sử dụng NFOV camera). Đối tượng có điểm số cao nhất sẽ là mục tiêu theo của các camera này. Mỗi tham số (parameters) dùng trong quá trình tính điểm này sẽ có một hệ số nhân và giá trị của chúng sẽ được đưa về trong một khoảng xác định trước.

Parameter	Factor	Clipping Range
Direction cosine	10	[−8, 8]
Speed (m/s)	10	[0, 20]
Capture attempts	−2	[−5, 0]
Face captures	−1	[−5, 0]
Times recognized	−5	[−15, 0]

Hình. Các tham số dùng để tính điểm cùng với hệ số nhân và khoảng giá trị xác định trước.

Toàn bộ quá trình xác định mục tiêu bằng các tham số trong bảng trên sẽ hướng đến những đối tượng di chuyển nhanh và chưa được chụp hình lại nhiều lần (bao gồm cả chụp thành công và chụp thất bại). Các đối tượng đã có nhiều ảnh khuôn mặt được chụp lại thì sẽ bị giảm mức độ ưu tiên.

Khi một đối tượng đã được chọn, hệ thống dùng Kalman filter để dự đoán vùng khuôn mặt của đối tượng này trong 0.5 – 1 s tới. Khoảng thời gian này gọi là **target time**. Camera NFOV sẽ hướng đến vị trí khuôn mặt cho đến khi **target time** đã trôi qua hết. Nhờ vào khoảng thời gian này, camera sẽ hoàn thành được các bước pan - tilt và các rung động cũng sẽ giảm lại.

Có một sự trade-off giữa hệ số zoom và xác suất chụp thành công ảnh khuôn mặt. Để thích ứng với vấn đề này, hệ thống sử dụng phương pháp: Nếu một đối tượng chưa có ảnh khuôn mặt nào được chụp thì goal ban đầu cho độ phân giải của ảnh khuôn mặt chỉ cần là 30 pixels eye-to-eye. Mỗi khi một resolution goal được vượt qua, thì resolution goal mới sẽ là resolution goal cũ tăng lên thêm 20%. Do mỗi đối tượng thường sẽ được chụp hình nhiều lần nên resolution goal sẽ tăng lên rất nhanh. Đối với một đối tượng bất kỳ, resolution goal và khoảng cách của đối tượng sẽ quyết định hệ số zoom của NFOV camera.

NFOV camera có 2 chế độ focus: thường (manual) và tự động (automatic). Khoảng cách của đối tượng cũng được dùng để set focus distance của NFOV camera. Automatic focus hoạt động tốt, nhưng thỉnh thoảng nó tập trung lên những vật gần hoặc xa hơn đối tượng được nhắm đến (targeted subject), làm đối tượng bị mờ (out of focus).

5.2. Recognition

Video có được từ NFOV camera được xử lý theo từng frame. Sau đó dùng Pittsburgh Pattern Recognition FT SDK để detect khuôn mặt trên từng frame. Nếu detect được từ 2 khuôn mặt trở lên trong 1 frame, nên chọn khuôn mặt nào nằm ở gần chính giữa frame đó nhất. Tiếp đến, crop phần khuôn mặt đó từ frame và đưa nó đến Face Recognition Manager. Thông tin của ảnh khuôn mặt này cũng được đưa đến Target Scheduler để record của đối tượng đang xét được update.

Khi Face Recognition Manager nhận được một ảnh khuôn mặt mới, một bản ghi (record) sẽ được tạo ra và ảnh đó sẽ được lưu lại. Tốc độ chụp ảnh nhanh hơn tốc độ nhận dạng nên các thuật toán nhận diện khuôn mặt được thực hiện một cách không đồng bộ. Các ảnh khuôn mặt chụp được sẽ liên tục được tiến hành nhận dạng và lưu vào facial image capture record.
Quá trình nhận dạng khuôn mặt có thể dùng gallery of images, manual enrollments, automatic enrollment hay kết hợp chúng lại với nhau.

Face Recognition Manager sẽ nhờ Target Scheduler để xác định subject ID của một ảnh khuôn mặt, dựa trên thời gian chụp của bước ảnh. Từ những thông tin này, các bản ghi cho các đối tượng (subject records) sẽ được tạo ra và chúng cũng được liên kết với facial image capture records.

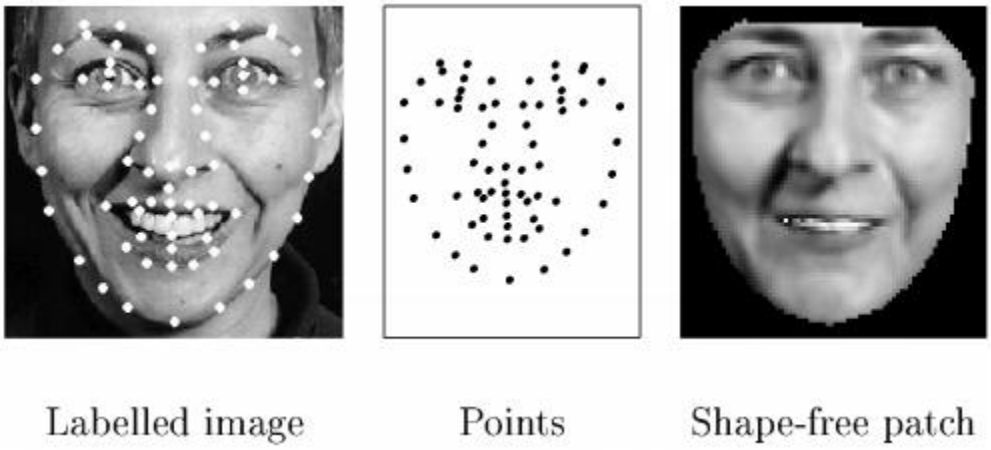
Cơ chế auto-enrollment sẽ tận dụng những subject record này để tự động “enroll” đối tượng khi biết chắc đối tượng đó chưa được enrolled và có ảnh khuôn mật phù hợp. Đối tượng được chọn phải có ít nhất một ảnh khuôn mặt có quality score vượt ngưỡng và quá trình nhận dạng sử dụng gallery (đã được cập nhật tính đến hiện tại) phải có ít nhất 4 lần nhận dạng thất bại, mỗi lần là một ảnh khuôn mặt khác nhau. Hai điều kiện trên giúp đảm bảo rằng đối tượng được chọn là chưa nhận diện được. Ảnh khuôn mặt được chọn phải xuất hiện ít nhất 4 giây trước. Tiêu chuẩn này để ngăn ngừa sự lựa chọn những đối tượng vẫn còn trong tầm nhìn. Đối với một đối tượng được chọn cho quá trình auto-enrollment, thì ảnh khuôn mặt tương ứng của đối tượng đó được đưa vào face recognition gallery sẽ là ảnh có quality score cao nhất.
Hiệu quả của hệ thống này trong thí nghiệm cho thấy: khoảng cách trung bình của lần phát hiện người đầu tiên (initial person detection) là 37 m, khoảng cách trung bình để capture được ảnh khuôn mặt đầu tiên (initial facial capture) là 34 m, khoảng cách trung bình để nhận diện là 17 m.

6. Low-Resolution Facial Model Fitting

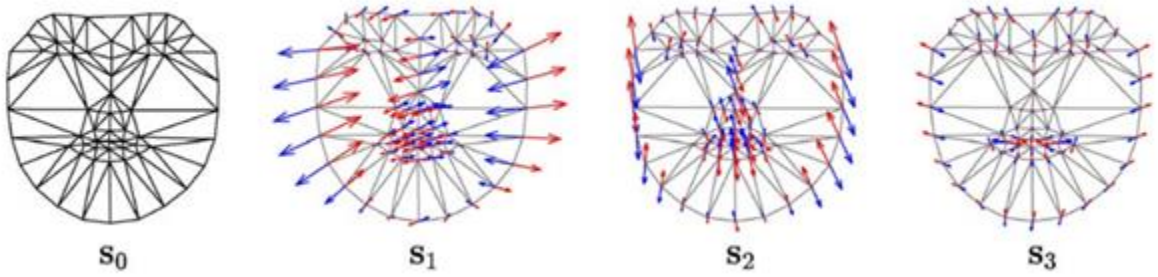
Mô tả: Face alignment (căn chỉnh khuôn mặt) là một phương pháp sử dụng một template (mô hình mẫu cho khuôn mặt) có thể biến dạng được để áp lên ảnh khuôn mặt và giúp lấy ra được các đặc trưng chính của khuôn mặt trong ảnh đó. Face alignment có thể được dùng trong rất nhiều các bài toán khác nhau như: face recognition, expression analysis (phân tích cảm xúc), face tracking,… Một trong những phương pháp face alignment phổ biến nhất là các Active Appearance Model (AAM). Đa số những nghiên cứu trước đây là chỉ fit AAMs lên các ảnh khuôn mặt có chất lượng tương đối tốt (tốt ở đây có nghĩa là có độ phân giải cao). Nhưng do nhu cầu cho bài toán FRAD này, chúng ta cần phải tìm ra cách để fit được AAMs lên cả các ảnh khuôn mặt có độ phân giải thấp.

Có một vài giải pháp đã được đề xuất để giải quyết vấn đề trên như: multi-resolution Active Shape Model, tích hợp công đoạn chỉnh đổi ảnh (image formulation process) vào quá trình fit

AAM. Trong quá trình fitting, các tham số cho image formulation và các tham số cho mô hình được sử dụng trong một framework thống nhất. Các tác giả cũng đã chỉ ra rằng multi-resolution AAM (sử dụng nhiều ảnh có độ phân giải khác nhau) tốt hơn rất nhiều high-resolution AAM (chỉ sử dụng 1 ảnh duy nhất có độ phân giải cao).



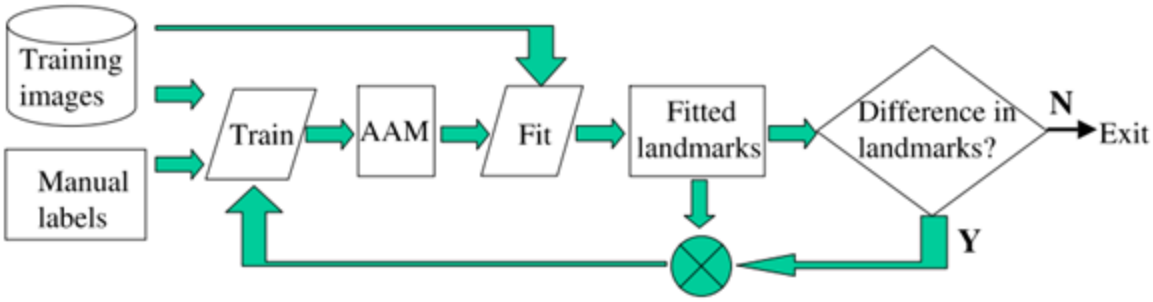
Hình. Từ ảnh input đầu vào, có được tập các point trên vùng khuôn mặt và cũng lấy ra được vùng có khuôn mặt như trong Shape – free patch



Hình. Shape model của AAM. Bao gồm base mesh s0 gồm các tam giác. S1, s2 và s3 là các shape vector

6.1. Face Model Enhancement

Một yêu cầu cho quá trình training AAM là chúng ta phải tự đặt facial landmark cho các ảnh training. Việc này thật sự rất tốn thời gian và có thể có lỗi (do người đặt sai). Lỗi sẽ gây ảnh hưởng xấu đến việc face modelling. Để vượt qua được lỗi, một quy trình cải thiện AAM đã được sử dụng.



Hình. : Sơ đồ quy trình cải thiện AAM

Giải thích quy trình: Input là các ảnh training và các label đặt sẵn, AAM sau đó sẽ được trained. Tiếp tục, AAM vừa mới học được sẽ được fit lại lên các ảnh training bằng cách sử dụng thuật toán Simultaneous Inverse Compositional (SIC), lúc này vị trí facial landmark ban đầu cho fitting sẽ là vị trí trong manual labels. Và tập các facial landmark mới sẽ tiếp tục được sử dụng tiếp cho quá trình face modelling, quá trình face modelling mới này sẽ sử dụng AAM đã được trained từ iteration trước đó. Cứ tiếp tục như vậy cho đến khi vị trí facial landmark trong 2 lần iteration kế nhau không còn khác biệt nhau quá lớn.

Việc cải thiện này cũng mang thêm một lợi ích nữa là tăng thêm sử nhỏ gọn (compactness) cho face model. Theo kết quả từ thí nghiệm, appearance và shape basis vector giảm từ 220 và 50 xuống còn 173 và 14. Hai lợi ích của một AAM nhỏ gọn hơn là ít thông số (parameter) phải sử dụng hơn trong model fitting và quá trình model fitting cũng diễn ra nhanh hơn.

6.2. Multi-Resolution AAM

Các phương pháp AAM ban đầu thường được trained bằng các ảnh “full-resolution” trong image dataset. Full resolution (hay còn gọi là high resolution) ở đây nghĩa là AAM được trained đúng với kích thước của ảnh input. Khi fit các AAM kiểu này vào ảnh có độ phân giải thấp, thì cần phải có thêm một bước upsampling cho các ảnh này . Điều này có thể gây ra vấn đề là high-resolution AAM có những component mà ảnh low resolution của khuôn mặt không có.

Ý tưởng của multi-resolution AAM: cho trước một tập các ảnh ban đầu, ta sẽ down-sample chúng xuống để có được các ảnh này nhưng với độ phân giải thấp hơn (các độ phân giải thấp hơn này sẽ được xác định trước). Sau đó tại mỗi độ phân giải, train AAM với các ảnh down-sample thì ta sẽ tạo ra được “pyramid of AAMs”.

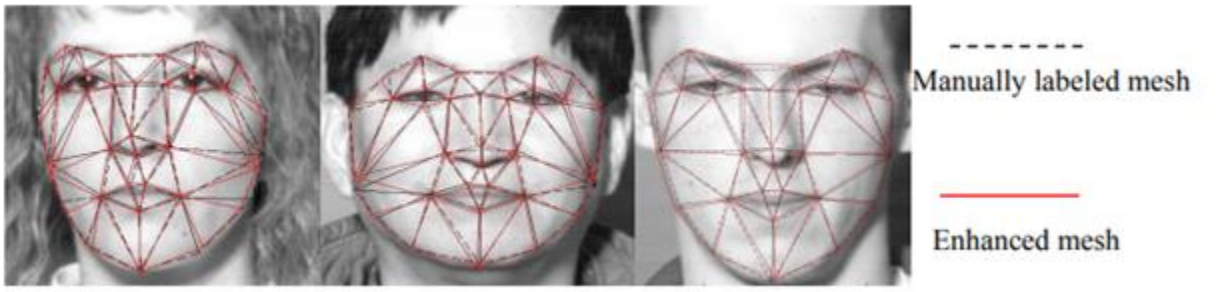
Ảnh AAM ở độ phân giải thấp hơn sẽ mờ (blurred) hơn và có ít basis vector hơn so với ảnh AAM ở độ phân giải cao. Những landmark được sử dụng trong train AAM cho các ảnh có độ phân giải cao, được lấy từ quy trình cải thiện AAM ở trên.

6.3. Experiments

Thí nghiệm trên ND1 face database gồm 534 ảnh của 200 đối tượng. Để đo mức độ thể hiện (performance), tác giả tính convergence rate (CR) theo các mức độ nhiễu loạn (perturbation) khác nhau của các vị trị landmark ban đầu. Việc fitting được xem là hội tụ (converged) nếu average mean squared error giữa landmark tìm được và groundtruth bé hơn một ngưỡng cho trước.

Một điều nữa cần quan tâm đó là số lượng ảnh/ đối tượng trong training set. Khi nhiều ảnh của một đối tượng được dùng để train AAM thì AAM đó sẽ được xem là person-specific, tức là AAM được train được dùng để nhận diện một đối tượng riêng biệt. Khi số lượng đối tượng trong dataset lớn, thì AAM được xem là generic, nghĩa là có mức độ tổng quát hơn và có thể được dùng đề nhận diện nhiều đối tượng khác nhau. Càng nhiều đối tượng trong dataset thì AAM càng generic.

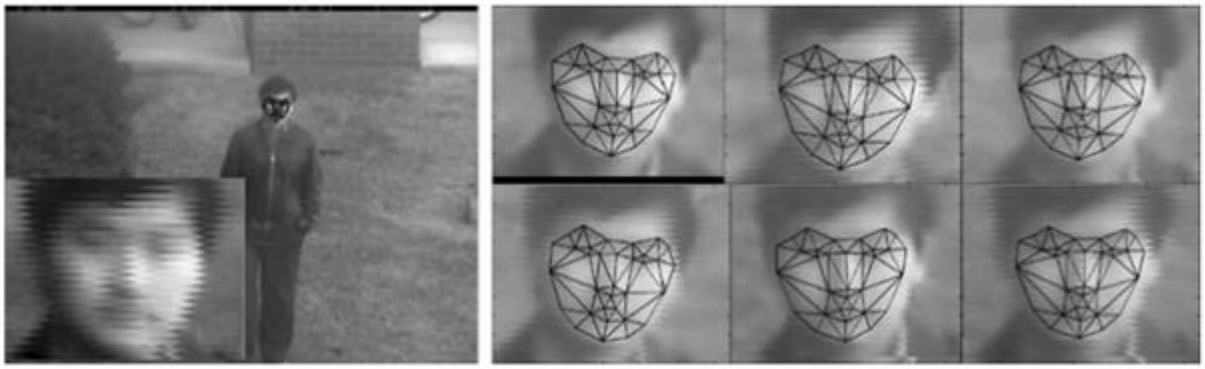
AAM đã qua giai đoạn model enhancement có performance tốt hơn hẳn so với khi chỉ được train bằng manual label. Sau khi model enhancement hoàn thành, quá trình fitting gần như hội tụ cho mọi test case.



Hình. So sánh giữa manual label và enhanced landmarks

Thí nghiệm thứ hai là để kiểm tra fitting performance của multi-res AAM. Dataset trên cũng được sử dụng tiếp, thêm vào đó là các ảnh down-sampled trong training set cũng được sử dụng khi test. Quá trình model fitting được diễn ra với tất cả sự kết hợp giữa độ phân giải của AAM và độ phân giải của ảnh. Đối với những ảnh low-resolution, fitting performance tốt nhất khi model resolution cao hơn một chút so với facial image resolution.

Thí nghiệm cuối cùng là model fitting trên đoạn video thu được từ PTZ camera đặt cách xa đối tượng 20m.



Hình. Kết quả fitting multi-res AAM lên một đoạn video được quay ngoài trời.

Mặc dù kích thước của frame là 480x640 (tương đối chi tiết) nhưng vùng có chứa khuôn mặt có độ phân giải rất nhỏ và còn bị blur rất nhiều. Điều này là một thách thức rất lớn. Multi-res AAM fit được gần khoảng 100 frame liên tiếp và cho ra kết quả rất khả quan. Ngược lại, high-resolution AAM chỉ fit được 4 frame đầu tiên.

7. Facial Image Super-Resolution

Độ phân giải của các ảnh quá thấp có thể ảnh hưởng không tốt đến việc nhận dạng, một giải pháp đối mặt với vấn đề này là multi-frame image super-resolution. Khi sử dùng camera, trong một khoảng thời gian ngắn có thể thu được liên tiếp nhiều ảnh của cùng một đối tượng. Và ta sẽ tận dụng tất cả các ảnh này để cải thiện việc nhận dạng. Một trong những phương pháp thực hiện điều này là super-resolution.

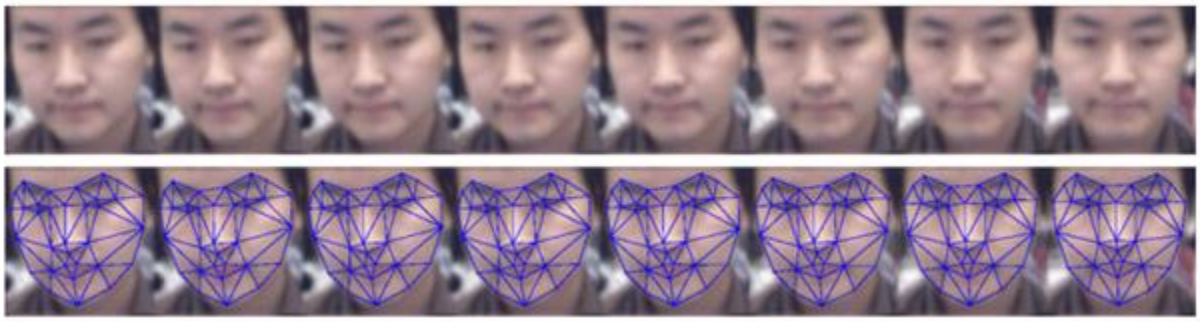
Super-resolution là quá trình tạo ra một ảnh high-resolution từ nhiều ảnh low – resolution của cùng một đối tượng. Việc cải thiện độ phân giải sẽ giúp khử răng cưa (dealiasing), khử mờ (deblurring) và khử nhiễu (noise reduction). Ý nghĩa của super-resolution là việc trích xuất ra được các sub-pixels từ các frame khác nhau của cùng một đối tượng. Các sub-pixels trong một ảnh sẽ có chứa các thông tin mới đối với các ảnh khác và từ đó có thể tạo nên được ảnh super-resolution.

Super-resolution thường gồm 2 phần: frame-to-frame registration và super-resolution image processing. Trong đề xuất này, tác giả sử dụng AAM cho registration. Còn về super-resolution image processing, tác giả sử dụng L1 fidelity component và Bilateral Total Variation regularization term.

7.1. Registration and Super-Resolution

Cho trước một chuỗi các ảnh đã được fit AAM của cùng 1 đối tượng. Hình dưới mô tả 33 vị trí landmark, đồng thời cũng là đỉnh của 49 tam giác. Việc registration khuôn mặt của 2 frame thực

chất là một phép biến đổi affine, các phép affine này sẽ tác động lên mỗi tam giác thông qua các đỉnh tương ứng của tam giác đó. Tập khoảng 10 frame kế tiếp nhau được kết hợp lại để tạo thành super-resolved image.



Hình. Fit AAM model trên một chuỗi các frame liên tiếp

Mô tả thuật toán super-resolution: xem các mỗi bức ảnh như là một vector chứa các pixel values. Quá trình super-resolution sử dụng các image formation model để liên kết các frame Y_i đến một ảnh super-resolution X. Quá trình image formation sẽ chịu trách nhiệm cho face motion, camera blur và detector sampling. Đối với mỗi frame, F_i là một registration operator để "bẻ cong" (warp) ảnh X để căn chỉnh nó với các frame Y_i . Blur operator H sử dụng Point Spread Function (PSF). Thường thì để xác định PSF của một camera là rất khó nên chúng ta sẽ giả định dùng Gaussian PSF. Cuối cùng là D, ma trận đại diện cho việc sampling, nhiệm vụ của nó là rút trích các pixel khác trong mỗi chiều để tạo ra được một ảnh giống với một ảnh thực mà chúng ta quan sát. Công thức cho toàn bộ quá trình image formation.

$$\hat{X} = \operatorname{argmin}_X \left[\sum_{i=1}^{N} \| DHF_i X - Y_i \|_1 + \lambda \Psi(X) \right].$$

Hình. Cost function

Về đại lượng regularization, ta sử dụng Bilateral Total Variation(BTV)

$$\Psi(X) = \sum_{l=-P}^{P} \sum_{m=-P}^{P} \alpha^{|m|+|l|} \left\| X - S_x^l S_y^m X \right\|_1.$$

Hình. Regularization term

S_{xl} và S_{ym} để shift ảnh theo hướng x và y một đoạn l và m pixel. Hệ số giảm alpha nằm trong khoảng (0,1). Regularization theo L_1 như ở đây là để giữ lại được các biên cạnh. Còn theo L_2 là cho làm trơn ảnh (smoothness).

Để tìm ảnh super-resolution, X được khởi tạo ban đầu bằng cách warping. Steepest descent search với gradient của cost function cũng được sử dụng. Đối với các frame được chuẩn hóa về khoảng pixel [0,1], tác giả nhận thấy rằng tham số lambda = 1 sẽ cho kết quả tốt nhất.

7.2. Results

Hình. Ảnh ví dụ khi thực hiện super resolution

Ảnh trên ví dụ kết quả sau khi thực hiện super-resolution. Cột 1 là ảnh gốc, cột 2 là ảnh sau khi đã qua Wiener filter, cột 3 là ảnh super-resolution sử dụng 10 frame liên tiếp. Ta thấy được ảnh mới sắc nét và rõ rành hơn. Trong một thí nghiệm khác, tác giả cũng tính được tỉ lệ chính xác trong nhận dạng tăng lên từ 50% lên 56% khi sử dụng super resolution.

8. Tổng kết

Nhận diện mặt người từ khoảng cách xa là một bài toán đầy thách thức với nhiều ứng dụng thực tiễn. Trong bài báo cáo này cũng đã có nhắc đến các thách thức chính, các hướng tiếp cận và cũng có mô tả sơ qua một vài phương pháp để tạo ra một hệ thống FRAD như: giới thiệu Biometric Surveilance System, cách fit một face model như AAM lên một ảnh khuôn mặt, Multi-resolution AAM và cách tạo ảnh bằng phương pháp Super-resolution,…. Tuy nhiên, vẫn còn rất nhiều vấn đề cần phải giải quyết trong bài toán này.

Các hệ thống nhận diện khuôn mặt hiện nay thường chỉ xử lý tốt khi có một tập ảnh chất lượng (độ phân giải cao; độ chiếu sáng thích hợp; tránh được noise, ngăn được độ rung của camera,…). Điều này đồng nghĩa với yêu cầu của phần cứng phải rất cao và tất nhiên giá thành cho những phần cứng này cũng cao ngất ngưỡng, dẫn đến các hệ thống này không thể được sử dụng phổ biến ở nhiều nơi được. Ngoài ra vẫn còn nhiều thách thức mà ngay cả sức mạnh của phần cứng cũng không thể giải quyết được. Vấn đề đặt ra là phải có một hệ thống FRAD hoạt động tốt và nhanh dưới nhiều điều kiện khác nhau.

Đối với thách thức về cử chỉ của đối tượng, một hướng tiếp cận khác từ bên ngoài việc phát triển hệ thống này là làm sao tìm ra cách nào đó thu hút được sự chú ý của đối tượng về phía camera. Trái ngược với WFOV và NFOV là hệ thống camera tĩnh, ta có thể sử dụng các bệ camera có thể di chuyển (moving camera platform), robot,… Đây là một số hướng tiếp cận khác để cải thiện hiệu quả của bài toán FRAD.

9. Demo:

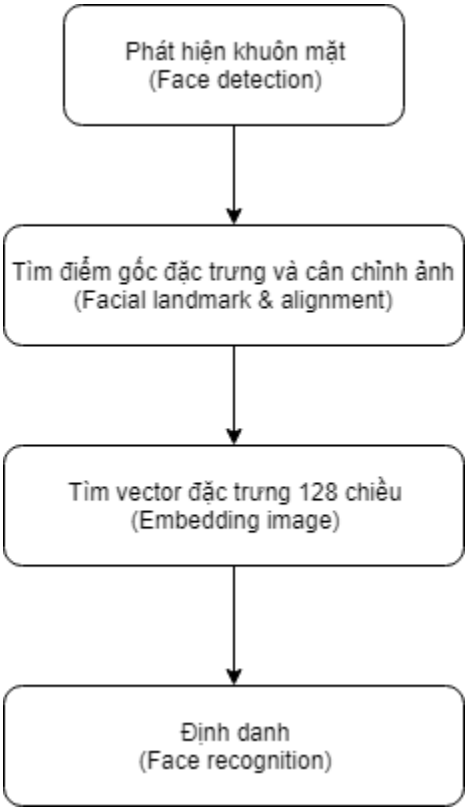


Figure 12: Các công việc liên quan đến hệ thống nhận dạng

Bài toán chỉ gồm 4 bước sau:

1) Phát hiện khuôn mặt (Face detection – FD): chúng tôi dùng mô hình mạng Single Shot Detector (SSD) với kiến trúc mạng cơ sở ResNet để đạt được kết quả chính xác cao trong phát hiện mặt người, đồng thời có thể chạy với thời gian thực.

2) Tìm điểm gốc đặc trưng trên khuôn mặt và cân chỉnh ảnh (Facial Landmark & alignment): Tập trung vào từng khuôn mặt và có thể hiểu rằng ngay cả khi một khuôn mặt bị quay theo hướng lạ hoặc trong điều kiện ánh sáng xấu, cần phải cân chỉnh bằng các phép biến đổi AFFINE vì nó vẫn là cùng một người.

- Facial landmark là 68 điểm đặc trưng trên khuôn mặt của một người bao gồm các điểm trải dài từ chân mày, hai bên hàm đến cằm dưới, các điểm dọc theo đường sóng mũi, vùng quanh mắt và miệng. Chúng ta sẽ dùng thuật toán Machine Learning để tìm 68 điểm này. Nhóm sẽ thực hiện cải tiến một chút trong việc tìm 5 điểm đặc trưng thay vì 68 điểm như trước dựa vào Davis King đề xuất – tác giả của thư viện Dlib.
- Nhưng lúc này có thể khuôn mặt nghiêng theo hướng nào đó, cần một bước cân chỉnh bằng dùng các phép AFFINE cơ bản: xoay, scale và shear mắt và miệng ở giữa hình nhiều nhất có thể. Bước này đóng góp một phần vào độ chính xác ở giai đoạn sau.

3) Tìm vector đặc trưng 128 chiều (Embedding image): Giai đoạn này giúp tìm ra các vector đặc trưng để tách bạch các lớp của dữ liệu và xác định danh tính của người đó chính xác hơn. Ở bước này, các chuyên gia đã tìm ra cách dùng mạng Deep learning để tìm ra các Embedding image này. Mô hình này được huấn luyện theo cách thức single triplet. Nghĩa là, dữ liệu huấn luyện sẽ gồm dữ liệu huấn luyện của một người, một ảnh khác của cùng người này và ảnh còn lại là của một người khác. Mô hình học sẽ điều chỉnh tham số bên trong để khiến độ đo của 2 ảnh đầu gần nhau và khác xa so với ảnh cuối. Sau đó, lặp lại quá trình huấn luyện này triệu lần trên hàng triệu ảnh huấn luyện để khởi sinh ra các vector đặc trưng và các vector đặc trưng này là duy nhất cho từng đối tượng. Mô hình học này được gọi là deep metric learning và được hỗ trợ trong thư viên Dlib và OpenCV.
4) Định danh (Face recognition): Khi ta đã có được các vector đặc trưng ở giai đoạn 3 thì ta sẽ dùng thuật toán phân lớp cơ bản Support Vector Machine (SVM) để phân lớp dữ liệu và theo khảo sát thì thuật toán này mang lại hiệu quả rất tốt. Ở bước này làm nhiều vụ tìm ảnh có độ đo gần nhất match với tập huấn luyện và dựa vào đó xác định danh tính của người đó.

Demo phục vụ: biểu diễn một hệ thống nhận dạng mặt người là như thế nào và đưa ra các giải pháp tiếp cận cũng như cái được và mất cũng với cải tiến mà nhóm đã tìm hiểu được. Đồng thời minh chứng cho việc một hệ thống nhận dạng mặt người hoàn toàn khả thi khi thực hiện ở thế giới thực nhằm vào nhiều mục đích khác nhau như bảo mật, an ninh, truy tìm tội phạm,… với độ chính xác gần như hơn cả người.

10. Tham khảo

Wheeler, F.W., Weiss, R.L., Tu, P.H.: Face recognition at a distance system for surveillance applications. In: BTAS (2010)
Cootes, T., Edwards, G., Taylor, C.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. 23(6), 681–685 (2001)
Liu, X., Tu, P.H., Wheeler, F.W.: Face model fitting on low resolution images. In: BMVC (2006)