PHÁT HIỆN VÀ TRUY VẾT ĐỐI TƯỢNG KHẢ NGHI SỬ DỤNG KỸ THUẬT HỌC SÂU

DETECTING AND TRACKING SUSPICIOUS HUMAN BASED ON DEEP LEARNING METHODS

SVTH: Nguyễn Nghĩa Thịnh

Lớp 18T1, Khoa Công nghệ Thông tin, Trường Đại học Bách Khoa – Đại học Đà Nẵng;
Email: nghiathinh2000@gmail.com

GVHD: TS. Ninh Khánh Duy

Khoa Công nghệ Thông tin, Trường Đại học Bách Khoa – Đại học Đà Nẵng; Email: nkduy@dut.udn.vn

Tóm tắt

Theo dõi đối tượng khả nghi là một trong những bài toán cực kỳ quan trọng trong cuộc sống khi nhiều vụ trộm cướp, các tội phạm truy nã ngày càng tăng gây nguy hiểm cho xã hội. Ngày nay, nhiều công nghệ tích hợp trí tuệ nhân tạo, thị giác máy tính cho phép theo dõi các đối tượng theo thời gian thực và đưa ra những kết quả chính xác về thông tin được trích xuất từ các camera công cộng. Trong nghiên cứu này, đề xuất phương pháp theo dõi đối tượng khả nghi hay cụ thể hơn là đối tượng đang bị truy nã bằng cách kết hợp các mô hình học máy, đồng thời đưa ra một giải pháp cải thiện kết quả truy vết đối tượng theo thời gian thực.

Từ khóa – Human Tracking; Deep learning; Machine Learning; Object Detection; Face Recognition; Human Pose.

Abstract

Tracking suspicious human is one of life's most pressing issues, especially as the number of robberies and wanted criminals rises, harming society. Many technologies that combine artificial intelligence and computer vision now enable for real-time object monitoring and reliable results about information retrieved from public cameras. This paper proposes a method for tracking suspicious humans. In particular, wanted humans are tracked by combining machine learning models, and we also propose a strategy to improve tracking results in real time.

Key words – Human Tracking; Deep learning; Machine Learning; Object Detection; Face Recognition.

1. Giới thiệu

Những năm gần đây, tỷ lệ tội phạm ngày càng tăng cao, đặt biệt là tội phạm bị truy nã, các tội nhân trốn trại vẫn đang sống ngoài xã hội gây ra nguy hiểm cho người dân xung quanh. Nhiều trường hợp thực hiện truy nã trong thời gian dài nhưng vẫn không mang lại kết quả cao vì đối tượng thường lẫn trốn trong đám đông hoặc che dấu đi thân phận của mình. Hiện nay, các camera công cộng với chất lượng cao được đặt khắp nơi như ở chợ, siêu thị, trường học và điều này có thể một phần nào đó giúp truy tìm được các đối tượng một cách nhanh chóng.

Việc ứng dụng các thuật toán học máy và trí tuệ nhân tạo đã trở nên phổ biến trong rất nhiều lĩnh vực. Không nằm ngoài xu thế đó, áp dụng trí tuệ nhân tạo mà cụ thể là học sâu (deep learning) và mạng nơron (neural network) để giải quyết các vấn đề liên quan đến thị giác máy tính nói chung và theo dõi đối tượng khả nghi nói riêng là một bước đi tất yếu. Dữ liệu đa dạng từ các nguồn video và camera khác nhau giúp cho con người khám phá ra các phương pháp về truy vết (Tracking) cũng như nhận diện đối tượng (Object Detection) một cách nhanh chóng nhằm đưa ra giải pháp kịp thời mà không cần con người trực tiếp giám sát.

Hiện nay đã có một số phương pháp nhận diện đối tượng khả nghi dùng Deep learning, và hầu hết trong số chúng đều là dựa vào hành động của các đối tượng như cướp, giật, đánh nhau, ... Tuy nhiên, những trường hợp này chỉ đúng khi camrea được đặt ở nơi có ít người và rất khó để có thể phân biệt được có phải là hành động khả nghi hay không bởi vì cách tiếp cận của các đối tượng ngày càng tinh vi khi thường lẫn vào những nơi đông người và thực hiện trộm cắp. Và những bài toán như vậy thường áp dụng trong trường hợp chưa biết đối tượng là ai trước đó vì chỉ dựa vào hành động và đưa ra cảnh báo khi sự việc đã xảy ra. Trong nghiên cứu này, tác giải muốn tiếp cận bài toán theo một cách khác đó là dựa vào trích xuất khuôn mặt của đối tượng thay vì dựa vào hành động để phát hiện và đưa ra cảnh báo, cụ thể ở đây là các đối tượng đang bị truy nã đã được cơ quan chức năng cung cấp hình ảnh khuôn mặt từ trước. Ngoài ra bài toán có thể giải quyết được ở những nơi có nhiều người qua lại như là các siêu thị, chợ, trường học, bệnh viện. Bằng cách kết hợp nhiều mô hình Deep Learning lại với nhau nhưng vẫn đảm bảo được tốc độ xử lý theo thời gian thực và độ chính xác tương đối cao.

Những đóng góp chính trong nghiên cứu này là:

(1) Xây dựng luồng xử lý kết hợp các mô hình Deep Learning lại với nhau nhằm tăng độ chính xác cũng như tốc độ xử lý đối với việc truy vết đối tượng khả nghi.
(2) Một phương pháp cải thiện mô hình Human Tracking so với lại các mô hình hiện tại bằng cách ứng dụng đặc trưng (feature) của khung xương người (Humn Pose).
(3) Ứng dụng giải pháp "tham số hóa lại" (re-parameterization) cho các mô hình Deep Learning với mục đích tăng tốc độ dự đoán của các mô hình bằng cách giảm kích thước của các tham số nhưng độ chính xác vẫn không đổi quá nhiều.
(4) Xây dựng mô hình nhận diện khuôn mặt của người Châu Á.

Phần còn lại của báo cáo này được sắp xếp như sau. Phần 2 là các nghiên cứu liên quan. Phương pháp được đề xuất trong nghiên cứu được đề cập tại phần 3. Phần 4 thể hiện kết quả thực nghiệm của phương pháp. Phần 5 là thảo luận những vấn đề khó khăn chưa giải quyết . Cuối cùng, phần 6 tổng hợp về toàn bộ nghiên cứu.

2. Các nghiên cứu liên quan

2.1. Phát hiện đối tượng từ camera tĩnh dựa vào mảng màu.

(Akdemir, et al., 2008)đã trình bày một cách tiếp cận có hệ thống để nhận ra các hoạt động của con người trong ngân hàng và sân bay dựa trên bản thể học. Tác giả đã sử dụng năm tiêu chí để thiết kế sự rõ ràng, tính liên kết bản thể học, độ lệch mã hóa tối thiểu, khả năng mở rộng và cam kết bản thể học tối thiểu. Hệ thống đã được đánh giá trên sáu video ngân hàng, trong đó bốn video bao gồm cướp ngân hang và hai video bao gồm hoạt động bình thường của con người. Các đối tượng chuyển động đã được theo dõi bằng cách sử dụng màu sắc dựa trên sự xuất hiện và chuyển động. Trong điều này, mô hình phân loại chính xác ba tình huống cướp nhưng có một nhược điểm của hệ thống này là nó không thành công trên một video nếu có nhiều hơn hai tên cướp.

(Ibrahim, et al., 2012) đã trình bày quy trình quyết định có phải là đối tượng khả nghi hay không thông qua hai giai đoạn bằng cách trích xuất thông tin từ luồng quang học để phát hiện trộm cướp hoạt động bất thường từ video giám sát tự động. Trong bước đầu tiên, kết quả luồng quang học sàng lọc hiện trường để tìm hoạt động tội phạm tiềm ẩn. Sau khi phát hiện hiện trường có khả năng xảy ra tội phạm, giai đoạn thứ hai sử dụng thống kê mô hình dòng chảy để phân tích để quyết định xem một sự kiện trộm cắp có xảy ra hay không. Tuy nhiên kết quả dựa trên ánh sáng quang học không mang lại độ chính xác cao, và rất dễ bị nhầm lẫn.

2.2. Phát hiện đối tượng từ camera tĩnh dựa vào trích xuất đặc trưng từ mô hình Deep Learning

(Phyo, et al., 2019) bài báo đề xuất một mô hình trích xuất khung xương của đối tượng sau đó dựa vào đó để tiến hành phân loại hành động của đối tượng đó có phải là khả nghi hay không. Tuy nhiên bài tác giả chỉ giải quyết bài toán chỉ có một hoặc hai người trong khung hình. Và camera được đặt trước mặt đối tượng để thực hiện quan sát. Điều này thiếu tính thực tế khi áp dụng ở các camera công cộng thường đặt ở những nơi tương đối cao.

(Amrutha, et al., 2020) cũng đưa ra một giải pháp tương tự nhưng thay vì dựa vào khung xương thì tác giả xây dựng một mô hình học theo chuỗi các khung hình liên tiếp, từ đó thực hiện nhận biết đối tượng đó có những hành động khả nghi hay không. Tuy nhiên giải pháp này tốn khá nhiều tài nguyên và không chính xác khi hình vi diễn ra ở những nơi đông người và camera nằm ở trên cao, khi đó các đối tượng có thể bị che khuất hoặc các hành động phạm tội đó chỉ chiếm một phần nhỏ đặc trưng trong khung hình vì phần lớn những đối tượng trong khung hình đó đều hành động bình thường.

2.3. Phát hiện và theo dõi nhiều đối tượng dùng Deep Learning.

Nhiều phương pháp đã được đưa ra để theo dõi con người (Human Tracking) đặc biệt là các đám đông, nhiều người qua lại. Bài toán này tập trung giải quyết chủ yếu hai phần chính, đó là nhận diện con người (human detection) và định danh con người (re-id).

Đối với bài toán human detection, (K, et al., 2017), hoặc (Ren, et al., 2015) xây dựng cấu trúc mạng CNN dựa vào trích xuất đặc trưng của các vùng ROI (region of interest) để nhận diện con người. Tuy nhiên kết quả rất chậm khi mô hình phải trích xuất thông tin từ rất nhiều vùng ROI không chứa người. (Redmon, et al., 2015) mô hình YOLO là một trong những mô hình phổ biến nhất hiện tại về nhận diện vật thể nói chung và nhận diện con người nói riêng. Đến nay có đến bảy phiên bản, trong đó YOLOv7 (Wang, et al., 2022) mới nhất hiện tại đã đạt tốc độ xử lý cực kỳ nhanh với độ chính xác rất cao. Các mô hình này đều có điểm chung là sử dụng tọa độ mục tiêu (anchor box) đã xác định từ trước sau đó sẽ thực hiện cập nhập tọa độ trong quá trình huấn luyện để khớp với vị trí người trong khung ảnh nhất có thể. Tuy nhiên điều này sẽ tăng tham số tính toán, cũng như thiếu tính tổng quát của kích thước con người khi có thể thay đổi dựa vào kích thước thực tế của khung hình. Ngoài ra, mô hình chỉ lấy tọa độ của người có độ tự tin cao nhất và loại bỏ đi những tọa độ có độ tự tin thấp hơn. Điều này sẽ gây ra hiện tượng trùng lặp (overlap), việc định danh người có thể bị nhầm lẫn trong mô hình theo dõi. YOLOX (Ge, et al., 2021) là mô hình sử dụng mục tiêu tự do (free-anchor) thay cho anchor-box của các mô hình tiền nhiệm, mô hình thực hiện tính toán nhanh hơn vì ít tham số hơn, ngoài ra mô hình dựa vào tâm của người (center point) để xác định tọa độ của người. Trong trường hợp nhiều người đứng gần nhau, mô hình vẫn sẽ lấy center point của người bị che khuất, ngược lại với các mô hình tiền nhiệm. Trong bài báo này, tác giả sử dụng YOLOX là một giải pháp cho human detection.

(P. R. Gunjal, 2018) là một mô hình dùng để định danh người, mô hình dùng chuyển động để đưa ra dự đoán cho vị trí tiếp theo của người trên khung hình, từ đó dựa vào tọa độ người thực tế để cập nhập lại thông tin của người đó. (Zhang, et al., 2022) là mô hình áp dụng phương pháp này và đạt được độ chính cao. Một cách tiếp cận khác là sử dụng trích xuất đặc trưng của người trong khung hình để thực hiện định danh (Zhang, et al., 2020), và (Du, et al., 2022) đã đưa ra giải pháp tương tự, mỗi người trên khung hình sẽ có đặc trưng riêng và dựa vào đó để thực hiện phân loại, mô hình có tốc độ xử lý rất nhanh. (Aharon, et al., 2022) đã kết hợp cả hai phương pháp trên để xây dựng được một mô hình có độ chính xác rất cao và thời gian thực thi tương đối nhanh. Tác giả dựa trên nghiên cứu của bài báo này để đề xuất một giải pháp cho bài toán định danh con người

Hình 1. Sơ đồ xử lý

3. Phương pháp

3.1. Xây dựng hệ thống truy vết Deep Learning.

Tác giả đã kết hợp nhiều mô hình học máy khác nhau để xây dựng ra một hệ thống chung (Hình 1) nhằm giải quyết bài toán truy vết đối tượng khả nghi một cách có rõ ràng và tối ưu nhất có thể.

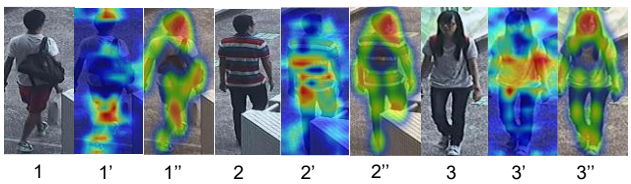
Các mô hình tác giả sử dụng gồm:

(1) Object Detection để nhận diện con người và đầu người
(2) Human pose để thực hiện trích cuất đặc trưng khung xương tối ưu cho bài toán Human Tracking
(3) Human Tracking thực hiện truy vết đối tượng là con người bằng cách sử dụng các mô hình nổi tiếng hiện tại như ByteTrack, BoTSORT.
(4) Face Recognition nhận diện đối tượng có phải là khả nghi hay không nhờ vào sự tương đồng của gương mặt của đối tượng đã được cho trước.

Dữ liệu đầu vào là các khung ảnh được lấy từ luồng camera và ảnh khuôn mặt đối tượng đang bị truy nã hoặc là đang bị theo dõi bởi cơ quan chức năng. Sau khi đi qua mô hình Object Detection để nhận diện được tọa độ của con người và đầu người. Tọa độ phần con người sẽ đi qua mô hình Human Tracking để thực hiện truy vết. Tại đây, những đối tượng có độ tự tin cao (> thresh) là xuất hiện trong khung hình sẽ đi qua mô hình Human Pose để thực hiện trích xuất đặc trưng khung xương người và xác định tọa độ điểm trung tâm của đầu người. Những đặc trưng cùng với những đối tượng có độ tự tin thấp sẽ được mô hình truy vết thực hiện tính toán để đưa ra danh sách các đối tượng đã được định danh trên khung hình (tracklets). Tọa độ đầu người (centroid_heads) đã được trích xuất sẽ được liên kết với tọa độ đầu người của tracklets thông qua bước lọc ảnh để xác định được những khuôn mặt nào là của tracklets nào. Phần khuôn mặt sau khi đã được lọc sẽ đi qua mô hình Face Recognition để so sánh độ tương đồng với ảnh thực tế của đối tượng đã được cung cấp từ trước, từ đó xác định được đối tượng nào trong tracklets là khả nghi để tiến hành cảnh báo và truy vết. Trong trường hợp, đã nhận diện được đối tượng khả nghi nhưng phần đầu của đối tượng không nhận diện được vì lý do quay đi chỗ khác hoặc bị ẩn sau một người khác thì hệ thống vẫn truy vết được đối tượng cho đến khi biến mất khỏi khung hình.

3.2. Trích xuất khung xương để định danh người.

Tác giả đã sử dụng đặc trưng khung xương để tiến hành Re-id, việc chọn đặc trưng từ các bộ phận của con người sẽ tốt hơn là đặc trưng của cả bức hình như cách mà các mô hình truy vết trước đó đã dùng, vì phần hình nền phía sau (background) thường sẽ thay đổi sau nhiều khung hình liên tiếp và gây ra nhiễu thông tin đặc trưng thứ thực chất chỉ dùng để định danh con người. Trong nhiều trường hợp, phần background của hai đối tượng gần giống nhau sẽ dẫn đến không phân biệt rõ ràng được 2 đối tượng



Hình 2. màu nhiệt biểu thị sự tập trung của các backbone là ResNet và HumanPose cho re-id

Trên Hình 2 có thể thấy các mô hình re-id (Zhang, et al., 2020) (các hình 1', 2' 3') sẽ tập trung vào một vài vị bức hình, tuy nhiên khi đi qua mô hình Human Pose của tác giả (hình 1'', 2'', 3'') thì những đặc trưng chỉ thực sự tập trung tại các vị trí của người như tay, chân, thân, ... Những đặc trưng này sẽ được cập nhập lại ở trạng thái sau

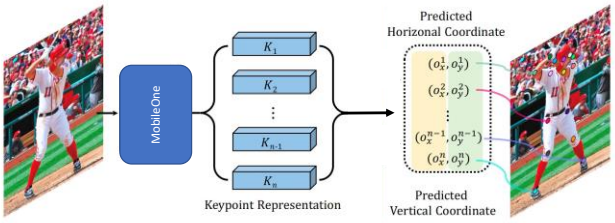
thông qua hệ số α để giữ lại một phần thông tin của đặc trưng của khung hình trước đó làm tránh mất mát thông tin khi người đó chuyển động.

$$feature_t = feature_t * \alpha + feature * (1 - \alpha)$$

$feature$: là trích xuất đặc trưng khung xương

t: là khung hình thứ t

Tuy nhiên mô hình Human Pose có một nhược điểm đó là thời gian thực thi khá lâu vì mô hình cần đủ lớn để có thể học được vị trí các tọa độ khung xương trên cơ thể con người. Điều này yêu cầu kiến trúc của mô hình gốc (backbone) phải được xây dựng có chiều sâu tốt và các thông tin đặc trưng cần phải bổ sung ngữ nghĩa cho nhau sau mỗi một khối. Đánh đổi lại là tốc độ xử lý sẽ rất chậm. Để giải quyết vấn đề này, tác giả thay backbone của bài báo gốc (Sun, et al., 2019) bằng mô hình MobileOne (Vasu, et al., 2022) có kích thước nhỏ chạy rất nhanh tương tự như các thế hệ MobileNet và độ chính xác cao. Mô hình sử dụng giải pháp tham số hóa lại sẽ được trình bày phần sau nhằm đổi cấu trúc mô hình nhẹ hơn khi thực thi.



Hình 3. Mô hình Human Pose

Kết quả của mô hình sẽ là một bộ tọa độ vị trí (O_x, O_Y) của 17 điểm trên cơ thể con người. Các điểm này sẽ được cập nhập thông qua hàm MSE loss theo công thức.

$$loss = \frac{1}{2 * N} * \sum_{i=0}^{N} \left(x_p^i - x_{gt}^i\right)^2 + \left(y_p^i - y_{gt}^i\right)^2$$

Trong đó:

N: là số các bộ phận trên khung xương

x_p^i : tọa độ dự đoán x tại bộ phận i

y_p^i : tọa độ dự đoán y tại bộ phận i

x_{gt}^i: tọa độ thực tế x tại bộ phận i

y_{gt}^i: tọa độ thực tế y tại bộ phân i

Trong phần nhận diện khuôn mặt, nếu như hai khuôn mặt gần nhau vì hai người đứng quá gần nhau thì việc định danh hai khuôn mặt đó thuộc cơ thể nào cần phải được giải quyết. Tác giả đã đưa ra giải pháp dựa vào vị trí xương đầu của người để liên kết với tọa độ đầu người đã được trích xuất từ Object Detection để xác định phần đầu người nào gần gần tọa độ xương đầu nhất sẽ thuộc về người đó.

Hình 4. Liên kết đầu vời và xương người

Trên Hình 4, chấm tròn ở giữa mặt người và viền đỏ xung quanh là tọa độ xương mặt người được trích xuất từ được trích xuất từ mô hình Human Pose và viền màu xanh thể hiện được vị trí của đầu người được trích xuất từ mô hình Object Detection. Hai bbox sẽ được liên kết với nhau để xác định đầu người thuộc đối tượng nào.

3.3. Ứng dụng tham số hóa lại mô hình Deep learning

Như đã đề cập ở trên thì tác giả đã sử dụng giải pháp tham số hóa lại mô hình bằng cách dùng mô hình MobileOne, ngoài ra đối với mô hình nhận diện khuôn mặt, tác giả cũng sử dụng giả pháp này cho mô hình RepVGG (Ding, et al., 2021) như là một phương án cho mô hình nhận diện khuôn mặt. Điểm đặc biệt của giải pháp này sẽ chia làm 2 giai đoạn đó là huấn luyện và thực thi.

Phần huấn luyện sẽ gồm có 3 nhánh gồm 2 lớp convolution là 3x3 ($W^{(3)}$), 1x1 ($W^{(1)}$) và một lớp định danh (identity). Như công thức phía dưới thì $\mu, \alpha, \gamma, \beta$ lần lượt là mean, stanndard deviation, scalling factor và bias của lần lượt 3 nhánh. Ba lớp này sẽ đi qua một lớp chuẩn hóa bn (Batch Normalization) sau đó cộng (add) các công thức trọng số của 3 lớp lại với cho ra kết quả là $M^{(2)}$, vì yêu cầu các lớp cùng kích thước nên phải thêm các trọng số bằng 0 (padding = 0) vào vùng biên của $W'^{(1)}$, $W'^{(0)}$ có kích thước lần lượt là 1x1 và 1x1 tương ứng với lớp convolution 1x1 và Identity trước khi add để cùng kích thước 3x3 của convolution 3x3.

Với $\forall 1 \leq i \leq C_2$,

C_2 *là số kênh đầu ra*, C1 là số kênh đầu vào

$$bn(M, \mu, \sigma, \gamma, \beta)_{:,i,:,:} = (M_{:,i,:,:} - \mu_i)\frac{\gamma_i}{\sigma_i} + \beta_i$$

$$= M_{:,i,:,:}\frac{\gamma_i}{\sigma_i} + (-\mu_i\frac{\gamma_i}{\sigma_i} + \beta_i)$$

Đặt $W'_{i,:,:,:} = \frac{\gamma_i}{\sigma_i} W_{i,:,:,:} b'_i = -\frac{\mu_i \gamma_i}{\sigma_i} + \beta_i$

$$M^{(2)} = bn(M^{(1)} * W^{(3)}, \mu^{(3)}, \sigma^{(3)}, \gamma^{(3)}, \beta^{(3)}) + bn(M^{(1)} * W^{(1)}, \mu^{(1)}, \sigma^{(1)}, \gamma^{(1)}, \beta^{(1)}) + bn(M^{(1)}, \mu^{(0)}, \sigma^{(0)}, \gamma^{(0)}, \beta^{(0)})$$

$$M^{(2)} = (M^{(1)} * W'^{(3)})_{:,i,:,:} + b_i'^{(3)} + (M^{(1)} * W'^{(1)})_{:,i,:,:} + b_i'^{(1)} + (M^{(1)} * W'^{(0)})_{:,i,:,:} + b_i'^{(0)}$$

$$M^{(2)} = M^{(1)} * (W'^{(1)} + W'^{(3)} + W'^{(0)}) + (b'^{(0)} + b'^{(3)} + b'^{(1)}) = M^{(1)} * W'^{(1+3+0)} + b'^{(1+3+0)}$$

Phần thực thi chỉ có 1 lớp convolution 3x3 và loại bỏ 2 lớp còn lại và đồng thời được bao bọc bởi Bacth

Normalization. Dù đã loại bỏ đi 2 lớp này nhưng kết quả đầu ra của mô hình vẫn không có sự hay đổi so với kiến trúc lúc huấn luyện $M^{(2)}$.

$$Ta\ có \quad M^{(2)} = bn(M * W, \mu, \sigma, \gamma, \beta)_{:,i,,,:}$$
$$= \left(M_{:,i,,,:}W_{:,i,,,:} - \mu_i\right)\frac{\gamma_i}{\sigma_i} + \beta_i$$
$$= \left(M^{(1)} * W'\right)_{:,i,,,:} + b_i'$$
$$= M^{(1)} * W' + b$$

Như vậy bài báo gốc đã thay đổi được kiến trúc của mô hình trước và sau khi huấn luyện để mô hình tính toán ít hơn khi có ít tham số tính toán hơn nhưng vẫn giữ được độ chính xác cao. Tác giả đã áp dụng cách thức này để xây dựng mô hình Human Pose và Face Recognition với mục tiêu giúp cho mô hình nhẹ nhất có thể để thực thi nhanh hơn nhưng vẫn giữ độ chính xác cao.

4. Thực nghiệm và đánh giá kết quả.

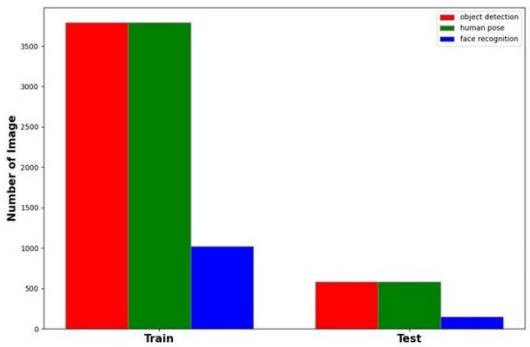
4.1. Dữ liệu.

Để giải quyết bài toán hiện tại, tác giả sử dụng 3 bộ dữ liệu khác nhau ứng dụng vào các bài toán khác nhau gồm: bộ dữ liệu COCO, bộ dữ liệu MPII, bộ dữ liệu AFAD.

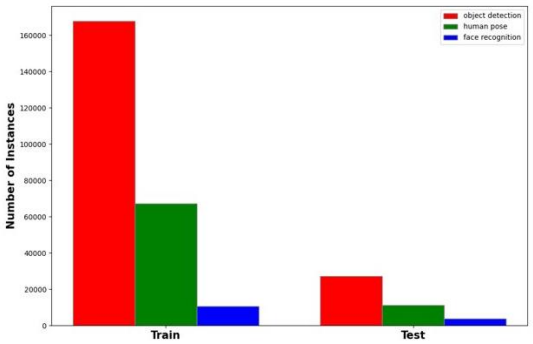
COCO (Common Objects in Context) là một tập dữ liệu rất lớn phục vụ cho các bài toán Object Detection, Segmentation, Image Captioning. Tập dữ liệu tổng cộng có khoảng 1.5 triệu objects thuộc về 80 class khác nhau. Tuy nhiên trong bài toán hiện tại tác giả chỉ sử dụng để phát hiện đối tượng là con người nên chỉ sử dụng 2 class: con người và đầu người. Vì đầu người không thuộc 80 class của tập dữ liệu gốc nên tác giả thực hiện sinh dữ liệu thông qua mô hình có sẵn và đã được tách ra từ những ảnh có hình con người.

MPII Human Pose là dữ liệu về khung xương người đã được chuẩn hóa dựa vào các khớp nối tại các điểm trên khung xương. Tập dữ liệu gồm 25 nghìn ảnh chứa hơn 40 nghìn người. Các hình ảnh được thu thập dựa trên hoạt động thường ngày của con người nên có độ bao quát cao. Tác giả sử dụng bộ dữ liệu để thực hiện huấn luyện khung xương người nhằm trích xuất đặc trưng để đưa vào giải quyết bài toán truy vết đối tượng (human tracking).

Tập dữ liệu tuổi khuôn mặt châu Á (AFAD) là tập dữ liệu mới được đề xuất để đánh giá hiệu suất ước tính tuổi, chứa hơn 160 nghìn hình ảnh khuôn mặt và các nhãn độ tuổi và giới tính tương ứng. Bộ dữ liệu này được định hướng để ước tính độ tuổi trên khuôn mặt Châu Á, vì vậy tất cả các hình ảnh trên khuôn mặt đều dành cho khuôn mặt Châu Á. Tác giả muốn dựa vào khuôn mặt để xác định danh tính của đối tượng khả nghi, đặc biệt đối tượng nhắm đến là người Châu Á nên đã sử dụng tập AFAD để giải quyết bài toán hiện tại.



Hình 5. Số lượng ảnh của 3 tập dữ liệu



Hình 6. Số lượng các instance trong từng ảnh của 3 tập dữ liệu

4.2. Ma trận đánh giá.

Để đánh giá bài toán Object Detection cac giả dựa vào các chỉ số Precision, Recall, mAP (mean average precision) để đưa ra góc nhìn rõ nhất

$$R = \frac{TP}{TP + FN} = \frac{succesful\ detections}{all\ ground\ truths}$$

$$P = \frac{TP}{TP + FP} = \frac{succesful\ detections}{all\ detections}$$

$$mAP = \frac{1}{c}\sum_{i=1}^{c}\frac{1}{11}\sum_{j}p(j)$$

Trong đó:

R (Recall): tỷ lệ nhận diện đúng vật thể trên tổng số vật thể mà mô hình nhận diện

P (Precision): tỷ lệ nhận diện đúng vật thể trên tập đã đánh nhãn

mAP: giá trị trung bình của quyết định dự đoán của các class.

Bài toán Tracking yêu cầu ma trận đánh giá tương đối phức tạp khi vừa phải xác định vị trí vật thể vừa phải định danh. Do đó tác giải dựa vào ba chỉ số chính là HOTA, MOTA, và IDF.

$$MOTA = 1 - \frac{|FN| + |FP| + |IDSW|}{gtDet}$$

$$IDF1 = \frac{|IDTP|}{|IDTP| + 0.5|IDFN| + 0.5|IDFP|}$$

$$HOTA = \sqrt{\frac{|TP|}{|TP| + |FN| + |FP|} * \frac{1}{|TP|}\sum_{c\,\epsilon\,TP}\frac{|TPA(c)|}{|TPA(c)| + |FNA(c)| + |FPA(c)|}}$$

Trong đó:

MOTA: là độ chính xác nhận diện vật thể trong quá trình tracking.

IDF1: là độ chính xác định danh cho vật thể trong quá trình tracking.

HOTA: là đô chính xác của mô hình tracking dựa trên nhận diện chính xác vật thể và định danh vật thể.

4.3. Kết quả Object Detection

Bảng 1:

Mô hình Yolox để nhận diện người và đầu người

	Image Size	mAP	Precision	Recall

Yolox-nano	416	32.498	0.325	0.381
Yolox-s	**640**	**42.985**	**0.430**	**0.504**

Tác giả thử nghiệm trên hai phiên bản YOLOX khác nhau là Yolox-nano và Yolox-s và cho ra kết quả khá khác biệt. Các chỉ số về Precision và Recall của của mô hình Yolox-nano là thấp hơn nhiều so với Yolox-s vì kích thước mô hình nhỏ hơn. Tuy nhiên tốc độ của phiên bản nano lại nhanh hơn 1.5 lần so với phiên bản s. Trong quá trình thực nghiệm, tác giả nhận thấy rằng, sử dụng Yolox-s là cân bằng nhất về độ chính xác cũng như là tốc độ nhận diện vật thể.

4.4. Kết quả Human Pose

Bảng 2

So sánh mô hình PELW và các mô hình hiện tại về trích xuất khung xương người.

	Image Size	AP	AR	Params	Time (s) CPU
Pose_ResNet_50	256x192	0.886	0.763	34M	0.23
Pose_ResNet_101	256x192	0.893	0.771	53M	0.33
Pose_ResNet_152	256x192	0.893	0.778	68.7M	0.5
Pose_HrNet_w32	256x192	0.905	0.798	28.5M	0.35
Pose_HrNet_w48	256x192	0.906	0.804	63.6M	0.6
PELW (our model)	**256x192**	**0.835**	**0.711**	**18M**	**0.1**

Kết quả cho thấy các chỉ số Accuracy Precision (độ chính xác trên tất cả dự đoán) và Accuracy Recall (độ chính xác trên tập đã đánh nhãn) của mô hình PELW là thấp hơn so với các mô hình còn lại tuy nhiên điểm khác biệt là không quá nhiều khi độ chính xác dự đoán của mô hình là 83% trên tập test COCO. Ngoài ra tốc độ của mô hình PELW là vượt trội hơn so với các mô hình còn lại vì số lượng tham số tính toán của mô hình sau khi được chuyển sang dạng thực thi là 18 triệu. Tác giả sử dụng PELW như là một giải pháp tối ưu về tốc độ xử lý nhưng độ chính xác vẫn không giảm quá nhiều để đưa vào thực hiện trích xuất khung xương cũng như trích xuất đặc trưng để hậu xử lý bài toán truy vết đối tượng khả nghi.

4.5. Kết quả Human Tracking

Bảng 3

So sánh kết quả của các mô hình tracking được thực thi trên NVIDIA GPU 1050 4G trên tập MOT17

	Base-model	Re-id	Image size	MOTA	IDF1	HOTA	FPS
Byte Track	Yolox-nano	No	416	26.29	37.29	31.54	14
SHTrack	Yolox-nano	HrNet	416	26.21	37.45	31.52	5
Byte Track	Yolox-s	No	640	50.96	58.64	46.66	12
SHTrack	**Yolox-s**	**Mobile One**	**640**	**51.07**	**59.64**	**47.31**	**10**

Mô hình được đánh giá trên tập MOT17, tác giả đã thực hiện so sánh với mô hình ByteTrack trên hai phiên bản YOLOx và đã cho ra kết quả khả quan. Đặt biệt là mô hình SHTrack khi dùng phiên bản Yolox-s để nhận diện và MobileOne để định danh cho ra kết quả cao hơn so với lại mô hình ByteTrack ở tất cả các chỉ số, đồng thời thời gian thực thi trên GPU lại không giảm quá nhiều. Kết quả cho thấy được mô hình khi thêm trích xuất đặt trưng của khung xương người để định danh cho kết quả tốt hơn.

4.6. Kết quả Face Recognition

Bảng 4

Kết quả mô hình nhận diện gương mặt trên tập ADFA

	Image Size	Acc Train	Acc Test	Thres
RepVGG	128X128	0.99	0.94	0.2

Vì tập dữ liệu AFDA tương đối lớn nên độ chính xác của mô hình rất cao đạt 99% cho tập train và 94% cho tập test. Ngoài ra qua đánh giá của tập test, tác giả rút ra được giá trị ngưỡng để phân biệt được đó có phải là gương mặt của đối tượng đang vị theo dõi là 0.2, nếu lớn hơn 0.2 thì đó là đối tượng khả nghi và ngược lại.

5. Thảo luận

Trong bài toán hiện tại, tác giả đã giải quyết được một phần nào đó bài toán truy vết đối tượng trong đám đông. Tuy nhiên vẫn có những khó khăn và thách thức chư giải quyết được.

Bài toán chưa giải quyết được vào ban đêm khi ánh sáng giảm và khả năng nhận diện được đối tượng là rất khó khăn. Đối với những ảnh có chất lượng không tốt gây ảnh hưởng phần nào đến kết quả trích xuất đặt trưng cũng như phát hiện đối tượng khả nghi. Ngoài ra, mô hình chưa hoạt động tốt nếu như ở đám đông quá nhiều người, và camera đặt ở góc cao. Khi đó chất lượng ảnh sẽ rất thấp nếu như phóng to khu vực chứa đối tượng khả nghi gây nhiễu cho mô hình truy vết.

Mô hình vẫn chưa hoạt động tốt khi số lượng người là quá đông và có sự chồng chéo lên nhau quá nhiều. Vì mô hình hiện tại được huấn luyện có kích thước nhỏ, nên việc nhận diện đối tượng trong đám đông vẫn còn hạn chế.

6. Kết luận

Hiện nay, có rất nhiều phương pháp khác nhau để phát hiện cũng như truy vết đối tượng khả nghi. Trong bài báo này, tác giả đã đưa ra một giải pháp tối ưu cho việc phát hiện và truy vết đối tượng khả nghi, đặc biệt là những đối tượng truy nã bởi các cơ quan chức năng. Qua đó cũng cho thấy rằng, việc trích xuất đặc trưng của đối tượng là vô cùng quan trọng và ảnh hưởng đến tổng thể của bài toán truy vết nói riêng và xử lý ảnh nói chung.

Tài liệu tham khảo

[1] Aharon, N., Orfaig, R. & Bobrovsky, B.-Z., 2022. BoT-SORT: Robust Associations Multi-Pedestrian Tracking.

[2] Akdemir, U., Turaga, P. & Chellappa, R., 2008. An Ontology based Approach for Activity Recognitionfrom Video.

[3] Amrutha, C, J. & J, A., 2020. *Deep Learning Approach for Suspicious Activity Detection from Surveillance Video.* s.l., s.n.

[4] Ding, X. et al., 2021. RepVGG: Making VGG-style ConvNets Great Again.

[5] Du, Y., Song, Y., Yang, B. & Zhao, Y., 2022. StrongSORT: Make DeepSORT Great Again.

[6] Ge, Z. et al., 2021. YOLOX: Exceeding YOLO Series in 2021 V100 batch 1 Latency (ms) YOLOX-L YOLOv5-L YOLOX-DarkNet53 YOLOv5-Darknet53 EfficientDet5 COCO AP (%) Number of parameters (M) Figure 1: Speed-accuracy trade-off of accurate models (top) and Size-accuracy curve of lite.

[7] Ibrahim, N., Mustafa, M., Mustafa, M. & Lee, Y. S., 2012. Detection of snatch theft based on temporal differences in motion flow field orientation histograms. *International Journal of Advancements in Computing Technology.*

[8] K, H., G, G., P, D. & R, G., 2017. Mask R-CNN.

[9] P. R. Gunjal, B. R. G. H. A. S. S. M. V. a. S. S. A., 2018. Moving Object Tracking Using Kalman Filter," 2018 International Conference On Advances in Communication and Computing Technology. *ICACCT.*

[10] Phyo, C. N., Zin, T. T. & Tin, P., 2019. Deep learning for recognizing human activities using motions of skeletal joints. *IEEE Transactions on Consumer Electronics.*

[11] Redmon, J., Divvala, S., Girshick, R. & Farhadi, A., 2015. You Only Look Once: Unified, Real-Time Object Detection.

[12] Ren, S., He, K., Girshick, R. & Sun, J., 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.

[13] Sun, K., Xiao, B., Liu, D. & Wang, J., 2019. Deep High-Resolution Representation Learning for Human Pose Estimation.

[14] Vasu, P. K. A. et al., 2022. An Improved One millisecond Mobile Backbone.

[15] Wang, C.-Y., Bochkovskiy, A. & Liao, H.-Y. M., 2022. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors.

[16] Zhang, Y., He, B. & Sun, L., 2020. Progressive Multi-stage Feature Mix for Person Re-Identification.

[17] Zhang, Y. et al., 2022. ByteTrack: Multi-Object Tracking by Associating Every Detection Box.

[18] Zhang, Y. et al., 2020. FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking.