Sử dụng thuật toán Yolov3 nâng cao chất lượng phát hiện đối tượng cho hệ thống giám sát, bảo vệ căn cứ trên đảo

Chu Văn Hoạt*, Vũ Minh Khiêm, Vũ Xuân Vượng, Nguyễn Đình Long

Viện Tự động hóa Kỹ thuật quân sự/Viện Khoa học và Công nghệ quân sự.
*Email liên hệ: sqchuhoat@gmail.com.
Nhận bài ngày 25/8/2021; Hoàn thiện ngày 28/10/2021; Chấp nhận đăng ngày 12/12/2021.
DOI: https://doi.org/10.54939/1859-1043.j.mst.76.2021.137-143

TÓM TẮT

Cải tiến, hiện đại hóa hệ thống giám sát an ninh, bảo vệ căn cứ trên đảo là một nhiệm vụ quan trọng của Quân đội ta hiện nay. Trước đây, phương pháp học máy được áp dụng để xây dựng bộ phát hiện đối tượng, tuy nhiên kết quả quá trình thực nghiệm ở biển đảo chưa đáp ứng được yêu cầu đặt ra, tỷ lệ phát hiện nhầm đối tượng còn cao. Trong bài báo này, đề xuất thuật toán Yolov3 tiến hành tự động phát hiện đối tượng xuất hiện trong khu vực giám sát.

Từ khóa: Tự động phát hiện; Hệ thống giám sát an ninh; Yolov3.

1. ĐẶT VẤN ĐỀ

Hệ thống giám sát an ninh hiện nay thường được tích hợp camera ảnh thường và camera ảnh nhiệt, được đặt trên các bệ pan-tilt quay quét để tăng cường phạm vi giám sát. Yêu cầu đối với các hệ thống giám sát biển đảo là phải tự động phát hiện được đối tượng đột nhập ở khoảng cách xa, có thể phát hiện đối tượng trong điều kiện sóng biển, sương mù phức tạp. Đối tượng có kích thước nhỏ, ảnh nhiều nhiễu, vì thế, khó trích xuất đặc trưng, gây khó khăn cho nhiệm vụ phát hiện đối tượng.

Hiện nay, các thuật toán phát hiện đối tượng chủ yếu được chia thành hai loại: phương pháp truyền thống dựa vào các đặc trưng thủ công và phương pháp học sâu dựa vào các đặc trưng được trích xuất từ mạng nơ-ron [1]. Phương pháp truyền thống dựa vào cách lựa chọn cửa sổ trượt và các đặc trưng ảnh theo một quy luật, sử dụng loại phương pháp này các đặc trưng ảnh tính khái quát không cao, làm ảnh hưởng tới hiệu quả của thuật toán. Bài báo [2] sử dụng thuật toán Adaboost và mô hình phân tầng Cascade để ứng dụng cho hệ thống giám sát an ninh, tuy nhiên khi hệ thống được thử nghiệm ở môi trường phức tạp như biển đảo, tỷ lệ phát hiện nhầm đối tượng vẫn cao. Phương pháp học sâu sử dụng mạng nơ-ron tích chập để trích xuất các đặc trưng ảnh, có thể mô tả đối tượng rất tốt, giúp nâng cao độ chính xác của thuật toán. Dựa theo ý tưởng thiết kế của thuật toán, có thể phân thành hai loại: Thuật toán một giai đoạn và thuật toán hai giai đoạn. Thuật toán hai giai đoạn chia quá trình phát hiện đối tượng thành hai thành phần chính là tạo khu vực dự đoán, sau đó từ những khu vực này tiến này phát hiện đối tượng. Các thuật toán tiêu biểu bao gồm R-CNN [3], Fast-RCNN [4], Faster-RCNN [5]. Thuật toán một giai đoạn trực tiếp tạo ra xác suất mục tiêu và tọa độ vị trí của đối tượng chỉ thông qua một mạng nơ ron, các thuật toán điển hình như SSD [6], DSSD [7]. Hiện nay một số nghiên cứu đã sử dụng phương pháp học sâu để phát hiện đối tượng tàu thuyền [8, 9], Tuy nhiên, chỉ tập trung xử lý cho một kênh ảnh nhất định, bài báo [8] tiến hành phát hiện và phân loại tàu thuyền trên nền ảnh thường, bài báo [9] phát hiện đối tượng trên nền ảnh vệ tinh. Vì thế, để giải quyết nhiệm vụ phát hiện đối tượng cho ba kênh ảnh là ảnh thường, ảnh hồng ngoại và ảnh nhiệt, bài báo này đề xuất thuật toán Yolov3 để nâng cao chất lượng phát hiện đối tượng cho hệ thống.

Bài báo gồm có 5 phần chính, bao gồm: Đặt vấn đề; Thu thập dữ liệu; Thuật toán tự động phát hiện đối tượng; Kết quả đạt được; Kết luận.

2. THU THẬP DỮ LIỆU

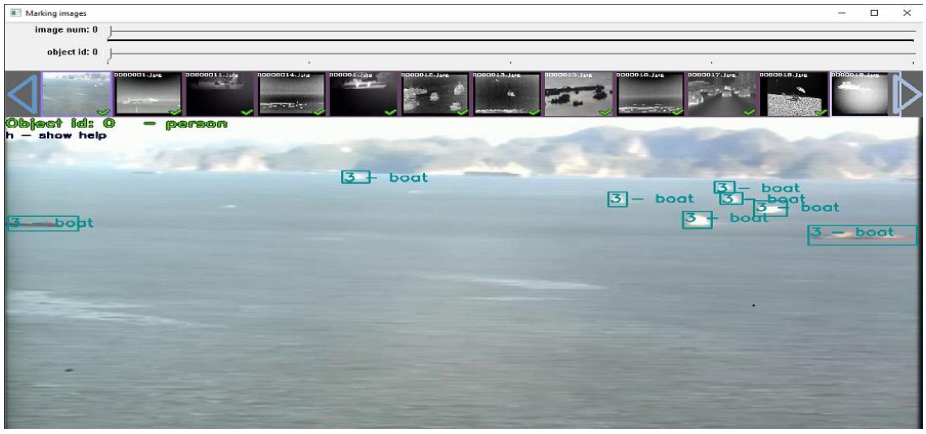
Tập dữ liệu chiếm một vị trí rất quan trọng trong sự phát triển các thuật toán phát hiện đối

tượng, tập dữ liệu đủ lớn, đa dạng là cơ sở để phát triển các thuật toán. Bài báo này sử dụng bộ dữ liệu MS COCO[10], ngoài ra, hình ảnh mục tiêu còn được thu thập từ các camera của hệ thống. COCO là một bộ dữ liệu lớn và đa dạng với hơn 200.000 hình ảnh. Tuy nhiên, bộ dữ liệu chỉ bao gồm ảnh chụp từ camera thường và thường được chụp từ khoảng cách gần, ít bị ảnh hưởng bởi nhiễu.

Đối với hệ thống giám sát, bảo vệ căn cứ trên đảo các đối tượng giám sát thường ở vị trí cách xa camera, hình ảnh bị tác động lớn bởi nhiễu. Đặc biệt đối với ảnh hồng ngoại và ảnh nhiệt, đối tượng có đường viền mờ, đặc trưng màu sắc và đặc trưng xám rất khó trích xuất, gây khó khăn cho nhiệm vụ phát hiện đối tượng. Vì thế, hình ảnh được lấy tại thực địa có vai trò rất quan trọng, giúp thuật toán có thể thích ứng được với những khó khăn thực tế của hệ thống.

Bộ dữ liệu tăng cường như hình 1 biểu diễn bao gồm ảnh nhiệt, ảnh thường và ảnh hồng ngoại với 4 lớp đối tượng: Tàu thuyền, người, xe, UAV. Trong đó, lớp tàu thuyền bao gồm 4450 ảnh với hơn 19000 đối tượng, lớp đối tượng người bao gồm 6430 ảnh với 22095 đối tượng, lớp đối tượng xe bao gồm 5024 ảnh với 20032 đối tượng, lớp đối tượng UAV bao gồm 2026 ảnh với 5061 đối tượng. Ảnh dữ liệu được thu thập ở nhiều khoảng thời gian khác nhau trong ngày, điều kiện sóng biển, thời tiết khác nhau, khoảng cách xa, đối tượng có kích thước nhỏ nhất 6*6 pixel.

Hình 1. *Tập dữ liệu.*



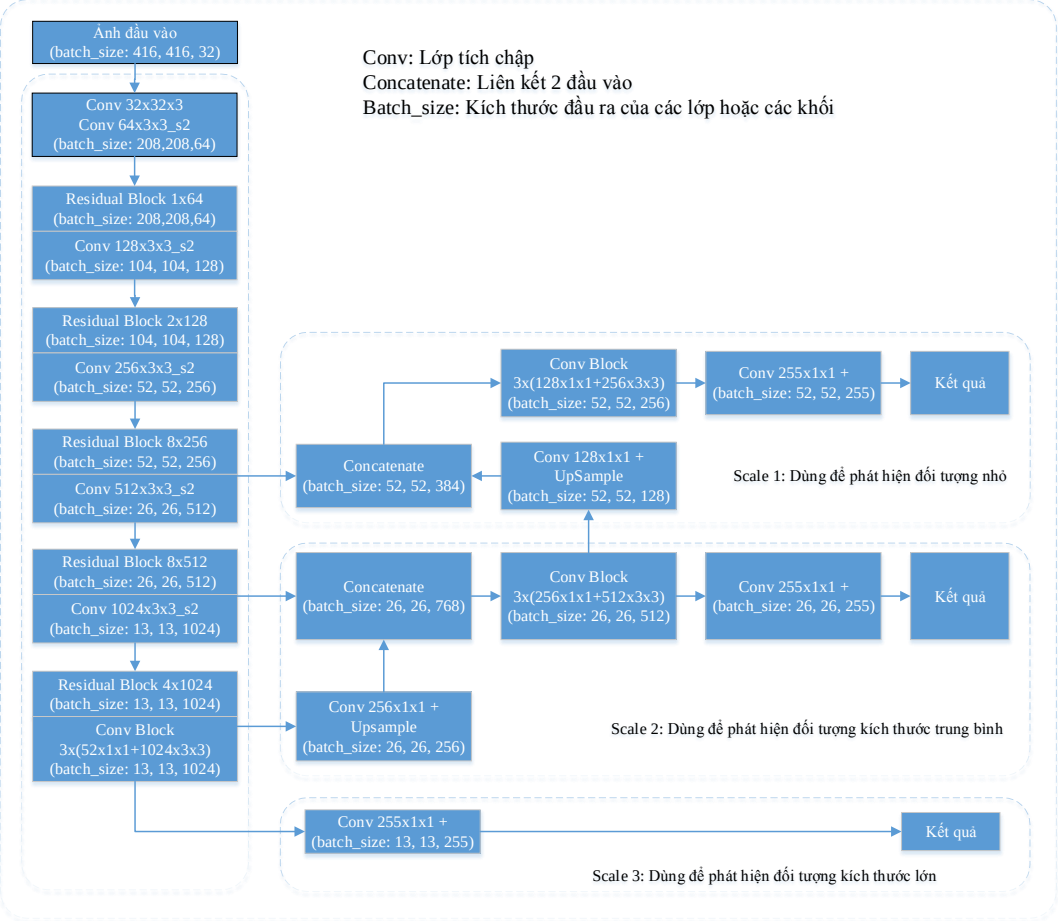
Hình 2. *Gán nhãn cho bộ dữ liệu.*

Sau khi thu thập dữ liệu cần tiến hành gán nhãn cho các đối tượng. Bài báo sử dụng phần mềm Yolo-mark để tiến hành gán nhãn cho bộ dữ liệu. Phần mềm này có chức năng lưu thông tin lớp đối tượng và thông tin vị trí được đánh dấu trong hình ảnh ở định dạng txt để tiến hành huấn luyện. Hình 2 cho thấy quá trình gán nhãn hình ảnh. Sau khi chọn lớp đối tượng và đánh dấu mục tiêu bằng hình chữ nhật, và phần mềm sẽ tạo ra văn bản nhãn định dạng txt có tên giống như tên hình ảnh.

3. THUẬT TOÁN TỰ ĐỘNG PHÁT HIỆN ĐỐI TƯỢNG

3.1. Cấu trúc thuật toán

Thuật toán Yolov3 là phương pháp chỉ sử dụng một mạng thần kinh để hoàn thành dự đoán và tính toán xác suất của các lớp đối tượng từ hình ảnh trong một lần chạy. Thuật toán sử dụng thông tin của toàn bộ bức ảnh một lần và chỉ sử dụng một mạng thần kinh duy nhất nên thuật toán được tối ưu hóa, cấu trúc đơn giản, có được hiệu suất phát hiện cao, và có thể xử lý được trong thời gian thực.



Hình 3. *Cấu trúc mô hình mạng của thuật toán.*

Sơ đồ cấu trúc mô hình mạng của Yolov3 như hình 3 biểu diễn, cấu trúc này bao gồm 53 lớp nơ ron tích chập kết nối liên tiếp, là lớp dùng để trích xuất đặc trưng của hình ảnh. Để giảm kích thước đầu ra sau mỗi lớp tích chập, tiến hành lấy mẫu xuống bằng các bộ lọc có kích thước là 2, qua đó có thể giảm số lượng tham số của mô hình, cải thiện thời gian quá trình trích xuất đặc trưng.

Các bức ảnh trước khi đưa vào mô hình, sẽ được đưa về một kích thước cố định, kích thước này là một tham số có thể thiết lập trong mô hình của thuật toán, có thể sử dụng các kích thước bao gồm 608x608, 416x416 và 304x304. Đối với mỗi kích thước đầu vào sẽ được thiết lập các lớp riêng phù hợp với kích thước của đầu vào. Để cân bằng giữa độ chính xác và tốc độ của thuật toán, bài báo sử dụng kích thước 416x416 để thiết lập kích thước đầu vào cho thuật toán. Sau khi đi qua các nơ ron tích chập thì kích thước giảm dần theo cấp số nhân là 2, sau đó, thu được một bản đồ đặc trưng có kích thước tương đối nhỏ để có thể dự đoán đối tượng trên từng ô của bản đồ đặc trưng. Đối với đầu vào 416x416, kích thước của bản đồ đặc trưng dùng để dự đoán đối tượng là 13x13, 26x26 và 52x52.

Đầu ra của thuật toán là một vector bao gồm các thành phần:

$$y^T = [p_0, x, y, w, h, p_1, p_2, ..., p_n] \quad (1)$$

Trong đó:

- p_0 là xác suất đối tượng xuất hiện trong khung;
- (x,y) là tọa độ tâm của khung, (w,h) là kích thước chiều rộng, dài của khung;
- $(p_1, p_2, ..., p_n)$ là dự báo xác suất của các lớp đối tượng.

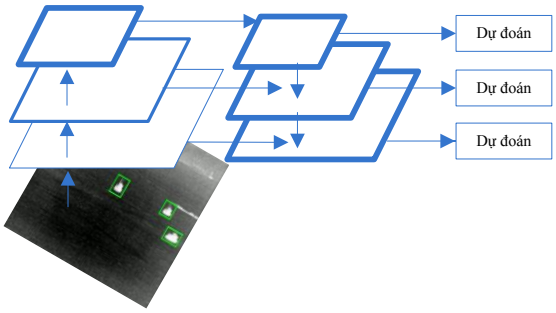
3.2. Nguyên lý dự đoán của thuật toán

Trong thuật toán Yolov3, một bức ảnh được chia thành S * S ô vuông. Nếu đối tượng cần phát hiện tồn tại trong bất kỳ một ô, thì ô vuông này có nhiệm vụ phát hiện mục tiêu. Kết quả dự đoán mỗi khung giới hạn gồm 5 phần tử (x, y, w, h, confidence), trong đó (x, y) là tọa độ tâm của khung dự đoán, (w, h) là chiều rộng và chiều cao của khung dự đoán, confidence là xác suất được định nghĩa bằng công thức sau:

$$confidence = \Pr(Object) * IOU_{pred}^{truth}(1) \quad (2)$$

Trong đó: Pr(*Object*) biểu thị trong ô vuông có chứa đối tượng hay không, có giá trị bằng 0 hoặc 1. $IOU_{pred}^{truth}(1)$ là hàm đánh giá độ chính xác kết quả dự đoán, được tính bằng tỷ lệ giữa diện tích phần giao nhau và phần hợp của khung dự đoán và khung chứa đối tượng được dán nhãn trong tập dữ liệu. Nếu *IOU*> 0.5 thì khung dự đoán được đánh giá là tốt.

Thông thường, trên một ảnh có thể chứa nhiều đối tượng có kích thước khác nhau, bộ phát hiện đối tượng cần phải phát hiện được các đối tượng ở mọi kích thước, vì thế, cấu trúc của thuật toán phải phù hợp để có thể phát hiện các đối tượng có kích thước khác nhau. Bản đồ đặc trưng được trích xuất từ mạng tích chập nông chủ yếu được sử dụng để phản ánh các đặc điểm chi tiết của đối tượng, phù hợp dùng để phát hiện các đối tượng có kích thước bé. Bản đồ đặc trưng được trích xuất từ mạng tích chập càng sâu thì có kích thước càng nhỏ, được dùng để mô tả các đặc điểm trừu tượng của đối tượng, phù hợp dùng để phát hiện các đối tượng có kích thước lớn. Như vậy lớp tích chập càng sâu, bản đồ đặc trưng có kích thước càng nhỏ, khả năng phát hiện các đối tượng có kích thước nhỏ càng khó. Cách tiếp cận của thuật toán là sử dụng bản đồ đặc trưng của mạng tích chập các lớp khác nhau để phát hiện đối tượng. Ngoài ra, thuật toán cũng thêm các liên kết giữa các lớp dự đoán, tiến hành lấy mẫu lên lớp dự đoán ở tầng sau và liên kết với lớp dự đoán ở tầng trước đó, như vậy, có thể kết hợp thông tin từ bản đồ đặc trưng ở các tầng khác nhau, tăng độ chính xác của thuật toán.



Hình 4. *Nguyên lý dự đoán của thuật toán.*

3.3. Hàm lỗi của thuật toán

Trong quá trình huấn luyện, mô hình sẽ tập trung vào những ô vuông có chứa đối tượng. Tăng

điểm phân loại chính xác đối với lớp đó lên. Sau đó, tiếp tục tập trung vào ô vuông đó, tìm vị trí khung dự đoán tốt nhất và tăng điểm vị trí của khung dự đoán đó lên, thay đổi thông tin khung dự đoán để gần đúng với nhãn đã được dán. Đối với những ô vuông không chứa đối tượng, giảm điểm tin cậy và không quan tâm đến điểm phân loại và điểm vị trí của những ô vuông này.

Hàm lỗi dùng để tính giá trị lỗi cho khung dự đoán so với vị trí thực tế từ tập dữ liệu. Bao gồm các thành phần: Độ lỗi của việc dự đoán loại nhãn của đối tượng và tính toán xác suất, độ lỗi vị trí và độ lớn của khung dự đoán. Hàm lỗi được tính như sau:

$$\zeta_{loc} = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left[\left(x_i - \hat{x}_i \right)^2 + \left(y_i - \hat{y}_i \right)^2 + \left(\sqrt{w}_i - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h}_i - \sqrt{\hat{h}_i} \right)^2 \right] \quad (3)$$

$$\zeta_{cls} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} \left(1_{ij}^{obj} + \lambda_{noobj} \left(1 - 1_{ij}^{obj} \right) \right) \left(C_{ij} - \hat{C}_{ij} \right)^2 + \sum_{i=0}^{S^2} \sum_{c \in C} 1_i^{obj} \left(p_i(c) - \hat{p}_i(c) \right)^2 \quad (4)$$

$$\zeta = \zeta_{loc} + \zeta_{cls} \quad (5)$$

Trong đó:

- ζ_{loc} là hàm lỗi của vị trí và độ lớn khung dự đoán so với thực tế khung được dán nhãn;
- ζ_{cls} là hàm lỗi của việc dự đoán loại nhãn của đối tượng và tính toán xác suất;
- 1_i^{obj} : Hàm indicator có giá trị 0 hoặc 1, nhằm xác định xem ô *i* có chứa đối tượng hay không. Bằng 1 nếu chứa đối tượng và bằng 0 nếu không chứa đối tượng;
- 1_{ij}^{obj} : Cho biết khung thứ *j* của ô *i* có chứa đối tượng hay không, bằng 1 nếu có chứa đối tượng, và bằng 0 nếu không chứa đối tượng;
- C_{ij} : Điểm tin cậy của ô *i;*
- C : Tập hợp tất cả các lớp đối tượng;
- $p_i(c)$: Xác suất có điều kiện của lớp $c \in C$ tại ô *i* mà mô hình dự đoán.

4. KẾT QUẢ ĐẠT ĐƯỢC

Hệ thống giám sát an ninh, bảo vệ căn cứ sử dụng ba kênh là ảnh thường và ảnh hồng ngoại và ảnh nhiệt. Video dùng để thử nghiệm hiệu quả của phương pháp đề xuất được quay bởi hệ thống giám sát tại khu vực khảo sát thực tế. Thuật toán Yolov3 được huấn luyện và thử nghiệm trên nền tảng máy tính hệ điều hành window, Intel i5-10400F, 2.9GHz, RAM 16GB, card đồ họa RTX 3060, ngôn ngữ lập trình C, sử dụng thư viện OpenCV 3.4.0, thư viện CUDA11.1 và CUDNN8.0. Máy tính được kết nối với bệ pan-tilt quay quét tích hợp camera ảnh nhiệt và camera thường. Các tham số của thuật toán được thiết lập như bảng 1 thể hiện.

Bảng 1. *Tham số của thuật toán.*

Tham số	Batch	Learning_rate	momentum	Decay	Max iteration
Giá trị	16	0.0025	0.95	0.0005	200000

Để đánh giá hiệu quả của thuật toán và tác động của bộ dữ liệu tăng cường, bài báo sử dụng thuật toán Yolov3 khi được huấn luyện bởi bộ dữ liệu COCO, thuật toán SSD và thuật toán Fast-RCNN để so sánh với hiệu quả với mô hình mà bài báo đề xuất. Kết quả so sánh như hình 5 và bảng 2 thể hiện.

Hình 5 và bảng 2 cho thấy, khi thuật toán Yolov3 được huấn luyện bởi bộ dữ liệu COCO, đối với nền ảnh thường và điều kiện ánh sáng tốt, thuật toán vẫn có hiệu quả phát hiện tốt, tuy nhiên, đối với điều kiện ánh sáng yếu, ảnh hồng ngoại, ảnh nhiệt thì hiệu quả của thuật toán giảm, độ chính xác của thuật toán là 0.55. Thuật toán SSD bỏ sót nhiều đối tượng, đặc biệt là các đối tượng có kích thước nhỏ và các đối tượng bị chồng lên nhau, thuật toán có độ chính xác là 0.76

và tốc độ xử lý là 21frame/s. Thuật toán Fast-RCNN có độ chính xác là 0.84, tuy nhiên, chi phí thời gian của thuật toán cao, tốc độ xử lý là 13frame/s. Thuật toán Yolov3 khi được huấn luyện bởi bộ dữ liệu COCO tăng cường đứng đầu về độ chính xác với độ chính xác là 0.94, thuật toán có tốc độ xử lý nhanh, đối với kích thước đầu vào thiết lập là 416x416 tốc độ xử lý lên tới 59frame/s. Có thể thấy, tăng cường bộ dữ liệu giúp cho mô hình sau khi huấn luyện có thể thích ứng tốt hơn với tình hình thực tế tại thực địa và thuật toán Yolov3 có ưu thế khi đối tượng có kích thước nhỏ, phù hợp yêu cầu về độ chính xác và xử lý thời gian thực của hệ thống giám sát, bảo vệ căn cứ trên đảo.

(a) Kết quả thử nghiệm thuật toán Yolov3 với bộ dữ liệu COCO.

(b) Kết quả thử nghiệm thuật toán SSD với bộ dữ liệu COCO tăng cường.

(c) Kết quả thử nghiệm thuật toán Fast-RCNN với bộ dữ liệu COCO tăng cường.

(d) Kết quả thử nghiệm của thuật toán Yolov3 với bộ dữ liệu COCO tăng cường.

Hình 5. *Kết quả thử nghiệm.*

Bảng 2. *Độ chính xác và tốc độ của các thuật toán.*

Thuật toán	AP Tàu thuyền	AP Người	AP xe	AP UAV	mAP	Fps Frame/s
YOLOv3-COCO	0.47	0.59	0.63	0.53	0.55	59
SSD	0.76	0.75	0.80	0.75	0.76	21
Fast-RCNN	0.85	0.83	0.83	0.84	0.84	13
YOLOv3	0.95	0.92	0.93	0.93	0.94	59

5. KẾT LUẬN

Bài báo đề xuất phương pháp Yolov3 để nâng cao chất lượng tự động phát hiện trong hệ thống giám sát. Kết quả thử nghiệm cho thấy phương pháp đề xuất có thể phát hiện đối tượng trong điều kiện phức tạp như mưa, sương mù, ánh sáng yếu, nhiễu lớn. Thông qua thực nghiệm cho thấy tính khả thi của phương pháp đề xuất. Do đó, thuật toán này có thể được áp dụng cho lĩnh vực công nghiệp hoặc quân sự như phát hiện và giám sát đối tượng trong các bến tàu hải

cảng, cứu nạn hàng hải và giao thông hàng hải. Đặc biệt, trong lĩnh vực quân sự không chỉ có thể ứng dụng vào việc phát hiện địch, mà còn dùng cho các hệ thống vũ khí để nâng cao độ chính xác của các thiết bị vũ khí, nâng cao hiệu quả chiến đấu của quân đội ta.

Lời cảm ơn: *Nhóm tác giả cảm ơn sự đóng góp ý kiến của phòng KHQS, Bộ Tham mưu Hải quân giúp hoàn thiện bài báo này. Nghiên cứu này được hỗ trợ từ nhiệm vụ cấp bộ mã số 2020.85.24.*

TÀI LIỆU THAM KHẢO

[1]. Kim C, Lee Y, Park J et al. *"Diminishing unwanted objects based on object detection using deep learning and image inpainting,"* 2018 International Workshop on Advanced Image Technology (IWAIT), 2018, 1-3.

[2]. Chu V H, Vũ M K. *"Xây dựng thuật toán tự động phát hiện đối tượng trên nền ảnh động cho bệ quay quét giám sát an ninh,"* Tạp chí Nghiên cứu KH&CN quân sự, Số Đặc san TĐH, 04 – 2019.

[3]. Uijlings J R R, van de Sande K E A, Gevers T, et al. *"Selective Search for Object Recognition,"* Int J Comput Vis 104(2013), 154–171.

[4]. Girshick R. *"Fast r-cnn,"* Proceedings of the IEEE international conference on computer vision, 2015, 1440-1448.

[5]. Ren S, He K, Girshick R, et. al. *"Faster r-cnn: Towards real-time object detection with region proposal networks,"* preprint arXiv:1506.01497, 2015.

[6]. Liu W, Anguelov D, Erhan D, et al. *"Ssd: Single shot multibox detector,"* European conference on computer vision, 2016, 21-37.

[7]. Fu C Y, Liu W, Ranga A, et al. *"Dssd: Deconvolutional single shot detector,"* arXiv preprint arXiv:1701.06659, 2017.

[8]. Cui H, Yang Y, Liu M, et al. *"Ship detection: an improved YOLOv3 method,"* OCEANS 2019-Marseille, 2019: 1-4.

[9]. Wang Q, Shen F, Cheng L, et al. *"Ship detection based on fused features and rebuilt YOLOv3 networks in optical remote-sensing images,"* International Journal of Remote Sensing, 2021, 42(2): 520-536.

[10]. Russakovsky O, Deng J, Su H, et al. *"Imagenet large scale visual recognition challenge,"* International Journal of Computer Vision, 2015, 115(3): 211-252.

ABSTRACT

USING THE YOLOV3 METHOD ENHANCED THE QUALITY OF OBJECT DETECTING FOR SURVEILLANCE SYSTEM, PROTECTION OF THE ISLAND FACILITIES

Improvement and modernization of the security surveillance system, protecting bases on the island is a vital duty to our military nowadays. Previously, machine learning methods have been used to construct object detectors, but the results of the experimental process in the ocean and islands did not meet the specified requirements, and the false detection rate was still high. In this paper, Yolov3 algorithm is proposed to automatically detect objects appearing in the surveillance area.

Keywords: Auto-detection; Security monitoring system; Yolov3.