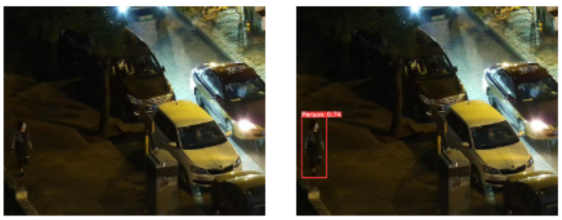
NHẬN DIỆN NGƯỜI ĐI BỘ VÀO BAN ĐÊM VỚI BỘ DỮ LIỆU LLVIP

1st Nguyễn Đỗ Quỳnh Như
Khoa Khoa học Máy tính, UIT
HCMC, Viet Nam
21521243@gm.uit.edu.vn

2nd Phạm Thị Trâm Anh
Khoa Khoa học Máy tính, UIT
HCMC, Viet Nam
21520146@gm.uit.edu.vn

***Tóm tắt nội dung*—Nhận diện người đi bộ là một bài toán quan trọng trong lĩnh vực thị giác máy tính và trí tuệ nhân tạo. Bài toán này có nhiều ứng dụng trong các lĩnh vực khác nhau như giám sát an ninh, quản lý giao thông, và tự động hóa. Hầu hết các bài nghiên cứu tập trung vào việc phát hiện người đi bộ sử dụng hình ảnh thấy được (visual image)vào ban ngày. Tuy nhiên, bài toán thú vị và thách thức hơn khi điều kiện môi trường thay đổi với ánh sáng yếu hoặc ban đêm. Gần đây, một số ý tưởng mới mới tập trung sử dụng cả ảnh thấy được và ảnh cảm biến hồng ngoại để phát hiện người đi bộ trong điều kiện ánh sáng yếu. Bộ dữ liệu chuẩn mực của bái toán này là LLVIP (Low-light Visible-Infrared Paired). Trong bài báo cáo này, chúng em sẽ áp dụng kiến thức máy học cơ bản và mô hình học sâu để cải thiện độ chính xác trên bộ dữ liệu benchmark LLVIP trong bài toán phát hiện người đi bộ ở môi trường yếu ánh sáng. Code tham khảo: https://github.com/IrisPham74/CS331.O11.KHCL.git**



Hình 1. Mô tả bài toán.

I. Giới thiệu

Việc phát hiện người đi bộ đã trở thành bài toán quan trọng với nhiều ứng dụng dựa trong Hệ thống giao thông thông minh hiện đại, an ninh và giám sát. Đây là quá trình xác định các chuyển động của con người trong dữ liệu đầu vào từ các thiết bị thu thập dữ liệu, như máy ảnh hình ảnh và cảm biến nhiệt, để hiểu biết ngữ cảnh của một cảnh. Nó giải quyết vấn đề an toàn của người tham gia giao thông, an ninh trật tự của đường xá cũng như tính an toàn của các loại xe tự động.

Việc thực hiện phát hiện người đi bộ trong môi trường thiếu ánh sáng rất thách thức do mất các khu vực mục tiêu hiệu quả. Một bộ dữ liệu benchmark đang được quan tâm gồm có ảnh visual và ảnh hồng ngoại. Hình ảnh nhìn thấy chứa rất nhiều thông tin về kết cấu và chi tiết, nhưng rất khó để phân biệt các đối tượng dưới điều kiện ánh sáng yếu. Hình ảnh hồng ngoại, không bị giới hạn bởi điều kiện ánh sáng, có thể đóng vai trò là thông tin bổ sung. Hình ảnh hồng ngoại được chụp thông qua trường nhiệt của bề mặt đối tượng, vì vậy chúng có thể làm nổi bật các mục tiêu như người đi bộ, nhưng thông tin về kết cấu bị thiếu. Do đó, bộ dữ liệu kết hợp cả hai loại ảnh với nhau để tăng tính thiết thực.

Một số phương pháp đã đạt kết quả mAP khá cao nhưng vẫn tận dụng tốt các thông tin về kết cấu của ảnh màu thấy được và ảnh hồng ngoại. Do đó, chúng em đề xuất kết hợp các kỹ thuật xử lý ảnh cơ bản là Histogram Equalization và CLAHE (áp dụng khác nhau với từng kiểu ảnh khác nhau) cùng mô hình phát hiện đối tượng mới nhất là YOLOv8 (so sánh với YOLOv5) để cải thiện độ đo mAP trên bộ dữ liệu benchmark trong bài toán này.

II. Phương pháp

A. Histogram Equalization (HE):

1) Khái niệm: Histogram Equalization là một kỹ thuật xử lý hình ảnh được sử dụng để cải thiện độ tương phản trong ảnh. Ý tưởng cơ bản là phân phối lại giá trị pixel trên toàn bức ảnh dựa trên histogram của nó.

2) Nguyên lý hoạt động::

- Tính histogram của cường độ pixel trong hình ảnh thông qua xây dựng biểu đồ thống kê phân phối cường độ pixel trong hình ảnh.
- Phân phối đồng đều các giá trị pixel phổ biến nhất (tức là những giá trị có số lượng lớn nhất trong histogram).
- Sử dụng hàm phân phối tích lũy CDF. Điều này giúp cải thiện đồng đều hóa phân phối cường độ pixel trong hình ảnh.

B. Contrast Limited Adaptive Histogram Equalization (CLAHE):

1) Khái niệm: CLAHE là một phương pháp xử lý hình ảnh được thiết kế để cải thiện độ tương phản trong ảnh một cách tự động và thích ứng. Phương pháp này là một biến thể của Histogram Equalization (HE), một kỹ thuật được sử dụng

để làm đồng đều độ tương phản của ảnh bằng cách phân phối lại giá trị pixel trên toàn bức ảnh dựa trên histogram của nó.

2) Nguyên lý hoạt động: : Hình ảnh được chia thành các khối nhỏ gọi là "tiles" (kích thước tile là 8x8).Cân bằng Histogra mỗi khối nhỏ này. Do đó, trong một khu vực nhỏ, histogram sẽ giới hạn trong một phạm vi nhỏ (trừ khi có nhiễu). Nếu có nhiễu, nó sẽ được tăng cường. Để tránh điều này, giới hạn độ tương phản được áp dụng. Nếu bất kỳ bin histogram nào vượt quá giới hạn độ tương phản được chỉ định (mặc định là 40), những pixel đó sẽ bị cắt và phân phối đồng đều vào các bin khác trước khi áp dụng cân bằng histogram. Sau đó, để loại bỏ các hiện tượng nhiễu ở biên của tile, phương pháp nội suy tuyến tính được áp dụng.

C. YOLOv5

YOLOv5 là một mô hình nhận diện đối tượng được giới thiệu vào năm 2020 bởi Ultralytics. YOLOv5 đạt được hiệu suất cao trên tập dữ liệu COCO, đồng thời hiệu quả trong quá trình inference. Mô hình gồm ba thành phần chính: backbone, neck, và head.

Backbone của YOLOv5 là Darknet53, một kiến trúc mạng mới tập trung vào việc trích xuất đặc trưng bởi sự kết hợp residual connection và filter windows.

Neck kết nối backbone với đầu, tập trung vào việc cải thiện thông tin không gian và ngữ nghĩa ở các tỉ lệ khác nhau. Một mô-đun Spatial Pyramid Pooling (SPP) loại bỏ ràng buộc kích thước cố định của mạng, từ đó loại bỏ nhu cầu biến đổi, tăng cường hoặc cắt ảnh. Sau đó là một mô-đun CSP-Path Aggregation Network, tích hợp các đặc trưng đã học trong backbone và rút ngắn đường dẫn thông tin giữa các tầng thấp và tầng cao.

Đầu của YOLOv5 bao gồm ba nhánh, mỗi nhánh dự đoán một tỷ lệ đặc trưng khác nhau. YOLOv5 sản xuất các hộp giới hạn, xác suất lớp và confidence score.

Cuối cùng, mạng sử dụng phương pháp Non-maximum Suppression (NMS) để loại bỏ các hộp giới hạn chồng lên nhau.

D. YOLOv8

YOLOv8 là phiên bản mới nhất của mô hình nhận diện đối tượng YOLO. YOLOv8 giữ nguyên kiến trúc giống như các phiên bản trước đó với một số cải tiến bao gồm một kiến trúc mạng nơ-ron mới sử dụng cả Feature Pyramid Network (FPN) và Path Aggregation Network (PAN), cùng với một công cụ gán nhãn mới giúp đơn giản hóa quá trình gán nhãn.

FPN hoạt động bằng cách dần giảm độ phân giải không gian của ảnh đầu vào trong khi tăng số lượng kênh đặc trưng. Điều này dẫn đến việc tạo ra các bản đồ đặc trưng có khả năng phát hiện đối tượng ở các tỷ lệ và độ phân giải khác nhau. Ngược lại, kiến trúc PAN tập trung nhóm các đặc trưng từ các cấp độ khác nhau của mạng thông qua skip connection. Bằng cách này, mạng có thể nắm bắt tốt hơn các đặc trưng ở nhiều tỷ lệ và độ phân giải, điều này quan trọng để nhận diện chính xác các đối tượng có kích thước và hình dạng khác nhau.

III. Thực nghiệm

A. Bộ dữ liệu

Hình 2. LLVIP Dataset.

- Bộ dữ liệu LLVIP bao gồm các cặp hình ảnh visible-infrared trong điều kiện ánh sáng yếu, được thu thập bằng một chiếc máy ảnh có khả năng chụp cả 2 loại ảnh hồng ngoại và ảnh số.
- Bộ dữ liệu bao gồm 15488 cặp hình ảnh visible-infrared từ 26 địa điểm khác nhau. Mỗi cặp hình ảnh đều có người đi bộ.
- 77.6 % bộ dữ liệu được dùng để huấn luyện và 22.4 % được dùng để kiểm tra.

B. Cấu hình thực nghiệm

Quá trình huấn luyện diễn ra trên Kaggle với 2 NVIDIA Tesla T4 với trung bình 14 tiếng đồng hồ với 200 epochs. Ảnh đầu vào có kích thước 640x640 với batch size là 8.

C. Chạy thực nghiệm

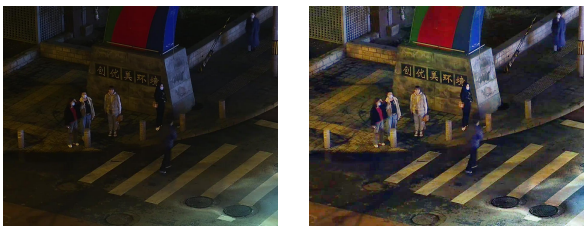
1) Tổ chức và định dạng dữ liệu: Bộ dữ liệu bao gồm hai folder: train và test. Mỗi folder sẽ gồm hai folder nhỏ là images và labels. Định dạng theo bộ dữ liệu COCO.

- folder images bao gồm *.jpg
- folder labels bao gồm *.txt

xml2txtyolov5.py sẽ hỗ trợ tổ chức và định dạng lại dữ liệu cho phù hợp với mô hình YOLO.

2) Tiền xử lý dữ liệu:

- Áp dụng CLAHE cho ảnh RGB



Hình 3. Ảnh gốc và ảnh sau tiền xử lý

Chúng tôi tiến hành tiền xử lý toàn bộ ảnh RGB với CLAHE OpenCV với các tham số:

1) clipLimit=3.0 (Ngưỡng tránh độ tương phản do nhiễu trong ảnh)

2) tileGridSize=(8,8) (kích thước cửa sổ để cân bằng histogram cục bộ)

- Áp dụng Histogram Equalization cho ảnh hồng ngoại: Chúng tôi tiến hành tiền xử lý toàn bộ ảnh hồng ngoại với Histogram Equalization OpenCV.

Hình 4. Ảnh gốc và ảnh sau tiền xử lý

3) Nhận diện người đi bộ: Trong giai đoạn 2, chúng tôi sử dụng hai mô hình YOLOv5n và YOLOv8n để quá trình huấn luyện diễn ra nhanh hơn. Sau đó, chúng tôi so sánh các kết quả thực nghiệm.

D. Kết quả thực nghiệm:

- Yolov5:

	mAP50	**mAP50-95**	infer
visible	0.898	0.513	2.9ms
infrared	0.942	0.61	2.2ms
visible (CLAHE)	0.902	0.516	69.4ms
infrared (HE)	0.941	0.607	124.9ms

- Yolov8:

	mAP50	**mAP50-95**	infer
visible	0.896	0.528	2.9ms
infrared	0.966	0.655	3.0ms
visible (CLAHE)	0.901	0.522	74.7ms
infrared (HE)	0.914	0.567	130.4ms

IV. KẾT LUẬN

Kết quả thực nghiệm visual (CLAHE) với YOLOv5n và YOLOv8n không có sự chênh lệch nhiều. Trong khi, kết quả ảnh infrared (HE) trên YOLOv5n và YOLOv8n thì không cao bằng ảnh gốc được chạy trên mô hình YOLO. Như vậy, việc áp dụng kỹ thuật xử lý ảnh CLAHE trên ảnh RGB là hợp lý và đưa ra độ chính xác cao hơn trong nhận diện người đi bộ. Còn ảnh hồng ngoại trong tập dữ liệu LLVIP nên được giữ nguyên và trực tiếp thực hiện với YOLOv8n.

TÀI LIỆU

[1] LLVIP: A Visible-infrared Paired Dataset for Low-light Vision. Available at: https://bupt-ai-cz.github.io/LLVIP/.

[2] Ultralytics Home, Ultralytics YOLOv8 Docs. Available at: https://docs.ultralytics.com/.

[3] ultralytics/yolov5 Available at: https://github.com/ultralytics/yolov5.

[4] OpenAI. (2023). ChatGPT