컴퓨터 비전

기말고사 동향분석 리포트

주제 : Multispectral Pedestrian Detection

학과 : 인공지능학과

학번 : 21110373

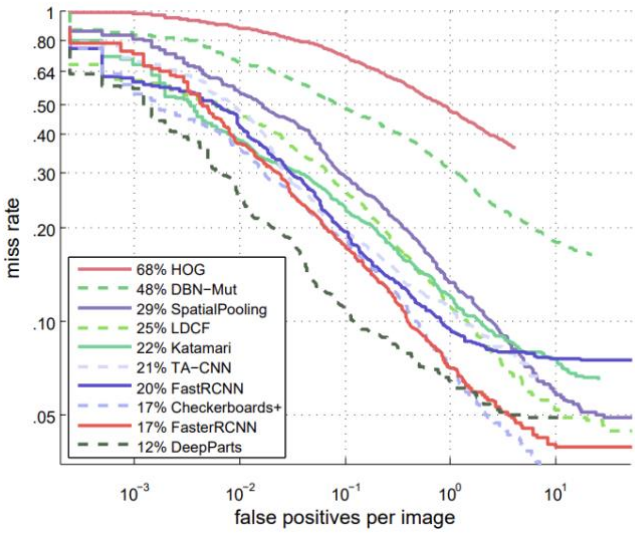
이름 : 차혜령

1. 서론

Multispectral Pedestrian Detection은 보행자 인식 기술에서 밤과 같이 어두운 환경이나, 구조물에 따라 작거나 가려진 부분에 대한 인식이 어려워지는 문제를 해결하기 위해 야간에서도 잘 찾을 수 있도록 Multispectral sensor를 사용한 기술이다. Multispectral Pedestrian Detection은 몇 년 동안 CNN을 기반으로 보행자 감지 성능을 향상시켜 왔다. 보행자 감지는 컴퓨터 비전의 필수적이고 중요한 요소로 활발한 연구 영역에 속한다. 객체 감지에 효과적인 CNN을 발전시켜 나온 다양한 방법들 중 Faster R-CNN, R-FCN과 같은 2단계 방법은 처리 속도가 상대적으로 느린 반면 성능이 좋은 결과를 얻을 수 있고, SSD, YOLO와 같은 1단계 방법은 실시간 작동이 가능하지만 결과가 비교적 만족스럽지 않다. 따라서 CNN을 기반으로 제안된 방법들에 대한 동향에 대해 다룰 것이다.

2. Multispectral Pedestrian Detection

Multispectral Pedestrian Detection 성능 구현을 위해서 많은 논문에서는 Kaist Dataset을 사용하여 독자적인 architecture를 구상해왔다. Kaist Dataset은 103,123개의 주석과 1,182개의 보행자 인스턴스가 포함된 95,328개의 정렬된 가시열 이미지 페어로 구성되어 있어 환경이 다양하고, 보행자 숫자, 밤과 낮이 다채롭다. 성능 측정에는 (그림 1)과 같이 miss rate를 기준으로 삼았다. FPPI (False positive per image)와 MR(miss rate)에 관한 그래프로, 성능 측정은 10^-2 ~10^0 사이의 구간의 miss rate를 사용한다. Miss rate는 오검출률과 잘못 찾은 것에 대한 지표를 동시에 보기 위함으로 정확도가 높아도 잘못 찾은 데이터가 많은 경우 성능이 좋다고 볼 수 없기 때문이다.



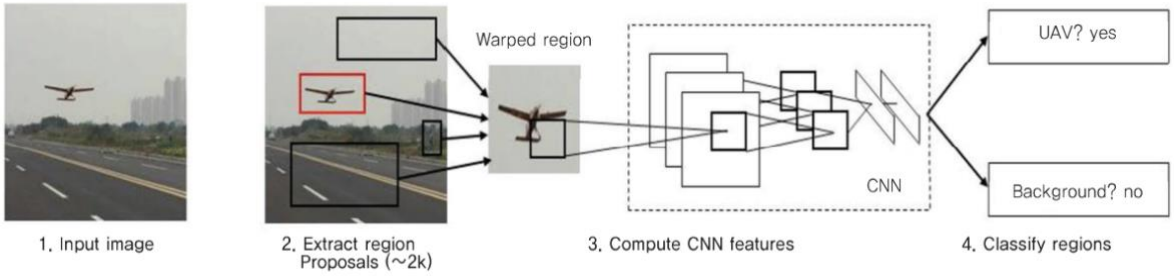
(그림 1) Vanilla ConvNet Comparison of detection results [1]

2.1. CNN(Convolutional Neural Network)

보행자 인식을 위해 사용되는 CNN은 객체 인식에 관한 방법과 모델이 많이 있다. CNN을 기반으로 한 객체 인식 방법으로는 R-CNN을 시작으로 SSD까지 다양하게 발전되어 왔고 사용되고 있다. 보행자 인식 이전에 기본적으로 제안되는 객체 인식 방법들로, 최근 많은 논문에서는 제안하는 Architecture의 성능 비교를 위해 사용되고 있다.

가. R-CNN

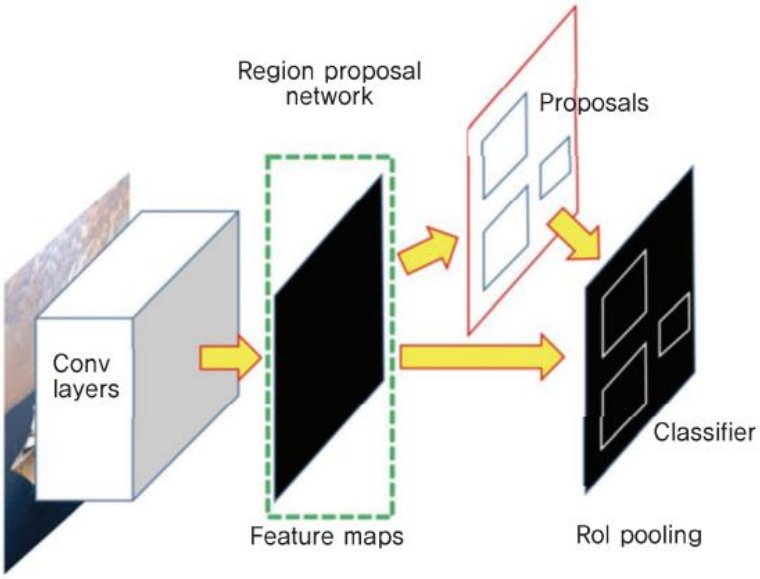
후보 영역을 생성하고 이를 기반으로 CNN을 학습시켜 영상 내 객체의 위치를 찾아내는 방법으로 CNN을 통해 특징을 추출하고, 추출된 특징을 이용해 후보 영역 내의 객체를 SVM(Support Vector Machine)을 이용하여 분류하게 된다. (그림 2)는 R-CNN의 객체 인식 시스템 과정이다.



(그림 2) R-CNN의 객체 인식 시스템

나. Faster R-CNN

Faster R-CNN 이전에는 Fast R-CNN이 있으며, Fast R-CNN은 CNN, SVM, 회귀의 학습 단계가 모두 분리되어 있어 많은 시간이 소요된다. 소요되는 시간을 감소시키기 위해 Softmax를 사용하였지만, 후보 영역을 생성하는 알고리즘이 CNN 외부에서 수행되는 비효율적인 면을 개선하기 위해 고안된 것이 Faster R-CNN이다. Faster R-CNN은 후보영역을 생성하는 데에 선택적 탐색 알고리즘을 이용하지 않고, (그림 3)과 같이 Feature Map을 추출하는 CNN의 마지막 층에 후보 영역을 생성하는 별도의 CNN인 영역 제안 네트워크(RPN)을 적용하였다. 해당 과정을 통해 Fast R-CNN 보다 훈련 시간을 10배 감소시키며 평균 정밀도(mAP) 또한 향상시켰다.



(그림 3) Faster R-CNN의 객체 인식 시스템

다. R-FCN

위치 정보를 포함하고 있는 Score Map을 이용하여 물체의 위치를 정확하고 효율적으로 찾아내는 방식으로 CNN을 통해 추출된 Feature Map으로부터 얻어지는 Score Map은 입력된 영상 내 특정 위치의 정보를 포함하고 있다. 특정 위치가 찾고자 하는 객체를 포함할 경우 Score Map의 반응이 커지고, 그렇지 않은 경우 반응이 작아지는 것을 이용한다. R-FCN은 훈련이 요구되지 않는 장점이 있어 Faster R-CNN보다 훈련 시간이 3배 빠르다.

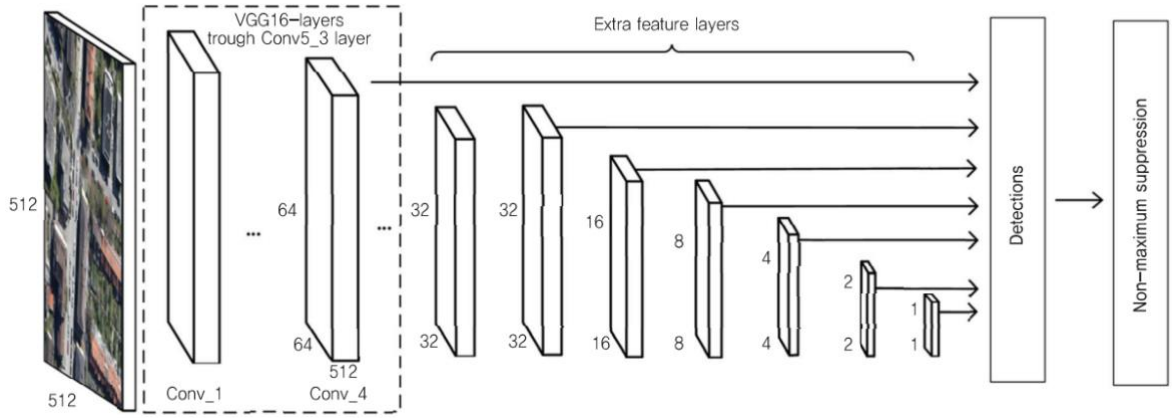
라. YOLO

객체 인식 문제를 하나의 회귀 문제로 접근하여 전체적인 구조를 간소화함으로써 훈련 및 검출 속도를 크게 향상시킨다. 입력된 영상을 CNN을 거쳐 Tensor 형태로 출력되어 영상을 격자 형태로 나눠 각 구역을 표현하여 해당 구역의 객체를 인식한다. YOLO는 영역 추출을 위한 별도의 네트워크가 필요하지 않아 Faster R-CNN보다 월등히 빠른 훈련 속도를 보이지만 인식 정확도가 떨어지며, 작은 물체 인식에 어려움이 있다.

마. SSD

후보 영역을 생성하기 위한 RPN을 따로 훈련시키지 않고 다양한 크기의 Feature Map을 이용하여 객체를 인식한다. (그림 4)와 같이 합성곱 층이 진행됨에 따라 크기가 줄어들며, 이 Feature Map들을 추론 과정에서 사용하여 객체를 인식한다. 크기가 큰 Feature Map은 작은 물체를 검출할 수 있고, 작은 Feature Map은 큰 물체를 검출

할 수 있다. RPN을 제거하였기에 Faster R-CNN보다 훈련 속도가 빠르며, 다양한 Feature Map을 사용함으로써 YOLO보다 정확하게 객체를 인식할 수 있다.

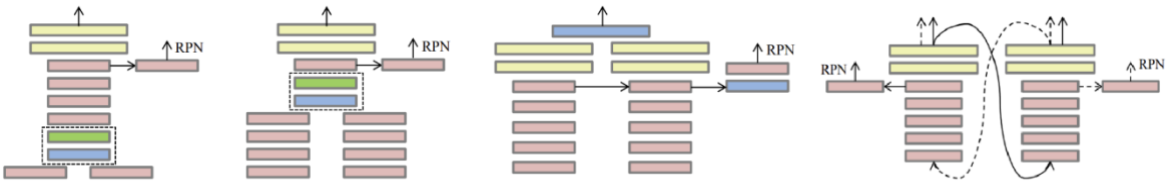


(그림 4) SSD의 객체 인식 시스템

2.2. Architecture

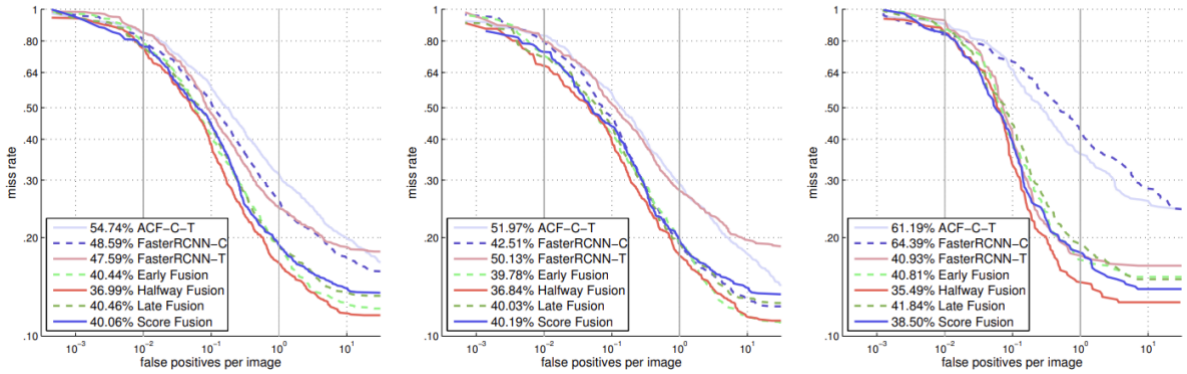
가. Halfway fusion

Halfway Fusion을 고안한 논문[2]에서는 (그림 5)와 같이 Early Fusion, Halfway Fusion, Late Fusion, Score Fusion도 같이 고안하였다.



(그림 5) 왼쪽에서 오른쪽으로 Early Fusion, Halfway Fusion, Late Fusion, Score Fusion

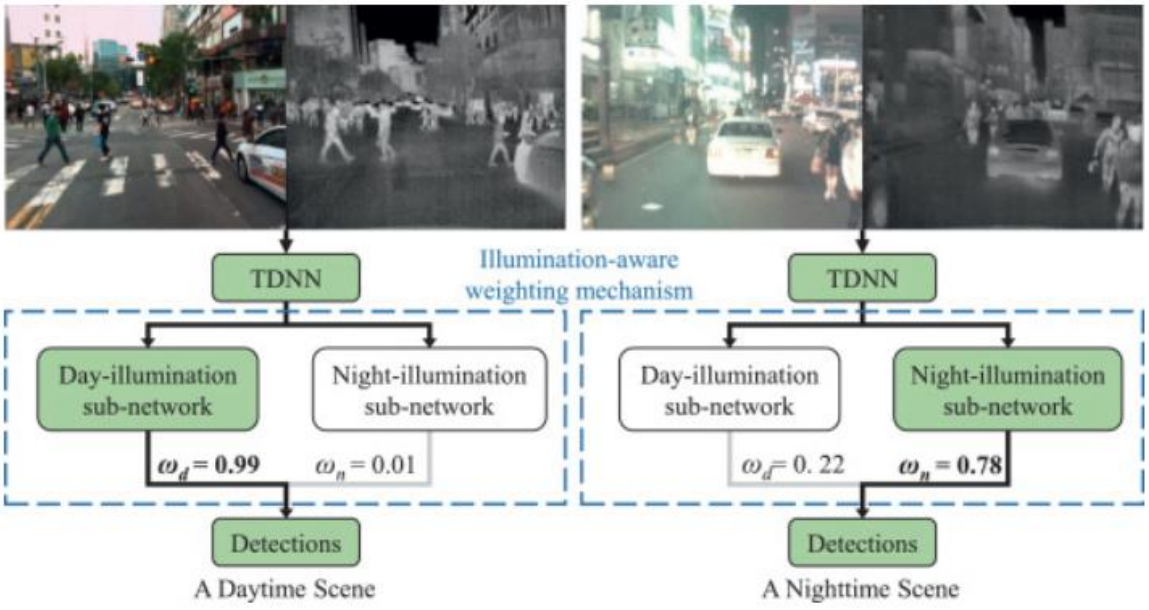
그 중 가장 성능이 높은 architecture는 중간 레벨에서 Fusion하는 Halfway Fusion architecture가 Baseline이 되었다. (그림 5)와 같이 Red Box와 Yellow Box인 컨볼루션 및 완전 연결 레이어들의 중간에서 Blue Box인 연결 레이어와, Green Box인 다음을 위해 사용되는 NIN(Network-in-Network)가 중간에 존재한다. Halfway Fusion은 서로 다른 DNN 단계에서 두 분기의 ConvNet을 통합하는 Fusion architecture로 ConvNet이 효과적인 결과를 보여준다는 것을 (그림 1)을 통해 증명하였다. Kaist Dataset을 사용하여 고안된 Halfway Fusion architecture의 성능은 (그림 6)과 같다. Faster R-CNN을 비롯하여 고안된 다른 architecture(Early Fusion, Late Fusion, Score Fusion)과 비교하여도 Halfway Fusion의 검출 결과가 좋음을 확인할 수 있다.



(그림 6) 검출 결과 비교 그래프, 왼쪽부터 오른쪽으로 All-day, Daytime, Nighttime

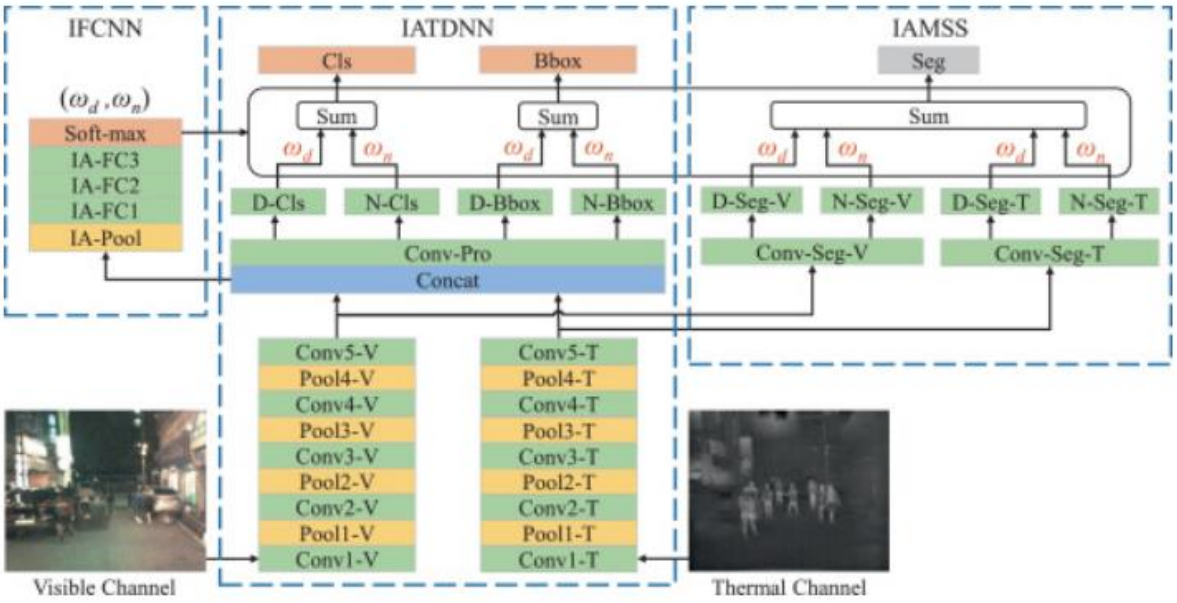
나. IATDNN

조명 인식 심층 신경망(IADNN)[4]을 통해 다양한 조명 조건에서 학습가능한 프레임워크를 제안한다. (그림 7)은 제안된 조명 인식 가중치 메커니즘이다. Multispectral 이미지가 주어지면 TDNN은 다중 스펙트럼 Feature Map을 생성한다. 주간 및 야간 조명 하위 네트워크의 출력을 계산된 조명 인식 가중치와 융합하여 감지를 생성한다.



(그림 7) 조명 인식 가중치 메커니즘

Kaist Dataset을 사용했으며, 다중 스펙트럼 집계 기능(ACF+T+THOG)을 추출하는 새로운 기술과 대상 분류를 위한 boosted decision trees(BDT)를 적용했다. 제안하는 모델에서는 테스트 단계에서 가시 이미지만 고려하기 때문에 색상 및 열 데이터를 모두 사용하는 다른 Multispectral 검출기와 비교하기 어렵다. 따라서 탐지 성능을 향상시키기 위해 CWF(Channel Weighting Fusion) 레이어를 개발했다. 해당 논문은 Multispectral 보행자 감지기의 성능을 향상시키기 위해 조명 정보를 탐색하는 작업이다.



(그림 8) 조명 인식 다중 스펙트럼 심층 신경망 아키텍쳐

(그림 8)은 조명 인식 다중 스펙트럼 심층 신명망의 아키텍처이다. Green Box는 컨볼루션 계층과 완전 연결 계층, Yellow Box는 풀링 계층, Blue Box는 융합 계층, Gray Box는 분할 계층, Orange Box는 출력 계층을 의미한다. IFCNN(Illumination fully connected neural networks)은 조명 완전 연결 신경망으로 IATDNN(Illumination-aware two-stream deep convolutional neural networks)은 조명 인식 2 스트림 심층 컨볼루션 신경망으로 융합 계층 Concat을 통해 생성된 Feature Map(TSFM)을 입력으로 받아 조명 인식 가중치를 계산해 조명 조건을 결정한다. 이 가중치는 다시 IATDNN에서 4개의 하위 네트워크의 출력과 결합되어 최종 감지 결과를 생성한다. IAMSS(Illumination-aware multispectral semantic segmentation)는 조명 인식 다중 스펙트럼 의미론적 분할로 multispectral 이미지에서 보행자 감지와 분할을 동시에 하기 위해 IATDNN과 통합한다.

Module	Layers	Ouput size	Operation
IATDNN	Conv5-V/T	60X48X512	
	Concat	60X48X1024	Concatenation
	Conv-Pro	60X48X512	3X3 conv
	D/N-Cls	60X48X18	1X1 conv
	Cls	60X48X18	sum
	D/N-Reg	60X48X36	1X1 conv
	Reg	60X48X36	sum
IFCNN	IA-Pool	7X7X512	interpolation
	IA-FC1	512	inner product
	IA-FC2	64	inner product
	IA-FC3	2	inner product

IAMSS	Conv-Seg-V/T	60X48X512	3X3 conv
	D/N-Seg-V/T	60X48X2	1X1 conv
	D/N-Seg-V/T	60X48X2	sum

(표1) 학습 정보

	Daytime	**Nighttime**
IFCNN-V	97.94%	97.11%
IFCNN-T	93.13%	94.48%
IFCNN	98.35%	99.75%

(표2) IFCNN 조도 예측 정확도

(표1)은 학습을 위한 각 단계의 세부 정보이다. 해당 조건으로 구성 후 가시화상(IFCNN-V), 열화상(IFCNN-T), IFCNN에 관한 조도 예측의 정확도는 (표2)와 같다. IFCNN의 성능을 확인 후 IATDNN의 성능이 계산된 MR은 (표3)과 같다. 마지막 단계인 IATDNN과 통합할 IAMSS에 관한 단계별 MR의 결과는 (표4)와 같다.

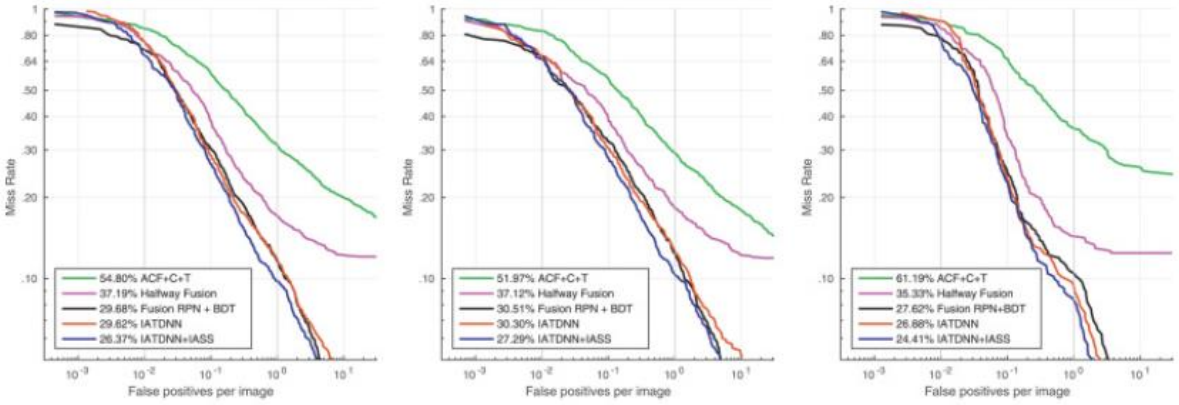
	All-day	**Daytime**	**Nighttime**
TDNN	32.60%	33.80%	30.53%
IATDNN	29.62%	30.30%	26.88%

(표3) IATDNN MR

	All-day	**Daytime**	**Nighttime**
IATDNN	29.62%	30.30%	26.88%
IATDNN+MSS-F	29.17%	39.62%	26.96%
IATDNN+MSS	27.21%	27.56%	25.57%
IATDNN+IAMSS-F	28.51%	28.98%	27.52%
IATDNN+IAMSS	26.37%	27.29%	24.41%

(표4) IAMSS MR

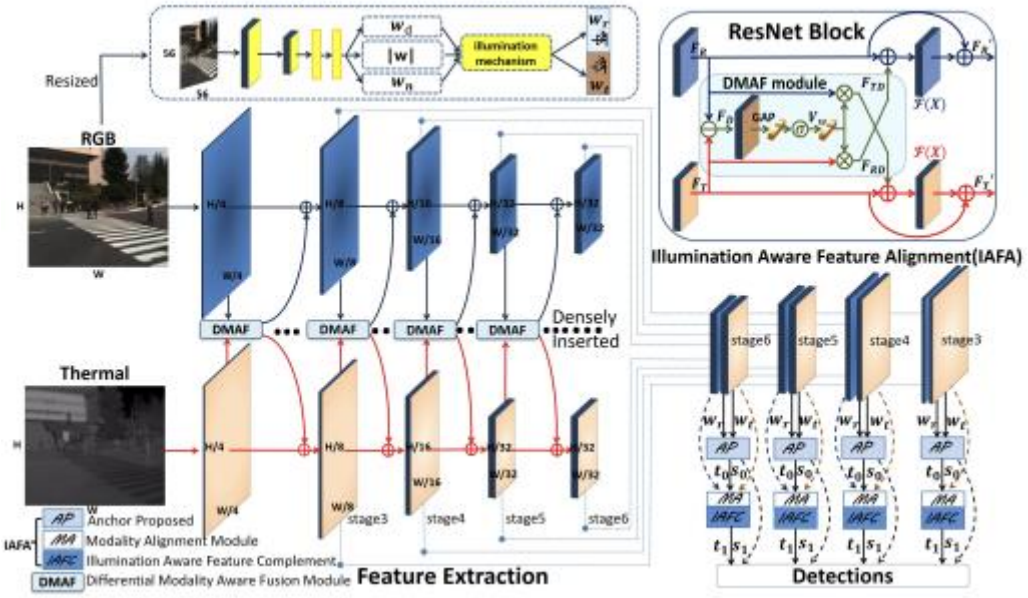
제안된 모델을 다른 모델과 비교한 그래프는 (그림 9)와 같다. 그래프와 같이 IATDNN+IAMSS 모델이 Halfway fusion에 비해 좋은 성능과 적은 런타임을 가진다.



(그림 9) 모델 비교 MR 그래프, 왼쪽부터 All-day, Daytime, Nighttime

다. MBNet

Multi-modality에서 객체 감지의 일반적인 최적화 프로세스에 관해서는 입력 불균형 문제가 있으며 대표적으로 전경과 배경의 불균형이 있다. 일반적으로 손실을 최소화하기 위해 각 손실 함수에 계수를 추가했다. 해당 논문은[5] Modality Balance를 구성하여 SSD를 기반으로 한 MBNet을 통해 두 가지 방식의 특성을 별도로 추출한다. 그 후 서로 다른 규모의 기능을 완전히 융합하기 위해 DMAF(Differential Modality Aware Fusion) 모듈을 제안하여 RGB와 열 감지의 Feature Map의 차이를 활용하여 보완한다.



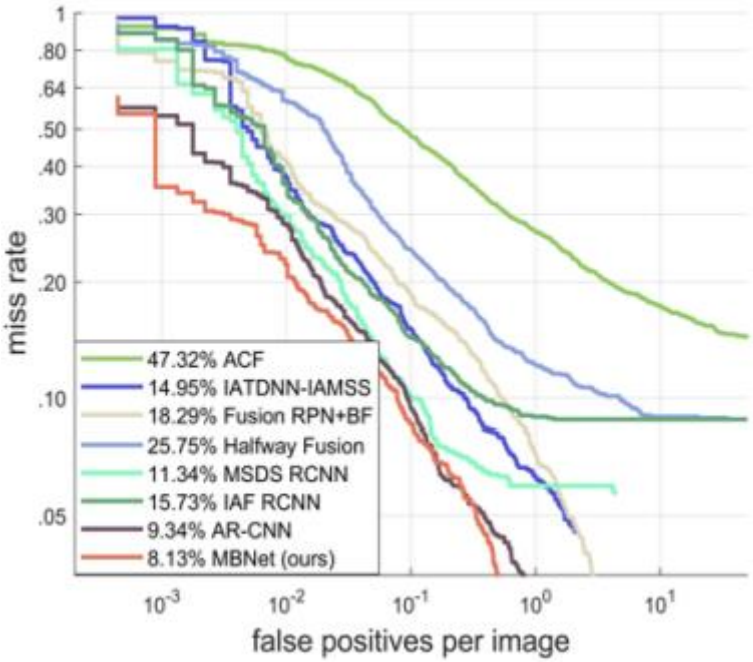
(그림 10) MBNet 개요 프레임 워크

(그림 10)은 Modality Balance Network(MBNet)의 개요 프레임워크이다. 특징 추출 모듈, 조명 인식 특징 정렬 모듈, 조명 메커니즘 모듈 3개의 부분으로 나뉘어진다. 특징 추출 모듈은 ResNet50을 backbone network로 채택하고 DMAF 모듈을 내장해 modality를 보완한다. 조명 메커니즘은 조명 값을 획득하도록 설계하여, 2개의 modality stream에 가중치를 할당한다. 조명 인식 특징 정렬 모듈에서 모델을 다양한 조명 조건에 적응시키는 역할을 하고 지역 제안 단계에서 두 가지 양식 기능을 정렬한다.

(표5)는 Kaist Dataset의 방법론과 MR을 비교한 것으로, 그래프로 나타내면 (그림 11)과 같다.

Methods	MR^{-2} (IoU = 0.5)			MR^{-2}(IoU = 0.75)			Plateform	Speed(s)
	All	Day	Night	All	Day	Night		
ACF	47.32	42.57	56.17	88.79	87.70	91.22	MATLAB	2.73
Halfway Fusion	25.75	24.88	26.59	81.29	78.43	86.80	TITAN X	0.43
Fusion RPN+BF	18.29	19.57	16.27	72.97	68.14	81.35	MATLAB	0.80
IAF R-CNN	15.73	14.55	18.26	75.50	72.34	81.12	TITAN X	0.21
IATDNN + IASS	14.95	14.67	15.72	76.69	76.46	77.05	TITAN X	0.25
RFA	14.61	16.78	10.21	-	-	-	TITAN X	0.08
CIAN	14.12	14.77	11.13	74.45	71.42	80.16	1080 Ti	**0.07**
MSDS-RCNN	11.34	10.53	12.94	70.57	67.36	79.25	TITAN X	0.22
AR-CNN	9.34	9.94	8.38	64.22	57.87	76.82	1080 Ti	0.12
MBNet(ours)	**8.13**	**8.28**	**7.86**	**60.12**	**54.90**	**68.34**	1080 Ti	**0.07**

(표5) 다른 방법론과 MR 비교 데이터



(그림 11) 다른 방법론과 MR 비교 그래프

	Methods	MR^{-2}				Methods	MR^{-2}		
		Day	Night	All			Day	Night	All
Visible	SVM	37.6	76.9	-	Visible+ Thermal	MACF	61.3	48.2	60.1
	DPM	25.2	76.4	-		Choi et al.	49.3	43.8	47.3
	Random Forest	26.6	81.2	-		Halfway Fusion	38.1	34.4	37.0
	ACF	65.0	83.2	71.3		Park et al.	31.8	30.8	31.4
	Faster R-CNN	43.2	71.4	51.9		AR-CNN	24.7	18.1	22.1
						MBNet (ours)	**24.7**	**13.5**	**21.1**

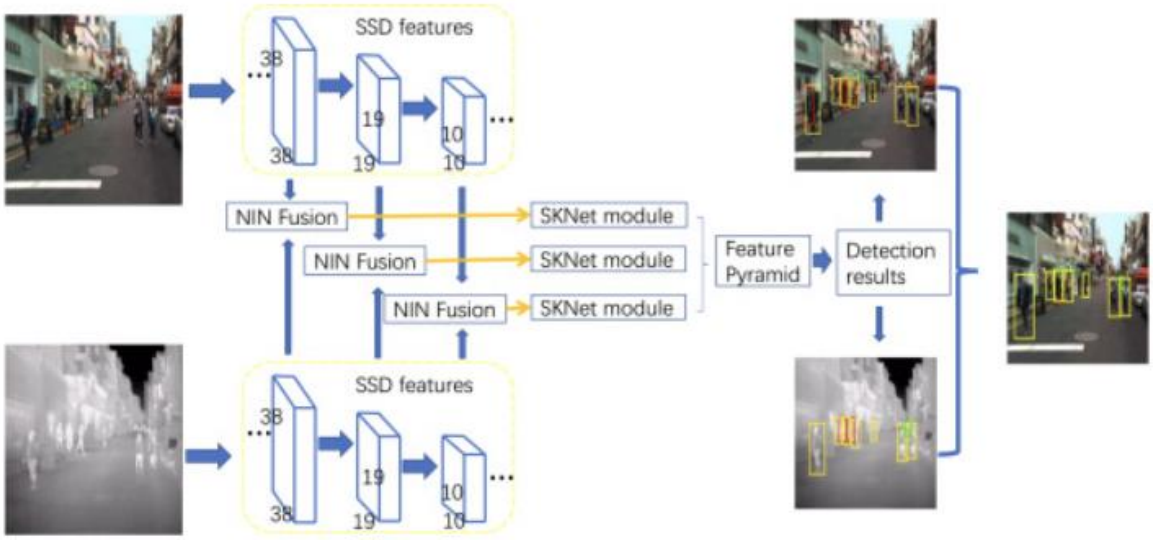
(표 6) CVC-14 Dataset 비교 데이터

CVC-14 Dataset에 관한 평과 결과는 (표6)와 같다. RGB 방식은 낮에 유리하고, 열 감지

는 밤에 유용한 것에 따라, 조명 인식 메커니즘으로 통해 양식 불균형 문제를 해결할 필요가 있다. 해당 문제를 완화하고자 제안된 MBNet으로 탐지한 결과 DMAF 모듈은 ResNet에 통합되며, MA 모듈은 RGB 및 열 탐지 데이터가 동일한 Feature를 가질 수 있도록 두 가지 양식을 정렬한다. Backbone 네트워크에 내장된 조명 게이트와 영역 제안 단계의 적응형 조명 인식 기능 보완으로 변형된 조명에 강한 것을 확인할 수 있다.

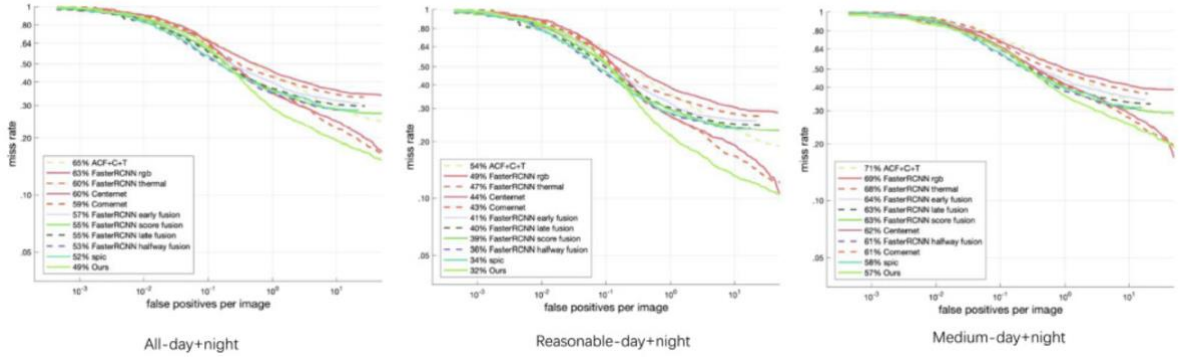
라. Reasonable Fusion

CNN 모델에서 각 레이어의 공유 수용 필드는 크기가 동일하기 때문에 다중 스케일 보행자 감지 결과가 제한된다. 이를 조정하기 위한 동적 선택 기법으로 NIN을 사용하였으며, 커널 크기가 다른 선택적 커널 단위를 사용하는 선택적 커널 네트워크(SKNets)를 사용했다. SSD를 사용하여 Feature를 추출하여 진행되는 방식으로 작업 흐름은 (그림 12)과 같다.[6]



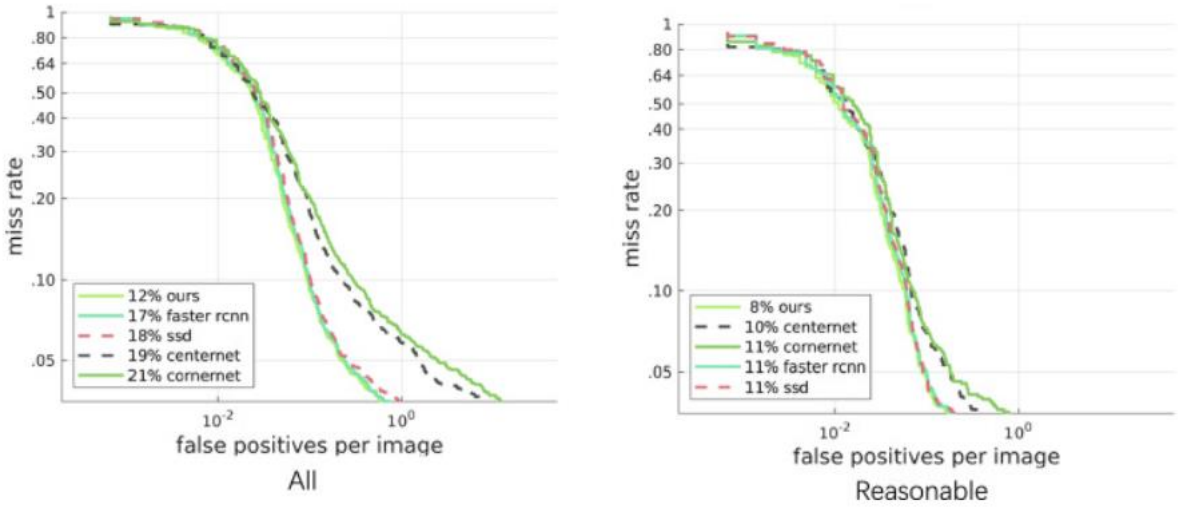
(그림 12) Feature 추출 작업 흐름도

훈련 및 테스트 이미지 크기는 600X600 픽셀로, 60,000번의 반복에 대해 학습률을 10^-3으로 설정했다. 단순한 융합이 아닌 합리적인 상황에서의 융합 검출 모델이 단일 방식보다 우수함을 3가지 상황을 통해서 최상의 성능을 달성했다. 먼저, 주야간 환경에서 제안된 방법과 베이스라인 방법을 비교한 결과는 (그림 13)과 같다.

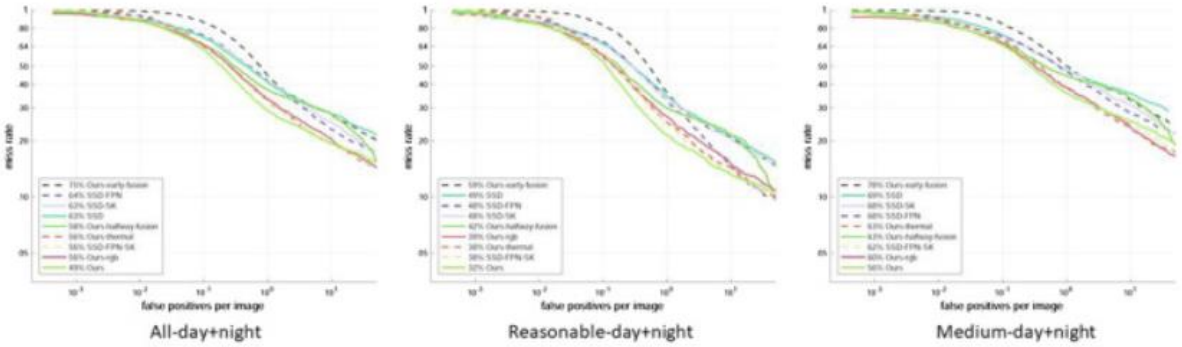


(그림 13) 주야간 환경 FPPI MR 비교 그래프

(그림 13)에 있는 Faster R-CNN halfway-fusion, Faster R-CNN early-fusion, Faster R-CNN late-fusion, Faster R-CNN score-fusion은 (가. Halfway Fusion)에서 언급된 알고리즘을 사용한 Faster R-CNN 기반 감지이다. 제안된 방식이 Faster R-CNN halfway-fusion보다 우수한 결과를 달성하고 있다. (그림 14)은 다른 베이스라인과의 비교 그래프로 모든 상황과 합리적인 상황에 대한 그래프이다. (그림 15)는 Kaist dataset에서 다양한 방식으로 융합한 모델의 비교 검출에 관한 그래프이다. SSD-FPN은 SKNet 구성요소가 없을 때의 제안된 모델, SSD-SK는 FPN 구성 요소가 없을 때의 제안된 모델이다. 3가지 그래프를 통해 각 계층의 기능 정보를 결합하는데 사용된 SKNet 모듈이 수용 필드를 적응적으로 조정하는데 효과적이며, 적응형 융합 알고리즘은 가시광선 정보와 열 정보를 효과적으로 통합할 수 있다.



(그림 14) All, Reasonable 상황에서 FPPI MR 그래프



(그림 15) 다른 Baseline 모델과의 MR 비교 그래프

3. 성능 비교

Multispectral Pedestrian Detection의 baseline이 되고, fusion에 관한 방향을 제안한 Halfway fusion을 사용하게 되면 Multispectral Pedestrian Detection을 구현하는 데에 있어 기본적인 성능을 재현해 볼 수 있다. Multispectral Pedestrian Detection의 새로운 발상을 위한 조명 인식 다중 스펙트럼 심층 신경망을 통해 Multispectral Pedestrian Detection에 조명에 관한 탐지가 굉장히 많은 영향을 미친다는 것을 확인할 수 있었다. Halfway fusion 에 비해 좋은 성능을 확인했다. 또한 modality의 불균형 문제를 제안한 MBNet 역시 조명에 관한 언급이 있었으며 모델 구성에도 포함되어 있었다. IATDNN에 비해 좋은 성능을 선보였다. 전체 상황이 아닌 합리적인 상황에서 가장 좋은 결과를 내는 것은 새로운 융합한 모델을 통해 결과를 낸 Reasonable fusion 모델로, 기존의 다른 모델과의 비교해봤을 때, Reasonable한 상황에 대한 분석에 대해 제안한다.

4. 결론

Multispectral Pedestrian Detection을 해결하기 위해 고려해야만 하는 것은 낮과 밤의 조명 차이와 함께 RGB와 열 센서의 차이점으로 인한 불균형 문제이다. 해당 문제들을 언급하고 해결하고자 제안하는 CNN 기반의 다양한 모델들이 있으며, 더욱 발전 가능한 양상을 보이고 있다.

5. 참고 사항

[1][Multispectral Pedestrian Detection: Benchmark Dataset and Baseline][2015]

[2][Multispectral Deep Neural Networks for Pedestrian Detection][2016]

[3][딥러닝 기반 객체 인식 기술 동향][2018]

[4][Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection][2019]

[5][Improving Multispectral Pedestrian Detection by Addressing Modality Imbalance Problems][ECCV 2020]

[6][A robust and fast multispectral pedestrian detection deep network][2021]