Đại học quốc gia TP.HCM
Trường đại học công nghệ thông tin

Khoa: Khoa học máy tính

Môn học: Nhập môn Thị giác máy tính

Lớp: CS231.O22

BÁO CÁO ĐỒ ÁN CUỐI KỲ
SINGLE OBJECT TRACKING

Sinh viên thực hiện:
Nguyễn Trung Kiên - 21521024
Võ Đức Dương - 21521992

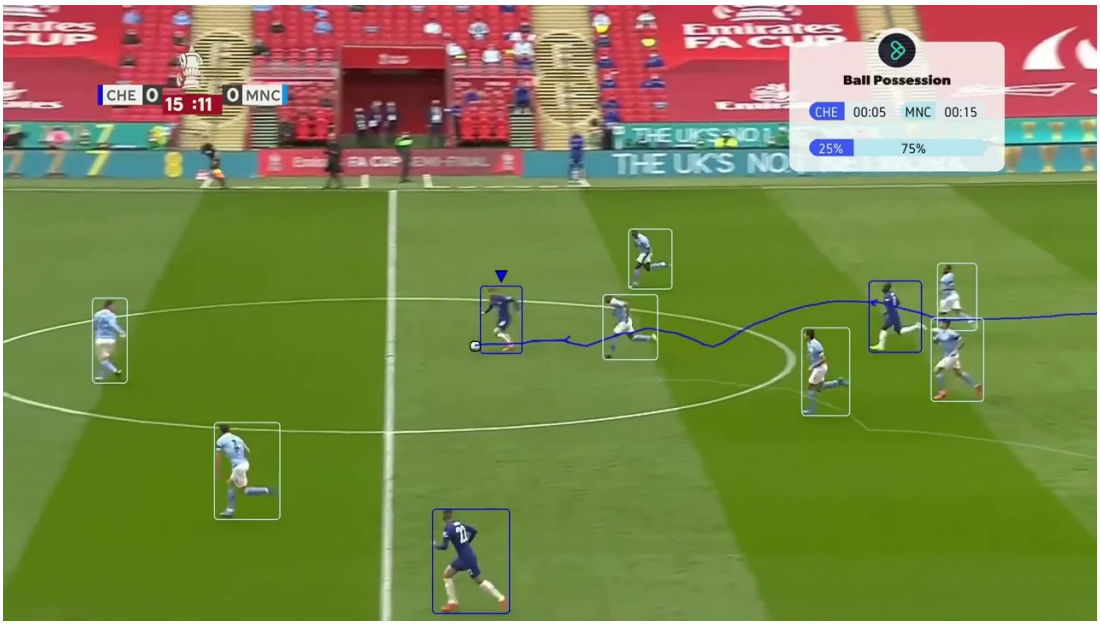
Giảng viên hướng dẫn:
TS. Mai Tiến Dũng

Mục lục

1 Lý do lựa chọn đề tài

Object Tracking (OT) là một trong những lĩnh vực nghiên cứu thuộc lĩnh vực Thị giác máy tính với mục tiêu theo dõi vị trí của một hoặc nhiều đối tượng trong một chuỗi các khung hình (frames) liên tiếp nhau.

Một ứng dụng của bài toán OT trong thực tế là các con số thống kê trong một trận đấu bóng đá. Chẳng hạn, ta có thể biết được thời lượng kiểm soát bóng của mỗi đội (hình 1) bằng cách theo dõi từng cầu thủ và cả quả bóng trên sân[1]. Việc này giúp các khán giả có được nhiều thông tin về trận đấu hơn, đồng thời giúp các đội bóng có thể dùng các thống kê có được để phân tích về những đối thủ và xây dựng chiến thuật hợp lý để đối phó.



Hình 1: Ứng dụng của OT trong lĩnh vực bóng đá

Single Object Tracking (SOT) là một nhánh của Object Tracking khi từ chuỗi các khung hình liên tiếp, ta chỉ cần theo dõi vị trí và trạng thái của một đối tượng cụ thể duy nhất. Chính vì những ứng dụng thú vị của bài toán OT và việc lựa chọn một đề tài phù hợp cho môn học là **hai lý do chính** mà nhóm mong muốn tìm hiểu và thực hiện các giải pháp để giải quyết bài toán SOT.

[1] Tryolabs | Automated soccer ball possession using AI

2 Giới thiệu bài toán

2.1 Định nghĩa bài toán

Cho một tập hợp F gồm các khung hình liên tiếp chứa vật thể cần theo dõi và tọa độ của bounding box bao xung quanh vật thể cần theo dõi ở khung hình đầu tiên dưới dạng xywh (x_min, y_min, width, height). Mục tiêu cần thực hiện là xác định tọa độ của bounding box trên toàn bộ các khung hình trong tập hợp F.

Input của bài toán:

- Chuỗi các frame liên tiếp nhau của một video có vật thể cần theo dõi.
- Tọa độ bounding box ở frame đầu tiên, được định nghĩa theo format xywh của vật thể cần theo dõi.

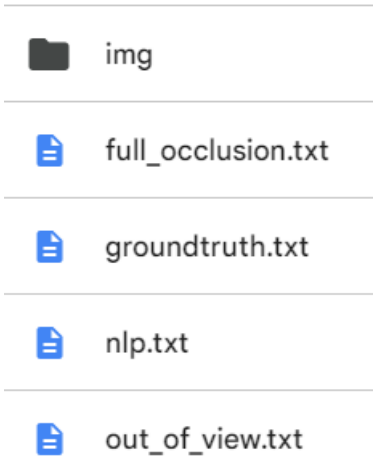
Output của bài toán: Tập hợp các tọa độ của bounding box bao xung quanh vật thể trên từng frame trong chuỗi frame.

2.2 Mô tả bộ dữ liệu

LaSOT (Large-scale Single Object Tracking) là một bộ dữ liệu dành riêng cho bài toán SOT, chứa khoảng 1400 chuỗi frames với hơn 3.5 triệu khung hình, gồm nhiều đối tượng khác nhau để theo dõi đơn vật thể. Mỗi chuỗi frames đối tượng trong bộ dữ liệu LaSOT chứa các thông tin đi kèm bao gồm:

- img: chuỗi các khung hình (frame) của video cần thực hiện tracking.
- ground_truth: tọa độ các bounding box ở các khung hình trong file *img*.
- nlp: một đoạn text mô tả ngắn gọn nội dung video.

- full_occusion: gồm nhiều số 0 và 1 liên tiếp nhau mô tả thông tin vật thể có bị che khuất bởi vật khác trong frame thứ i hay không. Giá trị 0 cho biết vật thể không bị che khuất và giá trị 1 cho trường hợp ngược lại.
- out_of_view: gồm nhiều số 0 và 1 liên tiếp nhau mô tả thông tin vật thể có bị rời khỏi frame thứ i hay không. Giá trị 0 cho biết vật thể vẫn còn ở trong frame và giá trị 1 cho trường hợp ngược lại.



Hình 2: Các thông tin trong chuỗi khung hình của một đối tượng

2 dạng vật thể được nhóm thực nghiệm với LaSOT là **person** và **mouse**.

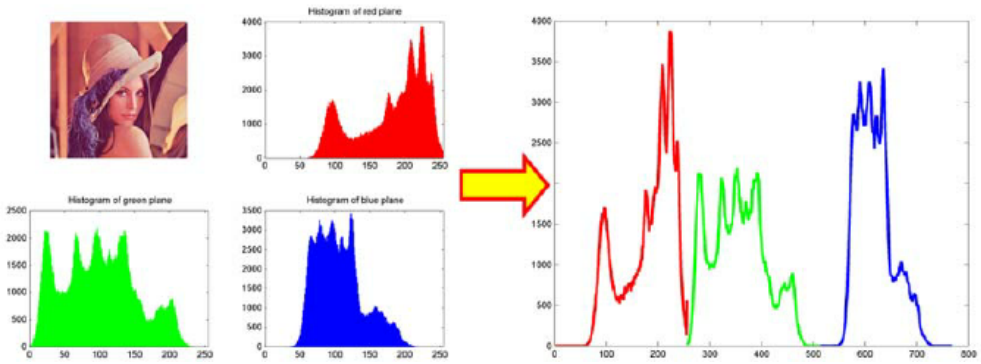
3 Các phương pháp thực hiện

Nhóm tách ra hai hướng tiếp cận để giải quyết bài toán:

- **Hướng thứ nhất**: Sử dụng kết hợp đặc trưng **Histogram** và **Mean Shift Tracker**. Khi đã biết tọa độ bounding box ở frame i - 1, ta sẽ sử dụng thông tin đó để dò ra vị trí của vật thể ở frame i. Từ frame đầu tiên, lần lượt tìm ra vị trí bounding box trong các frame kế tiếp theo trình tự quy nạp.
- **Hướng thứ hai**: Sử dụng kết hợp **YOLO** và **deepSORT**. Đối với phương pháp này, detector và tracker đều là các pre-trained model, trong đó detector là *YOLOv9* và tracker là *deep-sort-realtime 1.3.2*. Điểm khác của hướng tiếp cận trên so với ***detection-based tracking*** truyền thống là bounding box của mục tiêu cho frame đầu tiên cũng sẽ được dùng làm input. Bounding box này sẽ được sử dụng để lấy ra objectID của đối tượng cần theo dõi. Mỗi frame bắt đầu từ frame thứ hai sẽ đưa qua detector để thực hiện phát hiện vật thể, sau đó các detection này sẽ được sử dụng để cập nhật tracker. Quá trình lặp lại cho tới khi hoàn tất xử lý trên toàn bộ các frame.

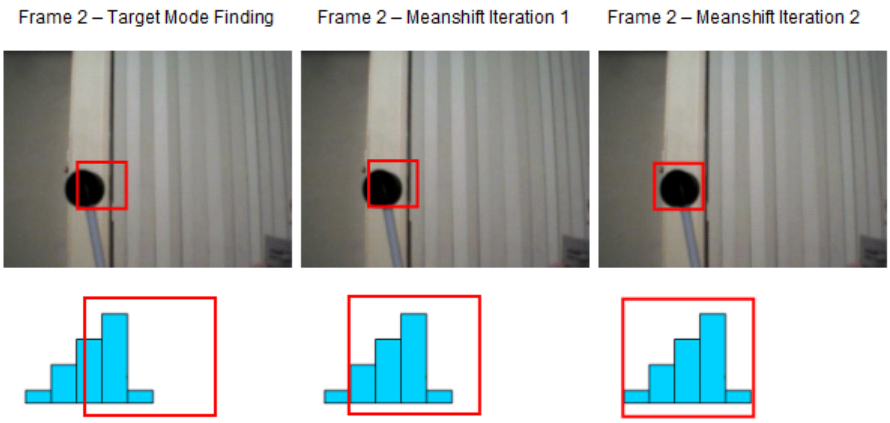
3.1 Histogram + Mean Shift Tracker

Bước 1: đưa vùng không gian của bounding box trong frame $i-1$, đưa về 3 vector đặc trưng Histogram ứng với 3 mô hình màu RGB và ghép lại thành 1 vector đặc trưng duy nhất theo thứ tự lần lượt là đỏ, xanh lá, xanh lục.



Hình 3: Ví dụ bằng ảnh Lenna được chuyển thành vector đặc trưng

Bước 2: Tìm những vùng của frame i "giống" với phần bounding box của frame $i-1$ sử dụng Histogram backprojection để tạo ra một ảnh mới từ frame i có kích thước tương đương. Trong đó, mỗi pixel trong ảnh mới là xác suất mà pixel tại vị trí đó trong ảnh gốc nằm trong phần bounding box của frame $i-1$. Từ phân bố xác suất các pixel ở bước 2, sử dụng thuật toán Mean Shift giúp tìm được vùng có phân bố xác suất cao nhất, cũng chính là vị trí có khả năng chứa vật thể cao nhất để cập nhật vị trí cho bounding box.



Hình 4: Minh họa cách hoạt động thuật toán Mean Shift

3.2 Yolov9 + DeepSort

Bước 1: Trước khi thực hiện tracking, ta cần khai báo các tham số cần thiết:

- **Tracking class:** classID tương ứng của mục tiêu. Giá trị này được khởi tạo dựa trên các class name của YOLO.
- **Object ID:** frame đầu tiên cùng với bounding box tương ứng sẽ được đưa trực tiếp qua tracker để lấy ra object ID của mục tiêu.

Bước 2: Lần lượt các frame bắt đầu từ frame thứ hai trong chuỗi video sẽ được đưa qua YOLO để phát hiện vật thể. Một frame có thể có nhiều vật thể. Khi quá trình phát hiện hoàn tất, ta thu được một danh sách các detection là danh sách các vật thể được detector (YOLO) phát hiện.

- Nếu classID của vật thể mà YOLO phát hiện trùng với tracking class thì detection này sẽ được lưu vào một danh sách chứa các detection với các classID trùng với tracking class.
- Tất cả các detection (kể cả các classID không hợp lệ) đều được lưu vào một danh sách khác.

figure 1: many detections in one frame

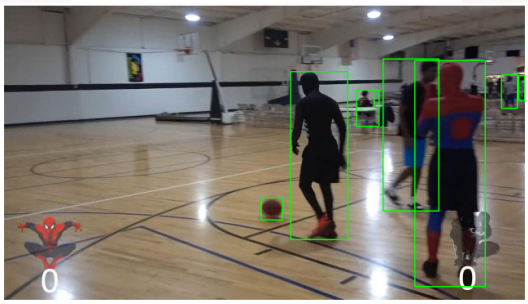
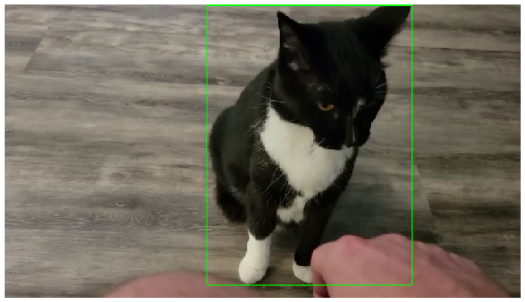


figure 2: one detection in one frame



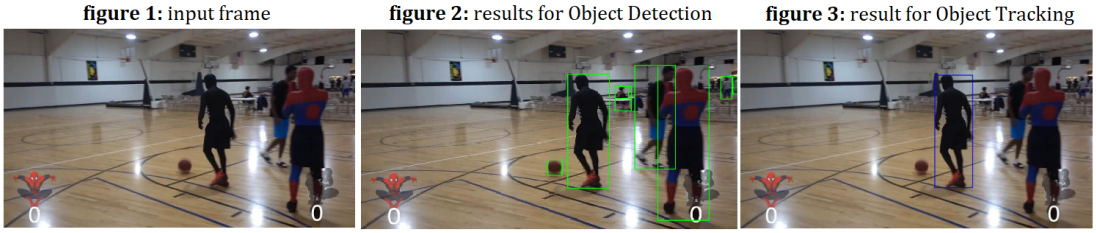
Hình 5: Một frame có thể có một hoặc nhiều detection phù hợp với tracking class

Một frame có thể có nhiều đối tượng thuộc cùng một class. Ngoài ra, YOLO cũng có thể nhầm lẫn một đối tượng thuộc classID khác thành đối tượng thuộc tracking class. Do đó, tất cả các detection mà YOLO phát hiện được đều sẽ được xem xét để cập nhật tracker.

Bước 3: Bounding box cho mục tiêu tại frame đầu tiên sẽ được sử dụng để cập nhật tracker (deepSORT). Nếu trong các detection có classID trùng với tracking class thì các detection đó sẽ được dùng để cập nhật tracker. Ngược lại, toàn bộ các detection sẽ được sử dụng.

Trong số các track sau khi được cập nhật, track được lựa chọn là track có trackID trùng với objectID, hoặc track có confidence score (độ tự tin của dự đoán) trên detection tương ứng cao nhất. Trường hợp phát hiện vật thể thất bại, tracking region (bounding box dự đoán vị trí mục tiêu) của frame trước đó sẽ được lựa chọn làm kết quả cho frame hiện tại.

Các tracking region sẽ được lưu lại để tạo video kết quả và đánh giá mô hình.



Hình 6: Ví dụ về quy trình tracking sử dụng YOLO và deepSORT

4 Đánh giá mô hình

4.1 IoU

Xét 2 bounding box từ 1 frame, tọa độ bounding box A là do phương pháp dự đoán và tọa độ bounding box B là tọa độ ground-truth ở frame đó. Để xác định phương pháp có dự đoán đúng hay không khi xét trên một frame, nhóm sử dụng giá trị IoU. Công thức của IoU (Intersection over Union) được tính bằng cách:

$$\text{IoU} = \frac{\text{Diện tích phần giao giữa A và B}}{\text{Diện tích phần hợp giữa A và B}}$$

Trong trường hợp lý tưởng, khi grouth-truth bounding box trùng với dự đoán của phương pháp thì IoU sẽ đạt giá trị là 1. Nhóm đặt ngưỡng cho giá trị IoU là 0.5. Khi IoU $>=$ 0.5 thì phương pháp được xem là dự đoán đúng và ngược lại.

4.2 Successful rate

Successful rate (tỉ lệ thành công) là tỷ lệ các frame được dự đoán đúng trên tổng số các frame. Công thức của Successful rate được tính bằng cách:

$$\text{Successful rate} = \frac{\text{Số các khung hình mà phương pháp dự đoán đúng}}{\text{Tổng số khung hình được xét}}$$

Successful rate và độ hiệu quả của phương pháp là 2 đại lượng tỷ lệ thuận.

5 Kết quả thực nghiệm

5.1 Bảng kết quả

Dưới đây là các giá trị successful rate (đơn vị %) của mỗi phương pháp cho từng chuỗi frames cụ thể được thực nghiệm. Các kết quả tốt hơn được **in đậm**.

LaSOT Dataset	**Histogram & MeanShift**	**YOLO & deepSORT**
mouse-1	**7.046**	2.973
mouse-2	**0.747**	0.093
mouse-3	**0.050**	0.050
mouse-4	**0.464**	0.258
mouse-5	**0.042**	0.042
mouse-6	0.727	**23.667**
mouse-7	0.167	**37.793**
mouse-8	0.157	**24.627**
mouse-9	0.248	**43.116**
mouse-10	**0.429**	0.322
mouse-11	**0.570**	0.057
mouse-12	0.045	**0.446**
mouse-13	0.087	**1.817**
mouse-14	**0.047**	**0.047**
mouse-15	**1.920**	0.427
mouse-16	0.022	**1.400**

LaSOT Dataset	Histogram & MeanShift	YOLO & deepSORT
mouse-17	0.062	**4.540**
mouse-18	3.488	**33.181**
mouse-19	1.212	**51.052**
mouse-20	0.050	**7.049**
person-3	0.400	**19.937**
person-4	0.041	**44.219**
person-5	0.034	**34.884**
person-6	1.044	**60.816**
person-7	2.909	**20.646**
person-9	3.413	**19.113**
person-10	6.639	**100.000**
person-11	0.171	**54.954**
person-12	0.050	**38.844**
Kết quả trung bình	0.011	**23.931**

5.2 Nhận xét kết quả

- Nhìn chung, phương pháp sử dụng YOLO kết hợp deepSORT cho kết quả tốt hơn so với Histogram kết hợp MeanShift.
- Vẫn còn một vài chuỗi frames mà cả 2 phương pháp cho kết quả rất tệ. Điểm chung của các chuỗi frames này là thách thức của bài toán SOT trong dữ liệu. Trên hầu hết các chuỗi frames, mục tiêu cần theo dõi nằm ngoài khung hình hoặc bị che khuất bởi vật khác hoặc bị lẫn vào background hoặc quá mờ, quá bé để detector phát hiện.
- Đối với những dataset có ít nhiễu, YOLO có thể dễ dàng phát hiện mục tiêu và do đó tracking cho ra kết quả tốt. Một số trường hợp có tỉ lệ thành công trên 50%.
- Phương pháp sử dụng đặc trưng Histogram kết hợp MeanShift có tỉ lệ thành công rất thấp. Nhóm cho rằng phương pháp này đã cũ hoặc đặc trưng Histogram không đủ tốt để giải quyết SOT tốt hơn.

6 Tài liệu tham khảo

1. LaSOT: A High-quality Benchmark for Large-scale Single Object Tracking
2. Single Object Tracking: A Survey of Methods, Datasets, and Evaluation Metrics
3. Computer Vision for object tracking
4. Single Object Tracking: Challenges, Techniques, and Future Directions
5. Hiểu về độ đo IoU trong nhận diện thực thể
6. Histogram Backprojection
7. MeanShift Tracking
8. YOLO trong bài toán Real-time Object Detection
9. YOLov9 description
10. YOLOv9 - models.predict
11. Simple Online Realtime Object Tracking
12. Simple Online and Realtime Tracking with a Deep Association Metric
13. deepSORT description