Journal of KIIT. Vol. 21, No. 11, pp. 115-124, Nov. 30, 2023. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2023.21.11.115

교차 어텐션 퓨전 기반의 멀티스펙트럴 객체 검출

양민석*, 손창환**

Multispectral Object Detection based on Cross-Attention Fusion

Min-Seok Yang*, Chang-Hwan Son**

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2020R1A2C1010405)

요 약

본 논문에서는 나이트 비전을 위한 RGB/IR 영상 기반의 멀티스펙트럴 객체 검출 모델을 소개하고자 한다. 기존의 멀티스펙트럴 객체 검출 모델의 퓨전 방식은 연결 계층을 통해서 단순히 RGB와 IR 특징을 스택으로 쌓는 방식을 취하고 있다. 그러나 이 퓨전 방식은 RGB와 IR 영상의 특징을 상호보완하도록 설계되어 있지 않다. 따라서 본 연구에서는 RGB/IR 특징 정보를 상호교환하여 특징의 구별력을 강화할 수 있는 교차 어텐션 퓨전 모듈을 새롭게 제안하고자 한다. 실험 결과를 통해, 제안한 멀티스펙트럴 객체 검출 모델이 기존의 단일 객체 검출 모델과 멀티스펙트럴 객체 검출 모델보다 정량적 평가인 AP 수치에서 더 우수한 성능을 달성할 수 있었다. 또한 애블레이션 실험을 통해, 제안한 교차 어텐션 퓨전 모듈이 AP 성능을 0.02로 향상할 수 있음을 확인하였다.

Abstract

In this paper, we would like to introduce a multispectral object detection model based on RGB/IR images for night vision. The fusion method of the existing multispectral object detection model simply stacks RGB and IR features through a connection layer. However, this fusion method is not designed to complement the characteristics of RGB and IR images. Therefore, in this study, we would like to propose a new cross-attention fusion module that can strengthen the distinguishing power of features by mutually exchanging RGB/IR feature information. Through experimental results, the proposed multispectral object detection model is able to achieve better performance in AP quantitative evaluation than the existing single object detection model and multispectral object detection model. Additionally, through ablation experiments, it is confirmed that the proposed cross-attention fusion module can improve AP performance to 0.02.

Keywords
multispectral imaging, object detection, attention model, image fusion

* 군산대학교 소프트웨어학부 소프트웨어학전공 학사과정
- ORCID: http://orcid.org/0009-0003-3353-5706
** 군산대학교 소프트웨어학부 소프트웨어학전공 부교수(교신저자)
- ORCID: http://orcid.org/0000-0001-7077-3074

· Received: Aug. 02, 2023, Revised: Sep. 07, 2023, Accepted: Sep. 10, 2023
· Corresponding Author: Chang-Hwan Son
Software Science and Engineering Major, School of Software, Kunsan National University, Republic of Korea
Tel.: +82-63-469-8915, Email: cson@kunsan.ac.kr

Ⅰ. 서 론

객체 검출은 컴퓨터 비전 분야에서 활발히 연구되는 분야 중의 하나로써, 이미지나 비디오에서 객체가 존재하는 위치를 바운딩 박스로 검출하고 객체의 부류를 예측하는 기술을 말한다. 이러한 객체 검출 기술은 무인자동차, 지능형 CCTV, 지능형 로봇에서 보행자 검출 및 안면 인식을 위해 널리 도입되고 있다. 하지만 이러한 객체 검출 기술은 객체의 스케일과 자세 그리고 조도 상태나 폐색 정도에 상당한 영향을 받는다. 특히, 저조도 환경에서, 예를 들면 야간이나 폭우[1]가 내리는 저조도 환경에서 객체 검출 기술은 색상 식별의 어려움으로 성능 저하가 수반된다. 따라서 저조도 환경에서의 객체 검출 성능 향상을 위한 기술 고도화 방안이 요구된다.

야간 환경에서 객체 검출 성능 향상을 위해, 저조도 영상의 대비 개선[2]과 같은 전처리 과정을 적용하거나 멀티스펙트럴 영상을 촬영하여 영상 퓨전 기법[3,4]을 적용할 수 있다. 본 연구에서는 멀티스펙트럴 영상 기반의 객체 검출 기법에 대해 자세히 소개하고자 한다. 나이트 비전(Night vision)을 위해, 주로 사용되는 멀티스펙트럴 영상은 RGB 영상과 IR(Infrared) 영상이다. 그림 1은 야간과 주간에서 촬영된 RGB 영상과 IR 영상의 샘플을 보여주고 있다. 그림 1(b)에 보듯이, IR 영상은 그림 1(a)의 RGB 영상에 비해, 야간 시 보행자 식별이 더 쉬운 것을 볼 수 있다. 따라서 IR 영상의 사용은 야간 객체 검출 성능 향상을 유도할 것으로 보인다. 반면 주간에는 그림 1(c)와 1(d)에 보듯이, RGB 영상이 IR 영상보다 보행자, 자동차, 보도와 같은 객체 분류 작업에 더 효과적임을 알 수 있다. 즉, RGB 영상과 IR 영상은 정보량 측면에서 상호 보완적인 특성을 지니고 있다. 이런 멀티스펙트럴 영상은 자동차나 보안 카메라에서 보행자 충돌 방지나 객체 검출 목적으로 활용되고 있다.

본 연구에서 멀티스펙트럴 객체 검출이란 RGB 영상과 IR 영상을 퓨전하여 보행자가 위치한 바운딩 박스를 찾는 기술을 말한다. 최근 멀티스펙트럴 객체 검출 기법은 단일 객체 검출에 사용되던 딥러닝 모델을 각각의 입력 영상에 적용하여 특징을 추출하고 퓨전하는 방식이 대세이다[5]. 멀티스펙트럴 객체 검출에 사용되는 단일 객체 검출 모델은 YOLO[6] 및 Faster RCNN(Regions with Convolutional Neural Networks)[7]이 있다. 하지만 최근 속도와 정확도 측면에서 YOLO나 Faster RCNN을 능가하는 단일 객체 모델이 소개되고 있다. 예를 들면, CenterNet[8], RetinaNet[9], YOLOX[10] 등이 있다. 이는 멀티스펙트럴 객체 검출을 위해 최신 단일 객체 검출 모델에 적합한 퓨전 모델을 개발할 필요가 있음을 의미한다.

그림 1. 멀티스펙트럴 영상 예시 (a) RGB 영상(야간), (b) IR 영상(야간), (c) RGB 영상(주간), (d) IR 영상(주간)
Fig. 1. Examples of multispectral images
(a) RGB image(night), (b) IR image(night),
(c) RGB image(day), (d) IR image(day)

따라서 본 연구에서는 최신 단일 객체 검출 모델인 CenterNet을 활용하여 멀티스펙트럴 객체 검출에 확장할 수 있는 새로운 퓨전 방식을 제시하고자 한다. CenterNet은 앵커 박스 후보를 생성하는 기존의 객체 검출 모델들과는 달리, 객체의 중심점을 추정하는 객체 검출 모델이다. 제안한 모델에서는 RGB Hourglass 백본과 IR Hourglass 백본을 공유(Sharing)하면서 각 백본에서 추출된 RGB 영상의 특징 정보와 IR 영상의 특징 정보를 서로 교환함으로써, 특징 추출 성능을 강화할 수 있는 교차 어텐션 퓨전 모듈(Cross-attention fusion module)을 새롭게 제시하고자 한다.

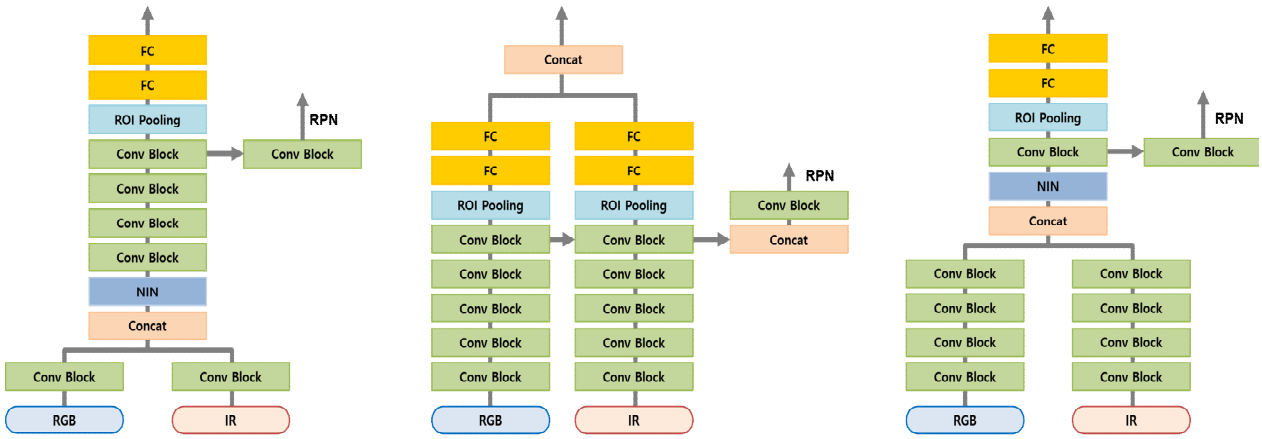
또한 Stacked Hourglass에서 앞단의 Hourglass 백본에서 추출된 히트맵, 즉 객체가 존재하는 위치에 화이트 색상을 지닌 맵을 활용하여 뒷단의 Hourglass 성능을 개선할 수 있는 공간 어텐션 모듈(Spatial-attention module)도 새롭게 제안하고자 한다. 그리고 실험 결과를 통해, 제안한 교차 어텐션 퓨전 모듈과 공간 어텐션 퓨전 모듈이 멀티스펙트럴 객체 검출 성능을 각각 제고할 수 있음을 보이고자 한다.

II. 기존 멀티스펙트럴 특징 퓨전 방식

그림 2는 기존의 Faster R-CNN 모델을 사용한 멀티스펙트럴 특징 퓨전 방식을 보여주고 있다. 그림에서 보듯이, 퓨전 방식은 크게 3가지로 초기 퓨전(Early fusion), 후기 퓨전(Late fusion), 중기 퓨전(Halfway Fusion)로 나눌 수 있다. 백본은 3가지 모델 모두 VGG-16을 사용했다고 가정했다. VGG(Visual Geometry Group) 모델은 5개의 합성곱 블록(Convolution block)과 2개의 완전연결계층(FC, Fully Connected layer)로 이루어져 있다.

멀티스펙트럴 객체 검출은 RGB 영상과 IR 영상, 즉 두 종류의 영상을 입력받기 때문에 퓨전 과정이 반드시 동반된다. 멀티스펙트럴 특징 퓨전이란 바로 이 두 종류의 영상에서 추출된 특징을 하나로 결합하는 과정을 말한다. 그림 2(a)와 같이, RGB 영상과 IR 영상을 하나의 합성곱 블록을 통과한 후, 바로 퓨전 하거나 그림 2(b)와 같이 백본의 말단 부분인 완전연결계층에서 퓨전 할 수 있다. 여기서 퓨전이란 일반적으로 연결계층(Concatenation layer)을 통해 두 종류의 특징맵을 스택처럼 쌓는 방식을 말한다. 또한 초기 및 후기 퓨전처럼, 백본의 중간 단계에서도 두 종류의 특징맵을 퓨전 할 수 있다. 그러나 후기 퓨전과는 달리, 중기 퓨전은 공간 정보를 유지하고 있는 점이 다르다. 후기 퓨전은 완전연결계층을 통과하기 때문에 특징의 공간 정보가 사라진다. 그리고 중기와 후기 퓨전의 또 다른 차이점은 관심영역 풀링(ROI pooling)의 적용 방식이다. 중기 퓨전은 관심영역 풀링이 하나인 반면 후기 퓨전은 RGB와 IR 백본에서 두 개의 관심영역 풀링이 각각 적용된다.

그림 2에서 RPN(Region Proposal Network)은 영역 제안 네트워크로써 바운딩 박스 후보를 검출하는 역할을 담당하고, NIN(Network-in-Network)[11]은 1×1 합성곱 필터를 사용해서 채널 차원을 변경함으로써, 기존 VGG 모델의 다음 계층에 입력으로 사용될 수 있도록 조정하는 역할을 한다. 지금까지 연구된 결과에 따르면[12], 초·중·후기 퓨전 중에서 가장 좋은 성능은 중기 퓨전을 적용했을 때라고 보고되고 있다.



(a) (b) (c)

그림 2. 기존 멀티스펙트럴 객체 검출을 위한 퓨전 모델; (a) 초기 퓨전, (b) 후기 퓨전, (c) 중기 퓨전

Fig. 2. Conventional fusion models for multispectral object detection fusion; (a) early fusion, (b) late fusion, (c) halfway fusion

III. 배경 이론

3.1 CenterNet 모델

최근 단일 객체 검출 모델 중에서 추론 속도와 정확도 측면에서 우수한 성능은 갖춘 모델로 CenterNet을 들 수 있다. 본 연구에서는 이 단일 객체 검출 모델을 멀티스펙트럴 객체 검출 모델에 확장하는 방안을 제시하고자 한다. 먼저 CenterNet 모델의 기본적인 아키텍처와 특징을 제안하고자 한다.

그림 3에서 보듯이, CenterNet 모델은 Hourglass[13]를 스택으로 쌓은 형태를 취하고 있다. 그리고 입력단의 특징 정보를 뒷단의 Hourglass에 전파하기 위해 긴 스킵 연결(Long skip connection) 구조를 포함하고 있다. 무엇보다 Hourglass 백본을 통과한 예측 결과물이 3종류의 맵을 출력하는 것을 볼 수 있다. 히트맵(Heat map)은 객체의 중심 위치를 화이트 색상으로 표현한 키포인트에 대한 정보를 가지고 있다. 즉, 배경색은 검정색이고 객체가 존재하는 중심 위치에는 화이트 색상을 가지는 맵이다. 그리고 이 키포인트를 중심으로 바운딩 박스의 가로 및 세로 위치에 대한 정보를 예측하기 위해 바운딩 박스 맵(Bounding box map)이 생성된다. 옵셉 맵(Offset map)은 원본 히트맵을 다운샘플링 할 경우, 키포인트 위치 변경에 따른 양자화 오차를 보정하기 위해 필요하다.

Hourglass는 특징 추출을 위한 백본의 역할을 하고 인코더-디코더 타입의 구조를 가진다. 그리고 인코더와 디코더 간에는 특징 정보를 전달하기 위해 스킵 연결로 이어져 있다. Hourglass를 통과한 특징 맵은 합성곱 블록을 통과한 후, 최종 추정치인 히트맵, 바운딩 박스 맵, 옵셋 맵으로 변환된다.

3.2 손실함수

CenterNet 모델은 3종류의 맵을 예측하기 때문에 손실함수도 아래와 같이 3개의 항목으로 구성된다.

$$L = L_k + \lambda_s L_s + \lambda_{off} L_{off} \tag{1}$$

여기서 L_k, L_s, L_{off}은 각각 히트맵, 바운딩 박스 맵, 옵셉 맵에 대한 손실을 의미한다. 그리고 λ_s와 λ_{off}는 손실에 대한 가중치를 의미하며 각각 1과 0.1로 세팅된다.

$$L_k = \frac{-1}{N} \sum_{xyc} \begin{cases} (1-\hat{Y}_{xyc})^{\alpha} \log(\hat{Y}_{xyc}) & \text{if } Y_{xyc} = 1 \\ (1-\hat{Y}_{xyc})^{\beta} (\hat{Y}_{xyc})^{\alpha} & otherwise \\ \log(\hat{Y}_{xyc}) \end{cases} \tag{2}$$

식 (2)는 교차 엔트로피를 수정한 초점 손실(Focal loss)에 해당한다. 초점 손실은 긍정 샘플과 거짓 샘플의 클래스 불균형 문제를 해결하기 위해 고안된 손실이다. 즉, 쉽게 분류될 수 있는 샘플에 대한 패널티를 낮게 부여하여 특정 클래스로 바이어스되는 학습을 막을 수 있다.

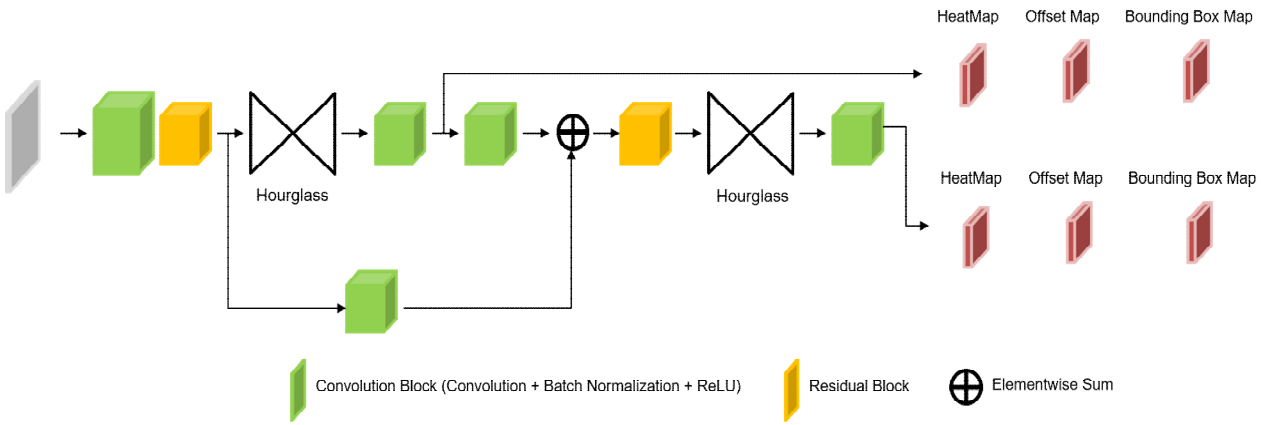


그림 3. 객체 검출을 위한 CenterNet 아키텍처
Fig. 3. Architecture of the CenterNet for object detection

식 (2)에서 $\hat{Y}$와 Y는 각각 예측한 히트맵과 정답 히트맵을 의미한다. 정답 히트맵은 라벨링된 바운딩 박스의 중심점에 화이트 색상을 할당해서 생성할 수 있다. 식 (2)에서 $Y_{xyc}=1$인 경우, 즉 키포인트 위치에서는 예측된 히트맵 $\hat{Y}_{xyc}$가 1에 가까워야 손실이 작아진다. 반면에 $Y_{xyc}\neq 1$인 경우, 예측된 히트맵 $\hat{Y}_{xyc}$이 0에 가까워야 손실이 최소화될 수 있다. 수식에서 α와 β는 초점함수의 곡률을 조정하기 위한 파라미터이고 N은 키포인트 개수를 나타낸다.

$$L_{off}=\frac{1}{N}\sum_{p}=\left|\hat{O}_{\tilde{p}}-\left(\frac{p}{R}-\tilde{p}\right)\right| \tag{3}$$

p는 키포인트이고 R은 다운샘플링을 위한 스트라이드(Stride)를 의미한다. $\tilde{p}$는 키포인트를 스트라이드로 나눈 후에 소수점 이하를 버림한 값이다. 영상 좌표는 정수로 표현되므로 양자화 오차가 발생한다. 따라서 식 (3)의 옵셋 손실 L_{off}는 양자화 오차를 보정하기 위한 용도로 사용된다.

마지막으로 키포인트 위치에서의 가로 및 세로 길이의 추정치 오차를 측정하기 위한 바운딩 박스 맵의 손실 L_s는 다음과 같이 정의된다.

$$L_s=\frac{1}{N}\sum_{k=1}^{N}\left|\hat{S}_{pk}-s_k\right| \tag{4}$$

여기서 $\hat{S}$는 바운딩 박스의 가로와 세로 길이를 저장하고 있는 추정된 바운딩 박스 맵이다. $\hat{S}$의 크기는 히트맵의 크기와 동일하고 채널의 개수는 2개이다. 식 (4)은 모든 키포인트에 대해 추정된 가로/세로 길이와 정답 세로/가로 길이 간의 오차의 합을 정량적으로 측정한 것을 의미한다.

Ⅳ. 제안한 교차 어텐션 멀티스펙트럴 CenterNet

4.1 제안한 접근 방법

본 논문에서는 기존 CenterNet을 RGB/IR 멀티스펙트럴 객체 검출에 확장하기 위해, 교차 어텐션 퓨전과 공간 어텐션 모듈을 고안하고자 한다. 기존의 멀티스펙트럴 객체 검출을 위한 퓨전 모듈은 연결 계층을 많이 활용하였다. 하지만, RGB/IR 영상의 경우, 그림 1에서 확인했듯이 정보량 측면에서 서로 상호보완해야 할 필요가 있다. 이를 위해, 본 연구에서는 RGB/IR Hourglass 백본에서 각각 추출된 특징 정보를 서로 교환하기 위한 새로운 CenterNet 기반의 퓨전 아키텍처를 설계하였다. 즉, RGB와 IR 백본에서 추출된 특징 정보의 차를 어텐션으로 모델링함으로써 교차 어텐션 퓨전 모듈을 완성하였다. 그리고 Stacked Hourglass에서 앞단의 Hourglass의 예측 값인 히트맵 결과를 공간 어텐션 모듈에 활용할 수 있는 방안도 마련하였다.

4.2 제안한 멀티스펙트럴 객체 검출 모델

그림 4는 제안한 RGB/IR 멀티스펙트럴 객체 검출을 위한 교차 어텐션 퓨전을 탑재한 CenterNet 모델의 확장판이다. 그림에서 보듯이, 제안한 모델은 이중 분기(Dual branch)로 구성되어 있다. 즉, RGB 영상을 입력으로 받아 특징을 추출하는 RGB 분기(윗단)와 IR 영상을 입력으로 받아 특징을 추출하는 IR 분기(아랫단)로 구성된다. RGB/IR 분기 모두 Stacked Hourglass를 백본으로 사용한다. Stacked Hourglass란 그림 4에서 보듯이, Hourglass 백본을 연속적으로 나열한 구조를 말한다. 그리고 RGB 분기와 IR 분기의 Hourglass 백본 간에 교차 어텐션 퓨전(Cross-attention fusion) 모듈을 통해 특징 정보를 상호 교환한다. 이를 통해, RGB 특징 정보와 IR 특징 정보 간의 부족한 부분을 상호 보완할 수 있다. 그리고 첫 번째 Hourglass를 통과한 특징은 합성곱 블록을 거쳐 1차 히트맵 결과를 생성한다. 이 히트맵은 객체의 중심점에 대한 정보를 갖고 있다. 따라서 히트맵은 공간적으로 특징의 중요도를 결정할 수 있는 중요한 단서가 될 수 있다. 제안한 모델에서는 이 히트맵을 공간 어텐션 모듈로 활용하고자 한다. 즉, 그림 4에서와 같이, 히트맵을 공간 중요도를 나타내는 가중치 맵으로 간주하여 두 번째 Hourglass의 입력 특징 맵과 요소별 곱(Elementwise product)을 함으로써, 공간 어텐션 모듈을 구현하였다.

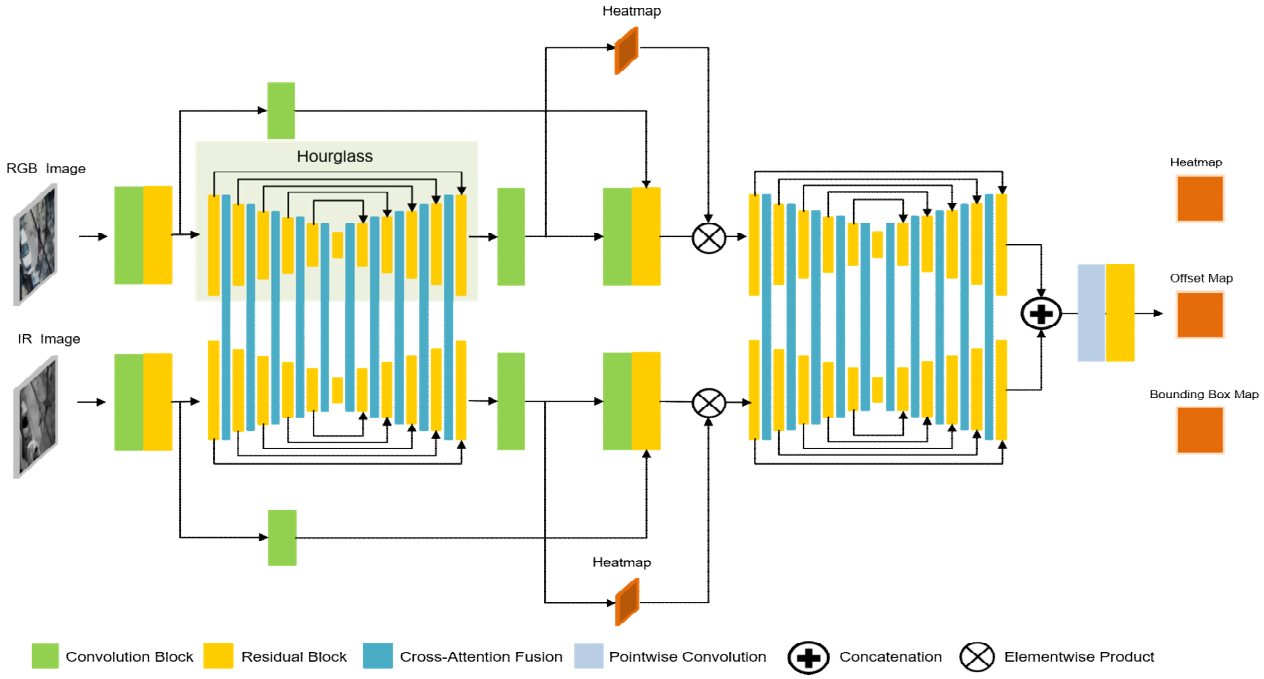


그림 4. 제안한 교차 어텐션 퓨전 기반의 멀티스펙트럴 객체 검출 모델
Fig. 4. Proposed multispectral object detection model based on cross-attention fusion

RGB와 IR 분기는 두 번째 Hourglass를 통과한 후, 연결계층을 통해 이중 분기에서 출력되는 결과를 결합한 후, 최종 히트맵, 옵셋 맵, 바운딩 박스 맵을 예측한다. 제안한 모델에서 사용된 손실함수는 3.2절에 소개된 CenterNet의 손실함수와 동일하다.

4.3 제안한 교차 어텐션 퓨전 모듈

그림 5는 그림 4에 소개된 교차 어텐션 퓨전 모듈의 상세한 구조를 보여준다.

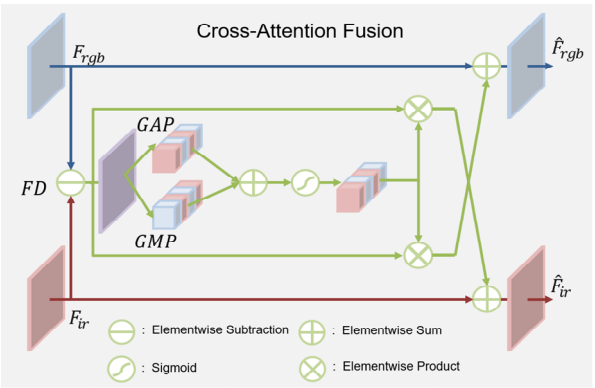


그림 5. 교차 어텐션 퓨전 모듈
Fig. 5. Cross-attention fusion module

교차 어텐션 퓨전 모듈에서는 먼저 RGB와 IR 분기의 Hourglass 백본에서 출력된 두 특징맵 간의 차이를 계산한다.

$$FD_{rgb} = F_{ir} - F_{rgb} \tag{5}$$

$$FD_{ir} = F_{rgb} - F_{ir} \tag{6}$$

여기서 F_{rgb}는 RGB 분기의 Hourglass에서 추출된 특징맵이고 F_{ir}은 IR 분기의 Hourglass에서 추출된 특징맵을 의미한다. 따라서 FD_{rgb}는 F_{ir}에서 F_{rgb}를 뺀 값이고, FD_{ir}은 F_{rgb}에서 F_{ir}를 뺀 값이다. 이러한 차연산은 두 특징맵 간의 차이점을 파악하는 데 용이하다.

$$\hat{F}_{rgb} = F_{rgb} \oplus \sigma(GAP(FD_{rgb}) \oplus GMP(FD_{rgb})) \odot FD_{rgb} \tag{7}$$

$$\hat{F}_{ir} = F_{ir} \oplus \sigma(GAP(FD_{ir}) \oplus GMP(FD_{ir})) \odot FD_{ir} \tag{8}$$

식 (7), (8)에서 GAP와 GMP는 각각 전역 평균 풀링(Global average pooling)과 전역 최대 풀링(Global max pooling)을 의미한다. σ는 시그모이드 함수이고 $\oplus$과 $\odot$는 각각 요소별 합(Elementwise sum)과 요소별 곱을 의미한다. 식 (7), (8)에서 GAP와 GMP를 적용하는 이유는 특징맵에서 채널의 중요도를 파악하기 위함이다. 채널 중요도를 모델링하기 위해, 각 채널에서 최대 값을 추출하거나 평균값을 취할 수 있다. 그리고 채널 중요도는 가중치로 활용되기 때문에 0~1사이의 값으로 변환되어야 한다. 이를 위해, 제안한 교차 어텐션 퓨전 모델에서는 시그모이드 함수를 적용하였다. 이렇게 생성된 가중치 맵은 각각 FD_{ir}와 F_{rgb}에 곱한 다음, 입력 F_{rgb}와 F_{ir}에 더함으로써 교차 어텐션 퓨전 을 수행할 수 있다. 즉, RGB/IR Hourglass 백본에서 각각 추출된 특징 정보를 서로 상호보완할 수 있다.

4.4 제안한 공간 어텐션 모듈

RGB와 IR 분기의 첫 번째 Hourglass로 추정된 히트맵은 객체의 중심점에 대한 정보를 포함하고 있다. 히트맵은 0~1사이의 값으로 표현되는데, 객체 영역에서는 1에 가까운 값, 즉 화이트 색상으로 표시되고 배경 영역에서는 0에 가까운 검정색으로 표현된다. 이는 히트맵이 공간적인 중요도로 활용될 수 있음을 의미한다. 따라서 본 연구에서는 공간 어텐션 모듈을 설계하기 위해 추정된 히트맵을 가중치로 바로 사용하고자 한다.

$$FI_{rgb} = FO_{rgb} \odot H_{rgb} \tag{9}$$

$$FI_{ir} = FO_{ir} \odot H_{ir} \tag{10}$$

여기서 FI_{rgb}와 FI_{ir}은 각각 두 번째 Hourglass에 입력될 RGB와 IR 특징맵을 의미하고, H_{rgb}와 H_{ir}은 첫 번째 Hourglass에서 추출된 히트맵 결과이다. 그리고, FO_{rgb}와 FO_{ir}는 각각 첫 번째 Hourglass에서 추출된 RGB와 IR 특징맵을 의미한다. H_{rgb}와 H_{ir}는 객체가 존재하는 위치 정보를 포함하기 때문에 공간적인 중요도로 활용될 수 있다. 따라서 본 연구에서는 식 (9), (10)과 같이 요소별 곱을 활용해서 공간 어텐션 모듈을 완성하였다.

Ⅴ. 실험 및 결과

5.1 실험 환경

LLVIP 학습 데이터셋은 저조도 환경에서의 보행자 검출을 위한 RGB/IR 멀티스펙트럴 영상 데이터로서, 15,488장의 RGB영상과 IR 영상 쌍으로 구성되어 있다. 학습을 위해, 7:3의 비율로 랜덤하게 훈련 집합과 테스트 집합으로 나누었고 가중치 갱신을 위해 아담 옵티마이저(Adam Optimizer)[14]를 사용했다. 배치 크기는 4, 에폭은 10, 학습률은 0.0001로 설정했다. 그리고 성능 평가를 위해, EfficientDet[15], CenterNet[8], RetinaNet[9]과 같은 SOTA 단일 객체 검출 모델과 후기 퓨전을 적용한 CenterNet 멀티스펙트럴 모델을 비교하였다. 그리고 제안한 모델의 교차 어텐션 퓨전 모듈과 공간 어텐션 모듈의 성능 효과를 검증하기 위해, 애블레이션 연구(Ablation study)를 수행하였다.

5.2 평가 척도

정량적 성능 평가를 위해, 객체 검출에서 널리 활용되는 AP(Average Precision) 척도를 도입했다. AP는 객체 검출 모델의 임계치 조정에 따른 정확도와 재현율(Precision and recall) 값을 선으로 연결했을 때, 생성된 곡선 아래의 면적에 해당한다. 여기서 재현율이란 모든 긍정 샘플 개수에서 모델이 정확하게 맞춘 긍정 샘플의 비율을 의미한다. 정확도란 검출된 긍정 샘플 중에서 실제 긍정 샘플의 개수를 의미한다. AP는 0~1사이의 값을 가지며 값이 클수록 모델의 검출 성능이 우수함을 나타낸다.

5.3 정량적 평가

표 1은 제안한 멀티스펙트럴 객체 검출 모델과 기존 모델과의 AP 성능 결과를 보여주고 있다. 표1에서 보듯이, 기존의 단일 객체 검출 모델은 IR 영상이 RGB 영상보다 더 우수한 객체 검출 성능을 가지는 것을 볼 수 있다. 이는 LLVIP 데이터셋이 대부분 야간 영상이므로 IR 영상이 나이트 비전에 더 효과적임을 말해준다.

표 1. 정량적 평가
Table 1. Quantitative evaluation

Model	Input image	AP
EfficientDet-D1[15]	RGB	0.862
EfficientDet-D1	IR	0.892
CenterNet[8]	RGB	0.868
CenterNet	IR	0.929
RetinaNet[9]	RGB	0.875
RetinaNet	IR	0.941
Conventional multispectral centerNet model based on later fusion	RGB + IR	0.940
Proposed multispectral object detection model	**RGB + IR**	**0.960**

그리고 기존 후기 퓨전 기반의 멀티스펙트럴 객체 검출 CenterNet 모델은 제안한 모델에서 교차 어텐션과 공간 어텐션 모듈을 제외한 모델을 의미한다. 즉, 그림 4에서 RGB 분기와 IR 분기를 말단에서 연결계층을 통해서 후기 퓨전만 적용한 아키텍처를 말한다. 표 1에서 보듯이, 제안한 멀티스펙트럴 객체 검출 모델이 기존 멀티스펙트럴 객체 검출 모델보다 성능이 더 우수함을 알 수 있다. 이는 제안한 모델이 RGB 분기와 IR 분기에서 추출된 특징 정보를 상호 교환하여 특징 구별력을 강화함으로써 객체 검출 성능을 향상했음을 말해준다. 표 2는 제안한 모델의 애블레이션 실험결과를 보여준다. 즉, 표 2는 교차 어텐션 퓨전 모듈과 공간 어텐션 퓨전 모듈의 적용 전후의 결과를 보여준다. 표 2에서 보듯이, 교차 어텐션 퓨전 모듈과 공간 어텐션 퓨전 모듈을 하나씩 추가할 때마다 객체 검출 성능이 향상되는 것을 확인할 수 있다. 특히 공간 어텐션 퓨전 모듈이 교차 어텐션 퓨전 모듈보다 성능 효과가 더 좋은 것을 볼 수 있다.

표 2. 애블레이션 연구
Table 2. Ablation study

Cross-attention fusion module	Spatial attention fusion module	AP
-	-	0.940
○	-	0.950
-	○	0.951
○	○	0.960

그림 6은 정밀도-재현율 곡선을 보여주고 있다. AP 결과에서 이미 예측했듯이, 제안한 멀티스펙트럴 객체 검출이 모델의 임계치 변화에도 객체 검출 성능이 가장 우수함을 알 수 있다. 즉, 그래프에서 면적이 가장 넓은 것을 의미하며 이는 재현율 변화에 따른 정확도 값이 가장 높게 유지되는 현상을 말해준다.

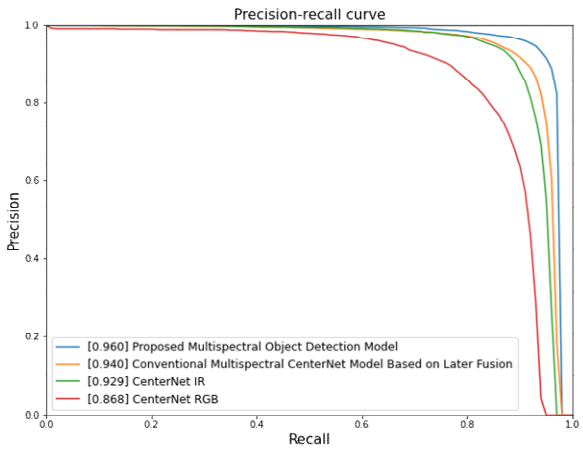


그림 6. 정밀도-재현율 곡선
Fig. 6. Precision-recall curve

5.4 바운딩 박스 검출 결과

그림 7은 단일 객체 검출 모델과 멀티스펙트럴 객체 검출 모델을 적용해서 검출된 바운딩 박스 결과이다. 그림에서 보듯이, 단일 객체 검출 모델인 CenterNet을 적용했을 때, RGB 영상에서는 오검출이 발견되고 IR 영상에서는 보행자가 미검출된 것을 알 수 있다. 반면, 제안한 멀티스펙트럴 객체 검출에서는 보행자를 모두 정확하게 검출한 것을 볼 수 있다. 이는 RGB/IR 멀티스펙트럴 객체 검출 접근 방식이 기존 단일 객체 검출 방식보다 나이트 비전에 더 효과적임을 뒷받침해 준다.

Ⅵ. 결 론

본 연구에서는 나이트 비전을 위한 RGB/IR 기반 멀티스펙트럴 객체 검출 모델을 새롭게 고안하였다. 기존의 RGB/IR 멀티스펙트럴 객체 검출은 이중 분기 구조로 연결계층을 활용해서 RGB와 IR 분기의 특징을 결합한다.

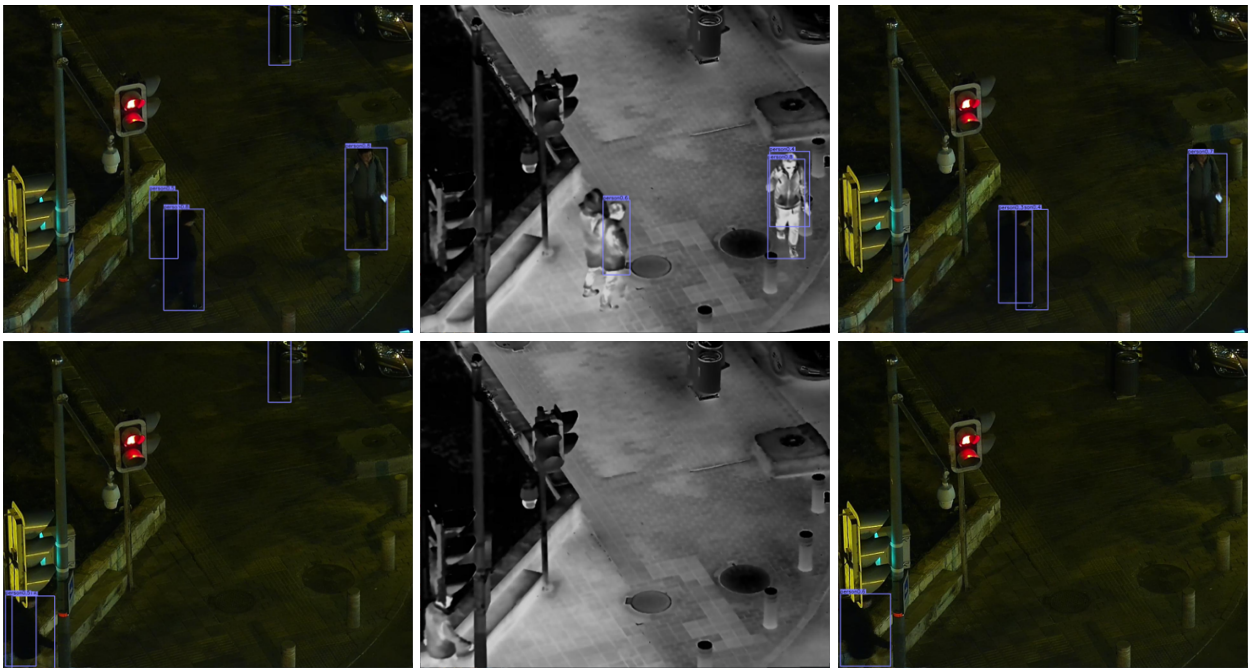


그림 7. 검출된 바운딩 박스 ; CenterNet (RGB영상), CenterNet (IR영상), 제안한 모델 (좌측에서 우측)
Fig. 7. Detected bounding boxes ; CenterNet (RGB image), CenterNet (IR image), proposed model (left to right)

그러나 이 연결계층 기반의 퓨전 모델은 RGB 정보와 IR 정보의 장단점을 상호보완하기에는 그 구조가 너무 단순하다. 따라서 본 연구에서는 RGB와 IR 특징 정보의 단점을 서로 보완할 수 있는 교차 어텐션 퓨전 모듈과 공간 어텐션 퓨전 모듈을 개발하였다. 실험 결과를 통해, 교차 어텐션과 공간 어텐션 퓨전 모델은 각각 AP 성능에서 0.01과 0.011의 성능 향상을 이끌 수 있었다. 또한 제안한 멀티스펙트럴 객체 검출 모델이 기존의 단일 객체 검출과 멀티스펙트럴 객체 검출보다 AP 수치에서 약 0.02의 성능 향상을 달성할 수 있었다.

References

[1] C.-H. Son and X.-P. Zhang, "Rain removal via shrinkage of sparse codes and learned rain dictionary", IEEE International Conference on Multimedia & Expo Workshop, Seattle, WA, USA, pp. 1-6, Oct. 2016. https://doi.org/10.48550/arXiv.1610.00386.

[2] A. Loza, D. Bull, and A. Achim, "Automatic contrast enhancement of low-light images based on local statistics of wavelet coefficients", IEEE International Conference on Image Processing, Hong Kong, China, pp. 3553-3556, Sep. 2010. https://doi.org/10.1109/ICIP.2010.5651173.

[3] C. H. Son and X. P. Zhang, "Near-infrared fusion via color regularization for haze and color distortion removals", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 28, No, 11, pp. 3111-3126, Nov. 2018. https://doi.org/10.1109/TCSVT.2017.2748150.

[4] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "Piafusion: A progressive infrared and visible image fusion network based on illumination aware", Information Fusion, Vol. 83-84, pp. 79-92, Jul. 2022. https://doi.org/10.1016/j.inffus.2022.03.007.

[5] J. Liu, S. Zhang, S. Wang, and D. Metaxas, "Multispectral deep neural networks for pedestrian detection", Proc. of the British Machine Vision Conference, York, UK, pp. 73.1-73.13, Sep. 2016. http://dx.doi.org/10.5244/C.30.73.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, realtime

object detection", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 779-788, Jun. 2016. https://doi.org/10.1109/CVPR.2016.91.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, No. 6, pp. 1137-1149, Jun. 2017. https://doi.org/10.1109/TPAMI.2016.2577031.

[8] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points", arXiv:1904.07850v2 [cs.CV], Apr. 2019. https://doi.org/10.48550/arXiv.1904.07850.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection", Proc. of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 2980-2988, Aug. 2017. https://doi.org/10.48550/arXiv.1708.02002.

[10] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding Yolo series in 2021", arXiv:2107.08430, Jul. 2021. https://doi.org/10.48550/arXiv.2107.08430.

[11] M. Lin, Q. Chen, and S. Yan, "Network in Network", arXiv preprint arXive:1312.4400, Dec. 2013. https://doi.org/10.48550/arXiv.1312.4400.

[12] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster RCNN for robust multispectral pedestrian detection", Pattern Recogniton, Vol. 85, pp. 161-171, Jan. 2019. https://doi.org/10.1016/j.patcog.2018.08.005.

[13] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation", in Proc. European Conference on Computer Vision, Amsterdam, Netherlands, pp. 483-499, Sep. 2016. https://doi.org/10.1007/978-3-319-46484-8_29.

[14] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization", in Proc. International Conference on Learning Representations, San Diego, USA, May 2015. https://doi.org/10.48550/arXiv.1412.6980.

[15] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection", in Proc. IEEE Conference on Computer Vision and Pattern Recognition(Virtual), pp. 10781-10790, Jun. 2020. https://doi.org/10.48550/arXiv.1911.09070.

저자소개

양 민 석 (Min-Seok Yang)

2018년 3월 ~ 현재 : 군산대학교 소프트웨어학부 소프트웨어학전공 학사과정

관심분야 : 컴퓨터 비전, 영상처리, 기계학습, 딥 러닝

손 창 환 (Chang-Hwan Son)

2017년 4월 ~ 현재 : 군산대학교 소프트웨어학부 소프트웨어학전공 부교수

관심분야 : 컴퓨터 비전, 영상처리, 기계학습, 딥 러닝