Một thuật toán bắt bám đối tượng sử dụng đa nguồn tín hiệu ảnh

Đào Vũ Hiệp*, Trần Quang Đức

Trường Công nghệ thông tin và truyền thông, Đại học Bách Khoa Hà Nội.
*Email: hiep.DVNCS18032@sis.hust.edu.vn
Nhận bài: 20/9/2022; Hoàn thiện: 24/10/2022; Chấp nhận đăng: 12/12/2022; Xuất bản: 28/12/2022.
DOI: https://doi.org/10.54939/1859-1043.j.mst.84.2022.32-41

TÓM TẮT

Hiện nay, có nhiều thuật toán bắt bám đối tượng đạt hiệu quả khá tốt trên ảnh nhìn thấy (visible hay RGB images) như KCF, CSRDCF, SiamFC, SiamRPN, ATOM, SiamDW_ST, DiMP. Tuy nhiên, các phương pháp này bị giảm chất lượng khi điều kiện chiếu sáng của môi trường bị kém đi. Các thuật toán sử dụng kết hợp ảnh nhìn thấy và ảnh nhiệt (thermal hay TIR images) như FSRPN, SiamDW_T, mfDiMP đã chứng minh hiệu năng bắt bám đối tượng được nâng cao đáng kể so với khi chỉ dùng riêng ảnh nhìn thấy hoặc ảnh nhiệt. Trong bài báo sẽ trình bày kết quả nghiên cứu một thuật toán bắt bám đối tượng sử dụng đa nguồn ảnh với trọng số xác định theo điều kiện môi trường. Kết quả thử nghiệm trên bộ dữ liệu VOT-RGBT cho thấy, thuật toán này có chỉ số EAO đạt 0,423, cao hơn so với một số thuật toán bắt bám đối tượng phổ biến hiện nay và đạt tốc độ khoảng 13 khung hình/giây trong điều kiện phần cứng phổ dụng.

Từ khoá: Mạng nơ-ron tích chập; Bắt bám đối tượng; Bộ lọc tương quan phân biệt; Kết hợp đa nguồn tín hiệu.

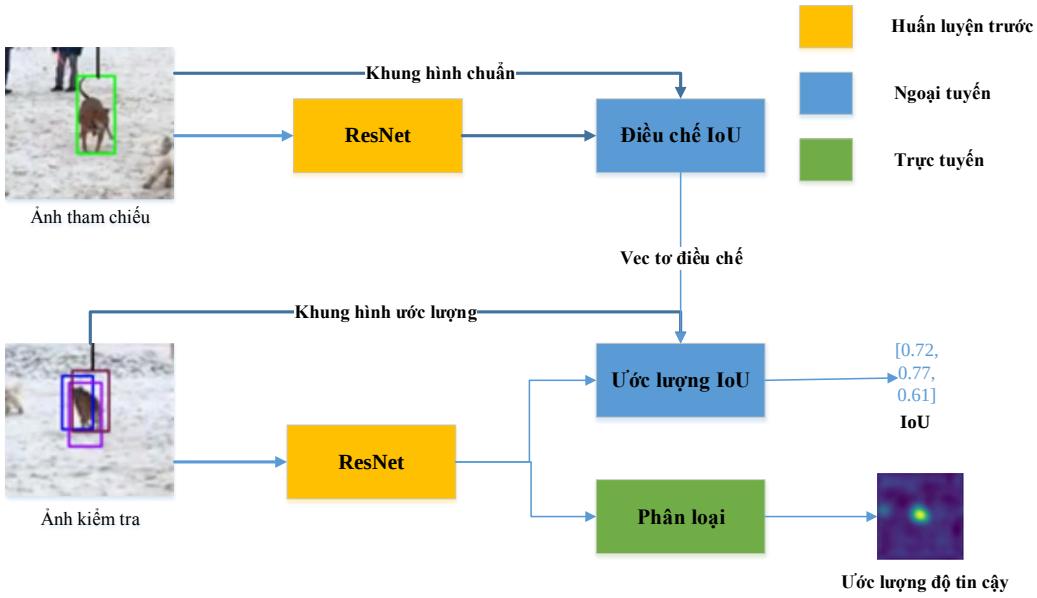
1. MỞ ĐẦU

Bắt bám đối tượng là bài toán cơ bản, quan trọng trong lĩnh vực thị giác máy tính và được sử dụng rộng rãi trong nhiều ứng dụng giám sát dựa trên video như: giám sát biên giới, phát hiện hoạt động khả nghi; giám sát giao thông, phân tích và trích xuất thông tin về giao thông công cộng, tắc đường, du lịch hoặc an ninh,… cũng như trong điều khiển tự động như: điều khiển bắt bám mục tiêu, xe tự lái và nhiều ứng dụng khác. Khó khăn chính của bài toán này là chỉ có được thông tin của đối tượng tại khung hình đầu tiên và sự thay đổi của điều kiện chiếu sáng (illumination change), sự che khuất bởi đối tượng khác (occlusion) hay sự chuyển động (motion). Ngoài ra, tốc độ xử lý cũng là một điều kiện quan trọng để áp dụng các thuật toán bắt bám đối tượng vào thực tế [1].

Ngày nay, có một số thuật toán bắt bám đối tượng trên ảnh nhìn thấy đạt hiệu năng khá tốt và đã được dùng trong các ứng dụng như KCF[3], CSRDCF[4], SiamFC[6], SiamRPN[7], ATOM[8], SiamDW[9], DiMP[10]. Các thuật toán bắt bám đối tượng này được phát triển theo hai hướng tiếp cận. Hướng thứ nhất dựa trên DCF (Discriminative Correlation Filters - Bộ lọc tương quan phân biệt), trong đó, trọng số của DCF được ước lượng trên miền Fourier bởi đối tượng tại khung hình đầu tiên [2]. Trong các khung hình tiếp theo, trọng số của bộ lọc được cập nhật khi định vị được đối tượng để phát hiện đối tượng chính xác hơn ở các khung hình tiếp theo. Thuật toán KCF nâng cao độ chính xác nhờ tăng số lượng đối tượng để ước lượng DCF bằng việc sử dụng ma trận tuần hoàn (Circulant Matrices) trên miền Fourier [3]. Trong khi đó, thuật toán CSRDCF nâng cao hiệu năng bằng việc kết hợp nhiều lớp DCF được ước lượng bởi nhiều đặc trưng của đối tượng như ảnh đa mức xám (Grayscale), HoG (Histogram of Gradient), ColorNames [4].

Gần đây, nhờ sự phát triển của các kỹ thuật học sâu (Deep Learning), các thuật toán có xu hướng sử dụng các đặc trưng tích chập nhiều lớp [5] để giảm sai số bám bắt đối tượng. Tiêu biểu theo hướng tiếp cận này có thể kể đến phương pháp SiamFC, SiamRPN dựa trên mạng nơ-ron so sánh (Similarity Learning - còn gọi là Siamese) với khả năng ước lượng chính xác vị trí của đối tượng. Hiện nay, các thuật toán bắt bám đối tượng trên ảnh cho hiệu năng cao kết hợp cả hai hướng tiếp cận này thành hai bước trong một thuật toán: (i) bước phân loại (Classification) sử

dụng DCF với đặc trưng tích chập để bóc tách đối tượng và phát hiện các vị trí có khả năng là đối tượng trong khung hình mới; (ii) bước ước lượng (Estimation) để từ các vị trí có khả năng là đối tượng trong khung hình mới, ước đoán vị trí chính xác của đối tượng. Thuật toán đầu tiên theo hướng này là thuật toán ATOM (Accurate Tracking by Overlap Maximization), trong đó, bước ước lượng vị trí xác định qua độ chồng lấn (Overlap) thay vì ước lượng trực tiếp vị trí như các các thuật toán SiamFC, SiamRPN [8]. Sơ đồ thuật toán ATOM mô tả tại hình 1. Sau đó, thuật toán SiamDW_ST nâng cao độ sâu của đặc trưng tích chập bằng cách sử dụng ResNet-50 thay vì ResNet-18 kết hợp một số kỹ thuật để nâng cao tốc độ tính toán [9]. Trong khi đó, thuật toán DiMP cải tiến bước phân loại để đạt độ chính xác cao hơn [10].



Hình 1. *Sơ đồ thuật toán ATOM.*

Ảnh nhìn thấy sẽ cung cấp nhiều thông tin về đối tượng như màu sắc, hình dạng, kích thước chính xác,... nên các thuật toán bắt bám đối tượng trên ảnh nhìn thấy sẽ hoạt động tốt với điều kiện môi trường tốt. Tuy nhiên, hiệu năng sẽ suy giảm đáng kể khi điều kiện môi trường kém như trong các trường hợp ánh sáng yếu, trời mưa hay sương mù. Trong khi đó, ảnh được tạo từ tín hiệu hồng ngoại bước sóng dài (8÷14 µm), gọi tắt là ảnh nhiệt, không bị ảnh hưởng trong các trường hợp này [12] (hình 2). Mặt khác, các loại thiết bị để tạo ảnh từ tín hiệu này cũng có thể tiếp cận dễ dàng với chi phí không cao. Vì vậy, hướng nghiên cứu bắt bám đối tượng sử dụng kết hợp ảnh nhìn thấy với ảnh nhiệt bắt đầu được chú ý thời gian gần đây [1].

Các thuật toán bắt bám đối tượng sử dụng kết hợp ảnh nhìn thấy và ảnh nhiệt như FSRPN (phát triển từ SiamRPN), SiamDW_T (phát triển từ SiamDW_ST) hay mfDiMP (phát triển từ DiMP) đạt hiệu năng cao hơn đáng kể so với các thuật toán chỉ sử dụng ảnh nhìn thấy hoặc ảnh nhiệt [12]. Tuy nhiên, khi sử dụng nhiều nguồn ảnh, các phương pháp này đều chưa xem xét tới các điều kiện môi trường như độ chiếu sáng (khi độ chiếu sáng yếu sử dụng ảnh nhiệt sẽ tốt hơn ảnh nhìn thấy) và nền nhiệt môi trường (khi nền nhiệt cao độ tương phản của ảnh nhiệt sẽ thấp hơn ảnh nhìn thấy). Vì vậy, trong nghiên cứu này sau khi phân tích, lựa chọn được phương án kết hợp ảnh nhìn thấy và ảnh nhiệt trong bài toán bắt bám đối tượng, chúng tôi sẽ đề xuất một thuật toán bắt bám đối tượng kết hợp ảnh nhìn thấy và ảnh nhiệt có trọng số phụ thuộc vào hai điều kiện môi trường nêu trên. Trong đó, độ chiếu sáng của môi trường có thể được xác định thông qua độ nhiễu của ảnh nhìn thấy, nền nhiệt của môi trường có thể được xác định thông qua cường độ sáng của ảnh nhiệt như đã mô tả trong [13]. Tuy nhiên, trong bài toán bắt bám đối tượng có thể xác định được vùng có khả năng là đối tượng nên thay vì xác định nền nhiệt của

môi trường, có thể xác định trực tiếp độ tương phản của đối tượng trên ảnh nhiệt. Do đó, chúng tôi đề xuất xác định trọng số dựa trên độ nhiễu của ảnh nhìn thấy và độ tương phản của đối tượng trên ảnh nhiệt. Cuối cùng, phương pháp đề xuất được thử nghiệm, đánh giá với bộ dữ liệu VOT-RGBT và kết luận.

Hình 2. *So sánh hiệu năng bắt bám đối tượng trên các nguồn ảnh khác nhau. Màu vàng kết quả chuẩn, màu xanh là bắt bám sử dụng ảnh nhìn thấy, màu đỏ sử dụng kết hợp ảnh nhìn thấy và ảnh nhiệt.*

Trong phần còn lại của bài báo này sẽ trình bày các nội dung sau:

- Phân tích, lựa chọn phương án kết hợp ảnh nhìn thấy và ảnh nhiệt trong bài toán bắt bám đối tượng tại mục 2.

- Đề xuất thuật toán kết hợp ảnh nhiệt và ảnh nhìn thấy sử dụng trọng số trong bắt bám đối tượng tại mục 3.

- Thử nghiệm và đánh giá kết quả tại mục 4.

- Cuối cùng là kết luận và hướng phát triển tại mục 5.

2. PHÂN TÍCH, LỰA CHỌN PHƯƠNG ÁN KẾT HỢP ẢNH NHÌN THẤY VÀ ẢNH NHÌN THẤY TRONG BÀI TOÁN BẮT BÁM ĐỐI TƯỢNG

2.1. Lựa chọn thuật toán bắt bám đối tượng kết hợp ảnh nhìn thấy và ảnh nhiệt

Như đã đề cập tại mục 1, các thuật toán bắt bám đối tượng sử dụng kỹ thuật học sâu kết hợp ảnh nhìn thấy và ảnh nhiệt có hiệu năng khá tốt và đang được sử dụng có thể kể đến như CISRDCF, FSRPN, SiamDW_T hay mfDiMP [1]. Để có thể sử dụng làm nền tảng phát triển thuật toán bắt bám đối tượng sử dụng đa nguồn tín hiệu ảnh có kết hợp trọng số phụ thuộc vào môi trường, thuật toán phải có hiệu năng tốt và tốc độ phù hợp.

Các thuật toán nêu trên được đo lường hiệu năng dựa trên phương pháp trong cuộc thi VOT [1]. Cuộc thi này cung cấp bộ dữ liệu và các công cụ đo lường các thuật toán bắt bám đối tượng phổ biến nhất hiện nay. Các phương pháp đo lường được sử dụng trong VOT là chỉ số độ chính xác (Accuracy, viết tắt là A), chỉ số độ bền vững (Robustness, viết tắt là R) và chỉ số độ chồng lấn trung bình kỳ vọng (EAO - Expected Average Overlap) là chỉ số tổng hợp của hai chỉ số trên. Độ chính xác là độ chồng lấn trung bình (Average Overlap) giữa khung dự đoán và khung chính xác trong quá trình bắt bám thành công qua các chuỗi ảnh. Độ chồng lấn được tính bằng tỷ lệ

diện tính giao nhau và diện tích hợp nhau của hai khung hình (IoU - Intersection over Union). Độ bền vững đo số lần thuật toán bắt bám đối tượng bị mất bám mục tiêu trung bình đối với chuỗi ảnh có chiều dài trung bình trong tập dữ liệu. Độ chồng lấn trung bình kỳ vọng được xác định bằng độ chồng lấn trung bình, khi thuật toán bị mất bám độ chồng lấn bằng 0 [1]. Do hiệu năng các thuật toán nêu trên được xác định trong các môi trường khác nhau, chúng tôi sử dụng dữ liệu và bộ công cụ trong VOT và cài đặt lại các thuật toán trên máy tính có CPU Ryzen5 3600 và GPU 1080Ti. Kết quả đo lường độ chính xác, độ bền vững và độ chồng lấn trung bình kỳ vọng được mô tả tại bảng 1 cho thấy thuật toán DiMP có hiệu năng cao nhất, thuật toán SiamDW_ST có hiệu năng cao thứ hai và hai thuật toán này có hiệu năng bắt bám đối tượng cao hơn nhiều so với các thuật toán khác. Tốc độ xử lý (khung hình/giây) của các thuật toán được mô tả tại bảng 2 cho thấy thuật toán SiamDW_T (19,2 khung hình/giây) có tốc độ xử lý cao hơn nhiều so với thuật toán mfDiMP (8,61 khung hình/giây).

***Bảng 1**. Kết quả đo lường EAO, A, R của một số thuật toán bắt bám đối tượng đa nguồn ảnh.*

Phương pháp	**EAO**	**A**	**R**
CISRDCF	0,346	0,502	0,412
FSRPN	0,387	0,630	0,372
mfDiMP	0,411	0,602	0,343
SiamDW_T	**0,413**	0,589	0,329

***Bảng 2.** Tốc độ xử lý (khung hình/giây) của một số thuật toán bắt bám đối tượng đa nguồn ảnh.*

Phương pháp	CISRDCF	FSRPN	mfDiMP	SiamDW_T
Tốc độ	32,1	38,1	8,61	19,2

Từ các so sánh ở trên có thể thấy, lựa chọn thuật toán SiamDW_T để tiếp tục nghiên cứu tích hợp trọng số phụ thuộc vào điều kiện môi trường là phù hợp vì đây là thuật toán có hiệu năng bắt bám đối tượng tốt và tốc độ phù hợp.

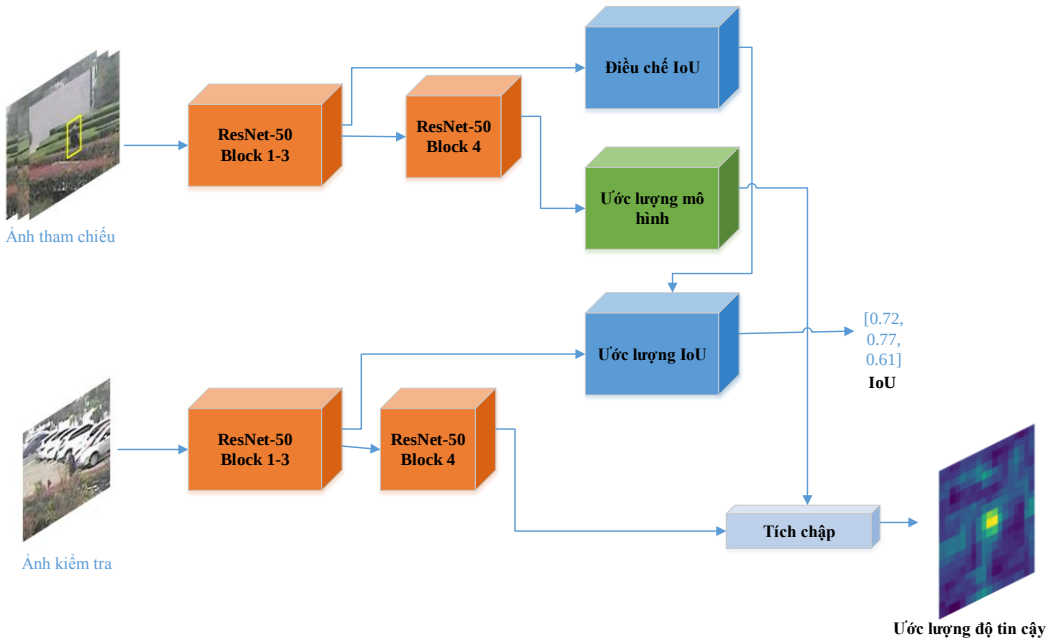
2.2. Lựa chọn mức kết hợp ảnh nhìn thấy và ảnh nhiệt trong bài toán bắt bám đối tượng

Như đã đề cập tại mục 1, thuật toán SiamDW_T được phát triển từ thuật toán SiamDW_ST được xây dựng từ kiến trúc được đề xuất trong thuật toán ATOM và sử dụng một số kỹ thuật nâng cao tốc độ xử lý, giảm sai số để mở rộng từ mạng ResNet-18 lên mạng ResNet-50. Kiến trúc của thuật toán SiamDW_ST được mô tả tại hình 3. Thuật toán SiamDW_ST cũng như ATOM bao gồm bước phân loại và bước ước lượng. Đối tượng được trích chọn đặc trưng thông qua mạng ResNet-50 đã được huấn luyện với bộ dữ liệu ImageNet. Bước phân loại lấy đặc trưng từ sau Block 4, bước ước lượng lấy đặc trưng từ sau Block 3.

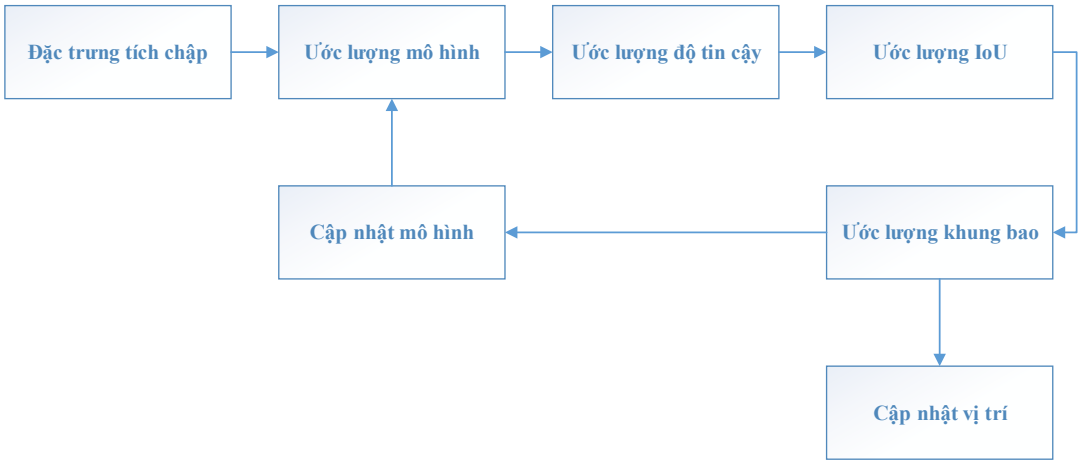
Đối với bước phân loại, từ đặc trưng của đối tượng ở khung hình trước thực hiện ước lượng mô hình (Model Prediction) chính là ước lượng trọng số của 02 lớp kết nối hoàn toàn (Fully Connected) như mô tả trong [8]. Trong quá trình bắt bám đối tượng, tương tự như các thuật toán dựa trên DCF, mô hình sẽ được tích chập với đặc trưng của khung hình hiện tại để tìm ra các vị trí có khả năng là đối tượng. Sau khi xác định được vị trí của đối tượng trong khung hình hiện tại, tiến hành cập nhật mô hình (Update Model Prediction) để sử dụng cho khung hình tiếp theo.

Đối với bước ước lượng, đặc trưng của khung hình trước $\boldsymbol{x_0}$ và vùng bao chuẩn (ground truth) $\boldsymbol{B_0}$ được điều chế IoU (IoU Modulation) để tạo ra vector điều chế $\boldsymbol{c(x_0, B_0)}$. Trong quá trình bắt bám đối tượng, đối với các vị trí có khả năng là đối tượng trích xuất ra các mảnh đặc trưng $\boldsymbol{z(x, B)}$, trong đó $\boldsymbol{x}$ là đặc trưng của khung hình hiện tại và B là các vùng bao dự đoán của các vị trí có khả năng là đối tượng. Từng mảnh đặc trưng $\boldsymbol{z(x, B)}$ này kết hợp với vector điều chế $\boldsymbol{c(x_0, B_0)}$ để ước lượng IoU (IoU Estimation) bằng công thức 1 với $\boldsymbol{g_{1,2,3}}$ là 03 lớp kết nối hoàn toàn (Fully Connected) để từ đặc trưng tích chập cho độ ra độ chồng lấn (IoU) [8]. Sơ đồ các bước thuật toán SiamDW_ST được mô tả tại hình 4.

$$IoU(B) = g_{1,2,3}(c(x_0, B_0) \cdot z(x, B)) \quad (1)$$



Hình 3. *Mô tả kiến trúc SiamDW_ST.*



Hình 4. *Sơ đồ các bước thuật toán SiamDW_ST.*

Để kết hợp ảnh nhìn thấy (gọi là RGB) và ảnh nhiệt (gọi là TIR), có thể kết hợp ở các mức khác nhau. Trong nghiên cứu này, tương tự như các mức kết hợp trong [12], các mức kết hợp đã thử nghiệm gồm: mức điểm ảnh (kết hợp từ ảnh đầu vào), mức kết quả và các mức đặc trưng: đặc trưng tích chập, ước lượng/cập nhật mô hình (gọi tắt là mức mô hình) và điều chế/ước lượng IoU (gọi tắt là mức IoU). Đối với mức IoU, ngoài thử nghiệm kết hợp dạng cộng Hadamard (cộng từng phần tử) như [12] được mô tả tại công thức 2, chúng tôi thử nghiệm nhân Hadamard (nhân từng phần tử) như mô tả tại công thức 3.

$$IoU(B) = FC_2\left(FC_1\left(g_{1,2}\left(c^{RGB} \cdot z^{RGB}\right)\right) \oplus FC_1\left(g_{1,2}\left(c^{TIR} \cdot z^{TIR}\right)\right)\right) \quad (2)$$

$$IoU(B) = FC_2\left(FC_1\left(g_{1,2}\left(c^{RGB} \cdot z^{RGB}\right)\right) \odot FC_1\left(g_{1,2}\left(c^{TIR} \cdot z^{TIR}\right)\right)\right) \quad (3)$$

Trong đó, $g_{1,2}(c^{RGB} \cdot z^{RGB})$ thể hiện phép ước lượng IoU chỉ bao gồm 02 lớp kết nối hoàn toàn. FC_1 là một lớp kết nối hoàn toàn để giảm số lượng chiều của đặc trưng từ 512 xuống còn 256 và FC_2 là một lớp kết nối hoàn toàn để từ 256 chiều của đặc trưng tính ra độ chồng lấn (IoU).

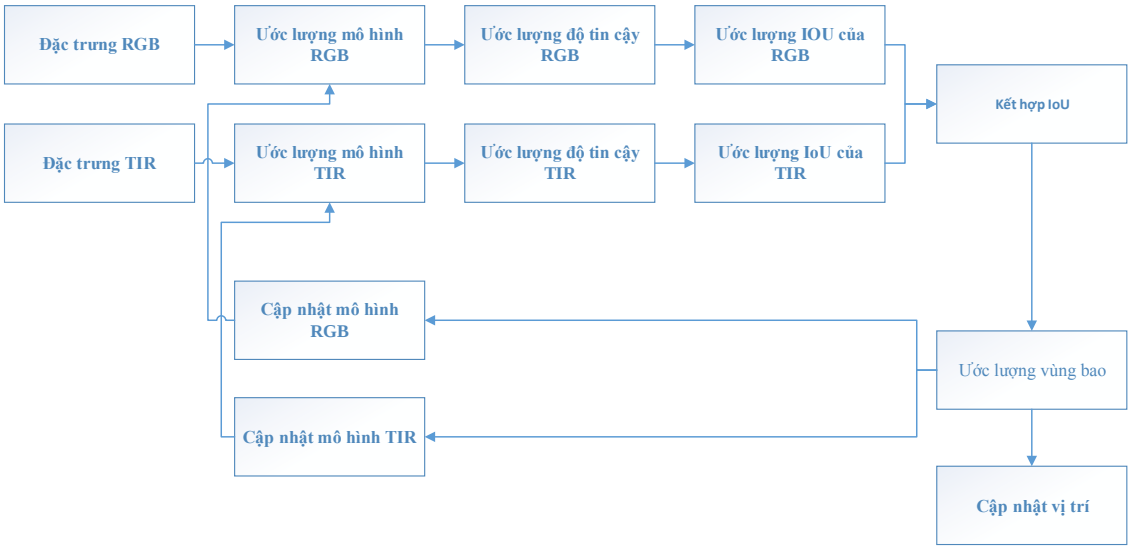
Bảng 3 mô tả kết quả của các thử nghiệm kết hợp cho thấy kết hợp ở mức ước lượng IoU dạng nhân Hadamard cho kết quả tốt nhất. Do đó, ở các bước tiếp theo sẽ tiếp tục nghiên cứu ở mức kết hợp này.

Bảng 3. *Kết quả ở các mức kết hợp ảnh nhìn thấy (RGB) và ảnh nhiệt (TIR).*

Mức kết hợp	Đặc trưng	Mô hình	IoU	Kết quả	EAO	A	R
Riêng biệt	TIR	TIR	TIR	TIR	0.257	0.354	0.898
	RGB	RGB	RGB	RGB	0.256	0.446	0.761
Mức điểm ảnh	RGB+TIR	RGBT	RGBT	RGBT	0.345	0.552	0.381
Mức kết quả	RGB/TIR	RGB/TIR	RGB/TIR	RGB+TIR	0.349	0.554	0.391
Các mức đặc trưng	RGB/TIR	RGB+TIR	RGB+TIR	RGBT	0.391	0.602	0.368
	RGB/TIR	RGB/TIR	RGB+TIR	RGBT	0.413	**0.589**	0.329
	RGB/TIR	RGB/TIR	RGB*TIR	RGBT	**0.419**	0.587	**0.299**

3. ĐỀ XUẤT THUẬT TOÁN KẾT HỢP ẢNH NHÌN THẤY VÀ ẢNH NHIỆT SỬ DỤNG TRỌNG SỐ TRONG BẮT BÁM ĐỐI TƯỢNG

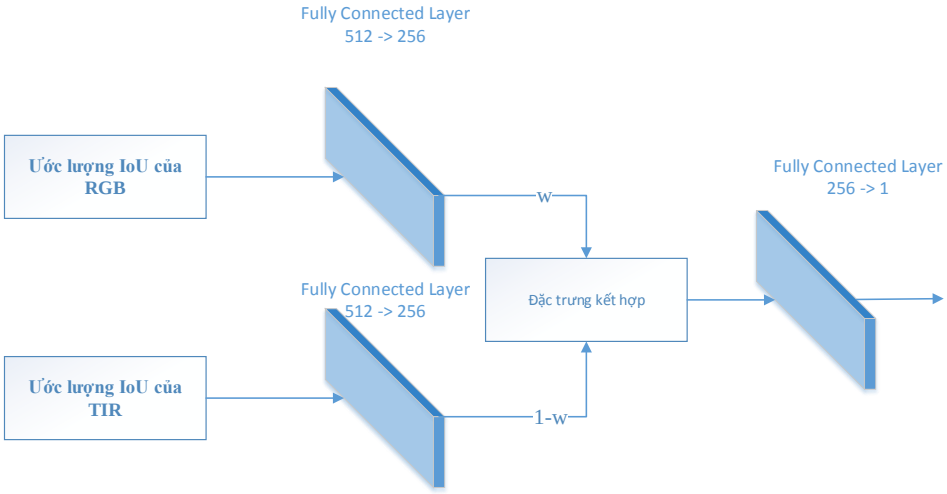
3.1. Phương án kết hợp có trọng số ảnh nhiệt và ảnh nhìn thấy trong bắt bám đối tượng



Hình 5. *Sơ đồ kết hợp ảnh nhìn thấy (RGB) và ảnh nhiệt (TIR) trong thuật toán bắt bám đối tượng.*

Như đã phân tích tại mục 2, phương pháp kết hợp cho hiệu năng cao nhất là kết hợp ở mức đặc trưng sau ước lượng IoU (Sơ đồ được mô tả tại hình 5). Sau lớp ước lượng IoU của ảnh nhìn thấy và ảnh nhiệt sẽ bổ sung một lớp kết hợp IoU để thực hiện kết hợp hai đặc trưng này. Lớp kết hợp IoU lấy đặc trưng ở lớp cuối của lớp Ước lượng IoU của RGB và Ước lượng IoU của TIR kết hợp với nhau. Mặt khác, để đưa trọng số môi trường vào việc kết hợp, từ công thức 3 ta có công thức 4, với $\boldsymbol{\omega}$ là trọng số đặc trưng ảnh nhìn thấy và $\mathbf{1} - \boldsymbol{\omega}$ là trọng số đặc trưng ảnh nhiệt. Sơ đồ của lớp kết hợp IoU có trọng số tại hình 7.

$$\boldsymbol{IoU(B)} = \boldsymbol{FC_2}\left(\boldsymbol{FC_1}\left(\boldsymbol{g_{1,2}}\left(\boldsymbol{c^{RGB}} \cdot \boldsymbol{z^{RGB}}\right)\boldsymbol{\omega}\right) \odot \boldsymbol{FC_1}\left(\boldsymbol{g_{1,2}}\left(\boldsymbol{c^{TIR}} \cdot \boldsymbol{z^{TIR}}\right)(\mathbf{1}-\boldsymbol{\omega})\right)\right) \quad (4)$$



Hình 6. *Sơ đồ lớp kết hợp IoU có trọng số.*

3.2. Xây dựng trọng số

Trong công bố [13], chúng tôi đã chứng minh được các điều kiện môi trường như độ chiếu sáng và nền nhiệt môi trường có ảnh hưởng để kết quả phát hiện đối tượng. Khi độ chiếu sáng yếu sử dụng ảnh nhiệt sẽ tốt hơn ảnh nhìn thấy và khi nền nhiệt cao độ tương phản của ảnh nhiệt sẽ thấp hơn ảnh nhìn thấy. Từ ý tưởng đó ta xây dựng trọng số của thuật toán bắt bám đối tượng. Tương tự như trong [13], có thể sử dụng độ nhiễu của ảnh nhìn thấy để xác định điều kiện ánh sáng. Tuy nhiên, trong sơ đồ kết hợp IoU có trọng số, ta đã có các vị trí có khả năng là đối tượng, do đó, có thể sử dụng trực tiếp độ tương phản để xác định chất lượng xác định đối tượng của ảnh nhiệt.

Như vậy, có thể tính trọng số $\boldsymbol{\omega}$ thông qua nhiễu độ nhiễu của ảnh nhìn thấy $\overline{\boldsymbol{noise}} \in [\mathbf{0}, \mathbf{1}]$ và chất lượng của mô tả đối tượng trong ảnh nhiệt thể hiện bằng độ tương phản Weber của đối tượng với nền gọi là $\overline{\boldsymbol{web}} = \frac{\boldsymbol{I}-\boldsymbol{I_b}}{\boldsymbol{I_b}}$, với I là cường độ sáng của đối tượng và $\boldsymbol{I_b}$ là cường độ sáng của nền, và được chuẩn hóa nằm trong dải $[\mathbf{0}, \mathbf{1}]$. Như vậy, công thức để tính trọng số được mô tả như sau:

$$\boldsymbol{\omega} = \boldsymbol{\alpha_1}\, \boldsymbol{e}^{\boldsymbol{\beta_1}\overline{\boldsymbol{web}}} + \boldsymbol{\alpha_2} \boldsymbol{e}^{-\boldsymbol{\beta_2}\overline{\boldsymbol{noise}}} + \mathbf{1} \tag{5}$$

Để tìm được các tham số $\boldsymbol{\alpha_1}, \boldsymbol{\beta_1}, \boldsymbol{\alpha_2}, \boldsymbol{\beta_2}$ cần sử dụng phương pháp để khớp đường cong phi tuyến (curve-fitting) Levenberg-Marquardt với bộ dữ liệu huấn luyện (tương tự như trong [13]).

Trong một số bộ dữ liệu, khi trời tối, camera tự động chuyển sang sử dụng chiếu sáng bằng tín hiệu hồng ngoại gần, ảnh nhận được là ảnh đa mức xám, do đó không áp dụng công thức 5 được. Vì vậy, trong thuật toán ta sẽ xác định nếu ảnh nhìn thấy ở dạng đa mức xám (khi camera tự động chuyển sang sử dụng tín hiệu hồng ngoại gần) bằng chỉ số MSE (Mean Squared Error). Nếu là ảnh đa mức xám, trọng số sẽ được xác định bằng độ tương phản giữa ảnh nhìn thấy (sau khi chiếu hồng ngoại) và ảnh nhiệt bằng công thức 6.

$$\boldsymbol{\omega} = \boldsymbol{\alpha_1}\, \boldsymbol{e}^{\boldsymbol{\beta_1}\overline{\boldsymbol{web}}_{\boldsymbol{RGB}}} + \boldsymbol{\alpha_2} \boldsymbol{e}^{-\boldsymbol{\beta_2}\overline{\boldsymbol{web}}_{\boldsymbol{T}}} + \mathbf{1} \tag{6}$$

Trong đó, $\overline{\boldsymbol{web}}_{\boldsymbol{RGB}}$, $\overline{\boldsymbol{web}}_{\boldsymbol{T}}$ là độ tương phản Weber của đối tượng dự đoán trên ảnh nhìn thấy và ảnh nhiệt.

4. THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

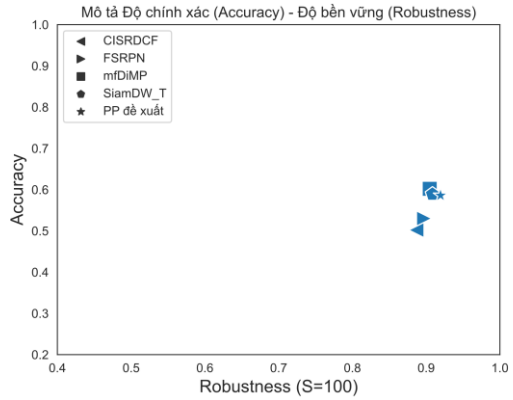
4.1. Điều kiện thử nghiệm

Thuật toán bắt bám đối tượng kết hợp ảnh nhìn thấy và ảnh nhiệt được thử nghiệm trên bộ dữ

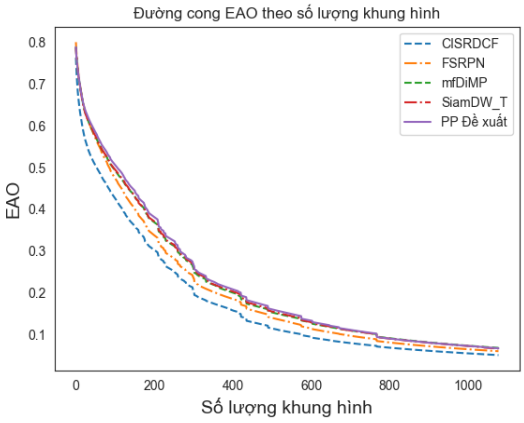
liệu VOT – RGBT 2019. Kèm bộ dữ liệu này có các công cụ để đo lường hiệu năng và tốc độ của thuật toán, được gọi là VOT Tookit. Bộ dữ liệu VOT-RGBT 2019 bao gồm 60 chuỗi và 20083 khung hình mô tả các hoàn cảnh bắt bám đối tượng [1]. Trong đó, có 43 chuỗi vào ban ngày, 17 chuỗi vào ban đêm; 2798 số khung hình có đối tượng bị che khuất; 0 số khung hình có độ chiếu sáng thay đổi; 17751 số khung hình có chuyển động thay đổi; 10927 số khung hình có đối tượng kích thước thay đổi; 2019 khung hình có camera chuyển động. Chiều dài chuỗi trung bình là 224; khoảng lấy mẫu chiều dài chuỗi để tính EAO (Expected Average Overlap) có phân bố xác xuất chiếm 50% là từ 46 đến 291. Giống như trong [1], các chỉ số được thử nghiệm là các các chỉ số chính trong VOT như: chỉ số độ chính xác (Accuracy, viết tắt là A), chỉ số độ bền vững (Robustness, viết tắt là R) và chỉ số độ chồng lấn trung bình kỳ vọng (EAO - Expected Average Overlap) là chỉ số tổng hợp của hai chỉ số trên.

4.2. Kết quả thử nghiệm

Thuật toán được mô tả ở mục 3 được so sánh với các thuật toán bắt bám đối tượng sử dụng đa nguồn ảnh điển hình như CISRDCF (phát triển từ CSR-DCF), FSRPN, mfDiMP, SiamDW_T. Kết quả phương án đề xuất cho hiệu năng cao nhất, EAO là 0,423 so với 0,413; 0,411; 0,387; 0,346 của các thuật toán SiamDW_T; mfDiMP; FSRPN và CISRDCF (hình 9) với tốc độ là khoảng 13,1 khung hình/giây đối với máy tính thử nghiệm có CPU: AMD Ryzen 5 3600 và GPU: GXT 1080 Ti. So với SiamDW_T có tốc độ 19,2 khung hình/giây; mfDiMP có tốc độ 8,61 khung hình/giây (bảng 2).



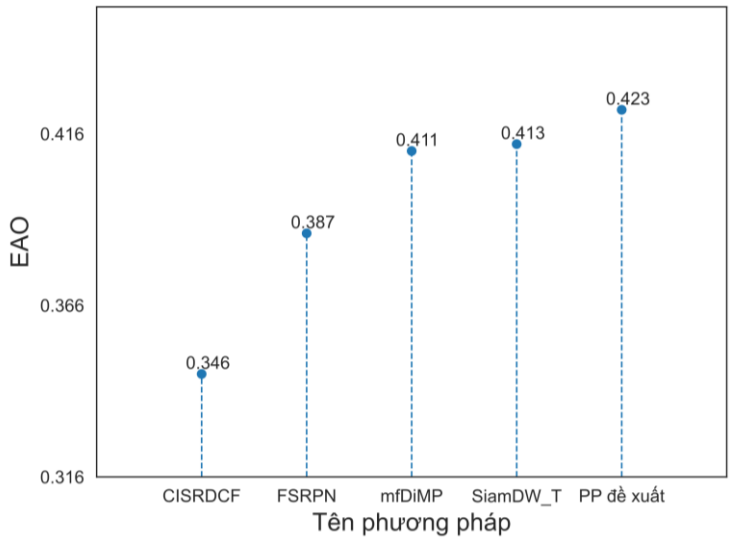
Hình 7. *Biểu đồ độ chính xác - độ bền vững.*



Hình 8. *Đường cong EAO theo số lượng khung hình.*

5. KẾT LUẬN

Trong bài báo, các tác giả đã trình bày các nghiên cứu, phân tích về các phương pháp kết hợp ảnh nhìn thấy và ảnh nhiệt trong bài toán bắt bám đối tượng và đề xuất thuật toán kết hợp sử dụng trọng số dựa trên độ nhiễu của ảnh nhìn thấy và độ tương phản của đối tượng trong ảnh nhiệt. Kết quả thử nghiệm trên bộ dữ liệu VOT-RGBT 2019 cho thấy, thuật toán có chỉ số EAO đạt 0,423 ở tốc độ 13,1 khung hình/giây đối với phần cứng thông dụng. Kết quả này cho hiệu năng cao hơn so với một số thuật toán bắt bám đối tượng sử dụng đa nguồn ảnh phổ biến như CISRDCF, FSRPN, mfDiMP, SiamDW_T. Trong tương lai, có thể bổ sung các phương pháp xác định điều kiện môi trường chi tiết hơn nữa để ước lượng trọng số kết hợp để nâng cao hiệu năng của bài toán bắt bám đối tượng đồng thời có thể tối ưu các phương pháp tính toán để tăng tốc độ xử lý của thuật toán.



Hình 9. *Thứ tự EAO trung bình của từng thuật toán.*

TÀI LIỆU THAM KHẢO

[1]. M. Kristan et al., *"The Seventh Visual Object Tracking VOT2019 Challenge Results,"* 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 2206-2241, (2019), doi: 10.1109/ICCVW.2019.00276.

[2]. D. S. Bolme, J. R. Beveridge, B. A. Draper and Y. M. Lui, *"Visual object tracking using adaptive correlation filters,"* 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2544-2550, (2010), doi: 10.1109/CVPR.2010.5539960.

[3]. Henriques, Joao & Caseiro, Rui & Martins, Pedro & Batista, Jorge. *"High-Speed Tracking with Kernelized Correlation Filters"*. IEEE Transactions on Pattern Analysis and Machine Intelligence. 37. 10.1109/TPAMI.2014.2345390, (2014).

[4]. Lukežič, A., Vojíř, T., Čehovin Zajc, L. et al. *"Discriminative Correlation Filter Tracker with Channel and Spatial Reliability"*. Int J Comput Vis 126, 671–688 (2018). https://doi.org/10.1007/s11263-017-1061-3

[5]. M. Danelljan, G. Häger, F. S. Khan and M. Felsberg, *"Convolutional Features for Correlation Filter Based Visual Tracking,"* 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), pp. 621-629, (2015), doi: 10.1109/ICCVW.2015.84.

[6]. Bertinetto, Luca & Valmadre, Jack & Henriques, Joao & Vedaldi, Andrea & Torr, Philip. *"Fully-Convolutional Siamese Networks for Object Tracking"*. 9914. 850-865. 10.1007/978-3-319-48881-3_56, (2016).

[7]. B. Li, J. Yan, W. Wu, Z. Zhu and X. Hu, *"High Performance Visual Tracking with Siamese Region Proposal Network,"* 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8971-8980, (2018), doi: 10.1109/CVPR.2018.00935.

[8]. Danelljan, Martin & Bhat, Goutam & Khan, Fahad & Felsberg, Michael. *"ATOM: Accurate Tracking by Overlap Maximization"*. 4655-4664. 10.1109/CVPR.2019.00479, (2019).

[9]. Zhang, Zhipeng & Peng, Houwen. *"Deeper and Wider Siamese Networks for Real-Time Visual Tracking"*. 4586-4595. 10.1109/CVPR.2019.00472, (2019).

[10]. Bhat, Goutam & Danelljan, Martin & Van Gool, Luc & Timofte, Radu. *"Learning Discriminative Model Prediction for Tracking"*. 6181-6190. 10.1109/ICCV.2019.00628, (2019).

[11]. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, *"ImageNet: A Large-Scale Hierarchical Image Database"*. IEEE Computer Vision and Pattern Recognition (CVPR), (2009).

[12]. Zhang, Lichao & Danelljan, Martin & Gonzalez-Garcia, Abel & Weijer, Joost & Khan, Fahad. *"Multi-Modal Fusion for End-to-End RGB-T Tracking"*. 2252-2261. 10.1109/ICCVW.2019.00278, (2019).

[13]. Hiep Dao, Hieu Dinh Mac, and Duc Quang Tran *"Noise-aware deep learning algorithm for one-stage multispectral pedestrian detection,"* Journal of Electronic Imaging 31(3), 033035, 16 June (2022). https://doi.org/10.1117/1.JEI.31.3.033035

[14]. S. Hwang, J. Park, N. Kim, Y. Choi and I. S. Kweon, *"Multispectral pedestrian detection: Benchmark dataset and baseline,"* 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1037-1045, (2015), doi: 10.1109/CVPR.2015.7298706.

ABSTRACT

Weighted Multi-Modal Fusion for RGB-T Tracking

As an important task in computer vision, visual object tracking, especially RGB tracking like KCF, CSRDCF, SiamFC, SiamRPN, ATOM, SiamDW, DiMP are commonly believed to be fast and reliable enough be deployed. However, RGB tracking obtains unsatisfactory performance in bad environmental conditions, e.g. low illumination, rain, and smog. It was found that thermal infrared sensors (8÷14 μm) provide a more stable signal for these scenarios. Some same level fusion modal algorithms such as FSRPN, SiamDW_T, mfDiMP obtain higher results while the environmental conditions are not considered. The paper describes a weighted multi-modal fusion for RGB-T tracking. Experiments are carried on VOT-RGBT dataset that demonstrate our algorithm achieve EAO of 0.423, higher than some popular tracking algorithms and can operate at speed of 13 fps on casual hardware.

Keywords: Visual Object Tracking; Multi-modal fusion; Convulutional Neural Network; Discriminative Correlation Filtes.