KIPS Trans. Softw. and Data Eng.
Vol.10, No.11 pp.439~448
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2021.10.11.439

Distracted Driver Detection and Characteristic Area Localization by Combining CAM-Based Hierarchical and Horizontal Classification Models

Sooyeon Go† · Yeongwoo Choi††

ABSTRACT

Driver negligence accounts for the largest proportion of the causes of traffic accidents, and research to detect them is continuously being conducted. This paper proposes a method to accurately detect a distracted driver and localize the most characteristic parts of the driver. The proposed method hierarchically constructs a CNN basic model that classifies 10 classes based on CAM in order to detect driver distration and 4 subclass models for detailed classification of classes having a confusing or common feature area in this model. The classification result output from each model can be considered as a new feature indicating the degree of matching with the CNN feature maps, and the accuracy of classification is improved by horizontally combining and learning them. In addition, by combining the heat map results reflecting the classification results of the basic and detailed classification models, the characteristic areas of attention in the image are found. The proposed method obtained an accuracy of 95.14% in an experiment using the State Farm data set, which is 2.94% higher than the 92.2%, which is the highest accuracy among the results using this data set. Also, it was confirmed by the experiment that more meaningful and accurate attention areas were found than the results of the attention area found when only the basic model was used.

Keywords : Distracted Driver Detection, Convolutional Neural Networks, Class Activation Maps, Attention Area Localization

CAM 기반의 계층적 및 수평적 분류 모델을 결합한 운전자 부주의 검출 및 특징 영역 지역화

고 수 연† · 최 영 우††

요 약

교통사고 원인 중 가장 큰 비율을 차지하는 것이 운전자의 부주의로서 이를 검출하는 연구가 꾸준히 진행되고 있다. 본 논문은 부주의한 운전자를 정확히 검출하고, 검출된 운전자의 모습에서 가장 특징적인 영역을 선정(Localize)하는 방법을 제안한다. 제안하는 방법은 운전자의 부주의를 검출하기 위해서 CAM(Class Activation Map) 기반의 전체 클래스를 분류하는 CNN 모델과 이 모델에서 혼동하거나 공통된 특징 영역을 갖는 클래스들에 대한 상세 분류가 가능한 네 개의 서브 클래스 CNN 모델을 계층적으로 구성한다. 각 모델에서 출력한 분류 결과는 CNN 특징맵들과의 매칭 정도를 표현하는 새로운 특징으로 간주해서 수평적으로 결합하고 학습하여 분류의 정확성을 높였다. 또한 전체 및 상세 분류 모델의 분류 결과를 반영한 히트맵 결과를 결합하여 이미지의 특징적인 주의 영역을 찾아낸다. 제안한 방법은 State Farm 데이터 셋을 이용한 실험에서 95.14%의 정확도를 얻었으며, 이는 기존에 동일한 데이터 셋을 이용한 결과 중 가장 높은 정확도인 92.2%보다 2.94% 향상된 우수한 결과이다. 또한 전체 모델만을 이용했을 때 찾아진 주의 영역보다 훨씬 의미 있고 정확한 주의 영역이 찾아짐을 실험으로 확인하였다.

키워드 : 운전자 부주의 검출, 합성곱신경망, CAM(Class Activation Map), 주의영역 지역화

1. 서 론

2018년 세계 보건기구(WHO)의 연구 결과에 따르면 전 세계에서 24초에 1명꼴인 135만명이 매년 교통사고로 숨지고 있다고 한다[1]. 2013년부터 2015년까지 교통사고를 다뤘던 이전 보고서에서는 연간 사망자 수가 125만명 선이었다. 이는 인구와 차량의 증가로 인해 교통사고 역시 꾸준히 늘어나고 있음을 보여주는 것이다. 또한 미국 고속도로 교통안전국(NHTSA)에서는 2015년에 운전자 부주의로 인한 교통사고로 3,477명이 사망하고 391,000명이 중상을 입었다고 한다[2]. 심지어 운전 중 가벼운 통화를 하는 것은 집중력의 40%를 낮추며 이는 혈중 알코올 농도 0.1%의 주취 상태에서 운전하는 것과 같은 수준으로 영향을 미친다. 또한 한국 도로교통공단에 따르면 운전 중 스마트폰 사용은 운전에 대한 집중도

※ 이 논문은 한국연구재단 기초연구과제에 의하여 연구되었음(No. NRF-2017 R1D1A1B04035633).
† 비 회 원 : 숙명여자대학교 컴퓨터과학과 석사과정
†† 정 회 원 : 숙명여자대학교 컴퓨터과학과 교수
Manuscript Received : August 27, 2021
First Revision : October 12, 2021
Accepted : October 12, 2021
* Corresponding Author : Yeongwoo Choi(ywchoi@sookmyung.ac.kr)

와 주의력을 감소시켜 평소보다 제동거리가 길어지며 급정지로 인해 교통사고 위험이 약 4배 정도 높아진다고 한다.

자율주행 자동차의 출현으로 운전자 부주의에 대한 문제가 궁극적으로는 해결되겠지만, 현재까지는 거의 대부분 운전자가 차량 운행에 중요한 역할을 담당하고 있기 때문에 운전자의 부주의를 검출하고 이를 안전 운행에 이용하는 것은 아주 중요한 연구개발 주제이다. 운전자 부주의 검출 연구를 위해서 중국 Southeast University에서 만든 SEU 데이터 셋과[3] 2016년 Kaggle 컨테스트에 사용된 State Farm 데이터셋[4], 이집트 American University에서 만든 AUC 데이터셋[5] 등을 중심으로 운전자의 부주의 검출 연구가 활발하게 진행되고 있다.

본 논문에서는 CAM(Class Activation Map)[6] 기반의 10개 클래스의 부주의 상태를 분류하는 CNN(Convolutional Neural Network) 모델과 이를 세분화한 4개의 서브 클래스 CNN 모델을 함께 결합하여 운전자의 부주의 검출 정확성을 높이며, 이 상태에서 운전자의 어떤 행동 모습이 분류 결과에 중요한 역할을 하는가를 결정하는 영역을 찾아내는(Localize) 방법도 제안한다. 제안하는 방법에서 상세 분류를 위한 서브 클래스 모델은 10개 클래스를 분류하는 CNN 모델이 자주 혼동하거나 핸드폰을 조작하는 등의 동일한 행동 패턴의 운전자 상태 등을 기준으로 만든다. 10개 클래스 모델과 서브 클래스 모델들의 마지막 층인 완전 연결(Fully Connected, FC)층을 GAP(Global Average Pooling) 층으로 변경하여 CAM을 구성한다. 이후 모든 모델에서 GAP 층을 통과한 값들을 입력 특징으로 하는 한 개의 은닉층으로 구성된 완전 연결 신경망을 구성하여 학습하고 최종 분류한다. 또한 GAP 층에서 CNN 특징맵들과의 매칭 정도를 집계한 후 이미지 내의 운전자의 움직임을 주의하는(attention) CAM을 이용하여 입력된 이미지에 관한 10개 클래스 모델과 서브 클래스 모델들의 특징맵을 각각 구한다. 이를 결합하여 새롭게 확장된 히트맵을 만들고 여기서 분류된 결과에 대한 이미지의 중요한 영역을 구체적으로 주의하도록 하였다. 이 과정을 통해서 10개 클래스 CNN 모델보다 훨씬 높은 정확성의 분류 결과와 구체적인 특징 영역을 얻을 수 있었다.

논문은 2절에서 관련 연구, 3절에서 제안 방법을 구체적으로 설명하고, 4절에서 제안한 방법에 적용한 실험 및 결과를 보여준다. 끝으로 결론과 향후 연구를 5절에서 언급한다.

2. 관련 연구

자동차 안에 카메라를 설치하고 이 영상을 판독해서 운전자의 부주의 여부를 판단하는 연구들이 있다[7-10]. Alotaibia[7]는 딥러닝 모델인 ResNet(Residual Network)[11], HRNN(Hierarchical Recurrent Neural Network)[12], Inception[13] 모듈을 모두 결합한 모델을 이용하여 State Farm 데이터 셋을 대상으로 운전자 부주의를 검출하는 연구를 수행하였다. 이 방법으로 기존의 ResNet, HRNN만을 각각 사용한 모델에 비해 96.23%라는 높은 분류 정확도를 얻었다. 또한 Masood[14]는 사전 학습된 VGG16[15] 모델과 VGG19[15] 모델을 이용하여 운전자 부주의를 검출하였으며, 각 모델에서 99.57%, 99.98%의 높은 분류 정확도를 얻었다. 그러나 이 논문들[7, 14]에서 사용한 테스트는 동일한 사람의 연속적인 움직임이 프레임별로 들어있는 학습 데이터 셋을 학습 데이터와 테스트 데이터로 나누어 성능을 평가하여, 제안한 방법의 성능이 우수하다고 평가하기에는 신뢰성이 높지 않다.

Lu[8]는 기존의 Faster-RCNN[16] 모델보다 변형이 가능하고 확장된 Deformable and Dilated Faster RCNN(DD-RCNN) 모델을 제안하여 운전자 부주의를 여부를 판단했다. 이 모델을 통해서 이미지 내의 크기가 작고 불규칙한 모양의 휴대폰이나 컵 등의 특징을 추출해 운전자의 동작을 분류하여 92.2%의 정확도를 얻었다. 기존의 Faster RCNN 모델의 Grad-CAM[17] 결과보다 새로운 모델의 Grad-CAM 결과에서 휴대폰이나 물병 등과 같은 이미지 내의 작은 요소들을 잘 찾아내는 것을 확인할 수 있다. Lu[9]와 Moslemi[10]는 연속적인 모션을 취하고 있는 동일한 사람의 프레임을 모아 운전자 부주의를 검출하였다. 그 결과 단일 프레임의 정확도는 높지 않지만 연속 프레임을 이용하여 해당 이미지 자체의 정확도를 높이는 결과를 만들었다.

Hesham[5]은 State Farm 데이터 셋과 비슷한 총 10개 클래스의 운전자 부주의 검출을 위한 AUC 데이터를 만들었다. 제안한 방법은 운전자의 얼굴과 손, 피부를 검출하고 앙상블 학습을 이용해서 운전자 이미지를 분류하여 높은 정확도를 얻었지만, 앙상블 학습을 위해 여러 개의 모델을 학습해야 하고 이미지에서 얼굴, 피부, 손 등을 정확히 검출해야 하는 이미지 분리 방법들에 크게 의존하는 단점이 있다.

본 논문에서는 운전자 부주의를 검출할 뿐만 아니라 분류된 운전자의 움직임에서 특징이 되는 영역을 찾는 CAM을 기반으로 한 WSOL(Weakly Supervised Object Localization) 연구를 소개한다.

WSOL은 이미지와 객체의 라벨만이 주어졌을 때 이미지 안에 있는 객체의 위치를 예측한다. 또한 이미지의 어떤 부분으로 인해서 특정 라벨로 분류 되는지를 CAM[6]으로 시각화하여 확인할 수 있다. 그러나 CAM은 주어진 라벨을 기반으로 부분적인 특징맵의 비중을 과도하게 의존하여 지역화 맵을 생성하는 것과 이미지 내의 레이블 된 객체의 부분 영역만을 찾는 등의 단점이 있다. 예를 들어 주어진 이미지의 운전자를 “옆 사람과 대화하는 중”이라고 분류하고자 할 때, CAM의 히트맵에서 운전자의 옆을 바라보는 상체가 특히 강조되고 옆을 바라보는 얼굴은 비교적 적게 주의되어 최종 지역화맵에서 이미지를 특정 클래스로 분류한 이유를 전체적으로 파악하지 못하게 되는 경우가 빈번하게 발생한다. 따라서 이미지 클래스를 분류할 때 그에 대한 부분적인 이유가 아닌 전체적인 이유를 알기 위한 방법들이 제안되고 있다[18-20]. 이 방법들을 통해서 지역화 맵에 표현된 주의 영역의 일부 또는 전부를 지우고, 지운 영역의 재학습을 통해 이미지 사이의 분류 기준이 되는 영역을 전반적인 영역으로 주시하고자 한다.

구체적으로 Hide and Seek(HaS)[18] 방법은 입력한 이미지 내의 일부를 임의로 지우고, 지운 이미지들로 재학습하여 전보다 더 넓은 주의 영역을 얻는다. CutMix[19]는 학습하는 동안 지워진 영역을 전혀 다른 이미지의 일부분으로 채운다. 이는 일반적으로 학습 이미지의 수를 늘리기 위해서 사용하는 방법이지만 이와 같은 방법으로 이미지 내 객체의 주의 영역을 확장하는 효과를 얻을 수 있다. ACoL[20]은 CAM으로 학습이 끝난 후 지역화 맵에 핵심 주의 영역으로 간주되는 영역을 픽셀값 0인 검정색으로 지운다. 이후 이렇게 만든 학습 데이터들을 재훈련 시켜 이미지 내의 객체에서 기존에 주목받지 못했던 부분들도 주의하도록 학습하여 주의 영역을 확장시킨다.

3. 제안 방법

제안하는 방법은 Fig. 1과 같이 2단계로 구성된다. 첫 번째 단계에서는 입력된 이미지를 전체 10개 클래스로 분류하는 CAM 기반의 CNN 분류 모델로 훈련하고, 이 모델에서 분류 결과가 자주 혼동하는 클래스들과 비슷한 객체를 검출하는 클래스들을 묶어서 서브 클래스로 나누어 이들 또한 CAM 기반의 상세 분류를 위한 CNN 모델을 각각 훈련시킨다. 서브 클래스 모델은 C0/C9, C1/C2/C3/C4, C5/C7, C6/C8의 4개의 모델로 이루어진다. 이 때 전체 클래스 모델의 GAP를 통과한 10개의 출력과 서브 클래스 모델 각각의 GAP를 통과한 10개의 출력은 주어진 이미지에 대한 특징맵과의 매칭 정도를 표현하는 새로운 특징값으로 생각하여, 20개의 특징을 입력으로 하는 완전 연결 신경망을 추가하여 학습하고 최종 분류한다. 한 개의 은닉층을 갖는 완전 연결 신경망으로 훈련한 결과 두 모델을 단순히 계층적 또는 수평적으로 결합하는 것보다 분류 성능이 크게 향상되는 것을 확인할 수 있었다.

다음 단계에서는 최종적으로 분류된 결과에 대하여 전체 클래스 모델과 서브 클래스 모델에서 각각 해당되는 클래스의 CAM 특징맵을 결합하여 하나의 히트맵을 구성한다. 이는 주어진 이미지의 분류 결과에 대한 특징적 주의 영역에 해당된다. Fig. 1의 예에서 C9 Talk to passenger 클래스 결과에 대 운전자가 옆에 있는 탑승객 방향으로 얼굴을 돌리고 있

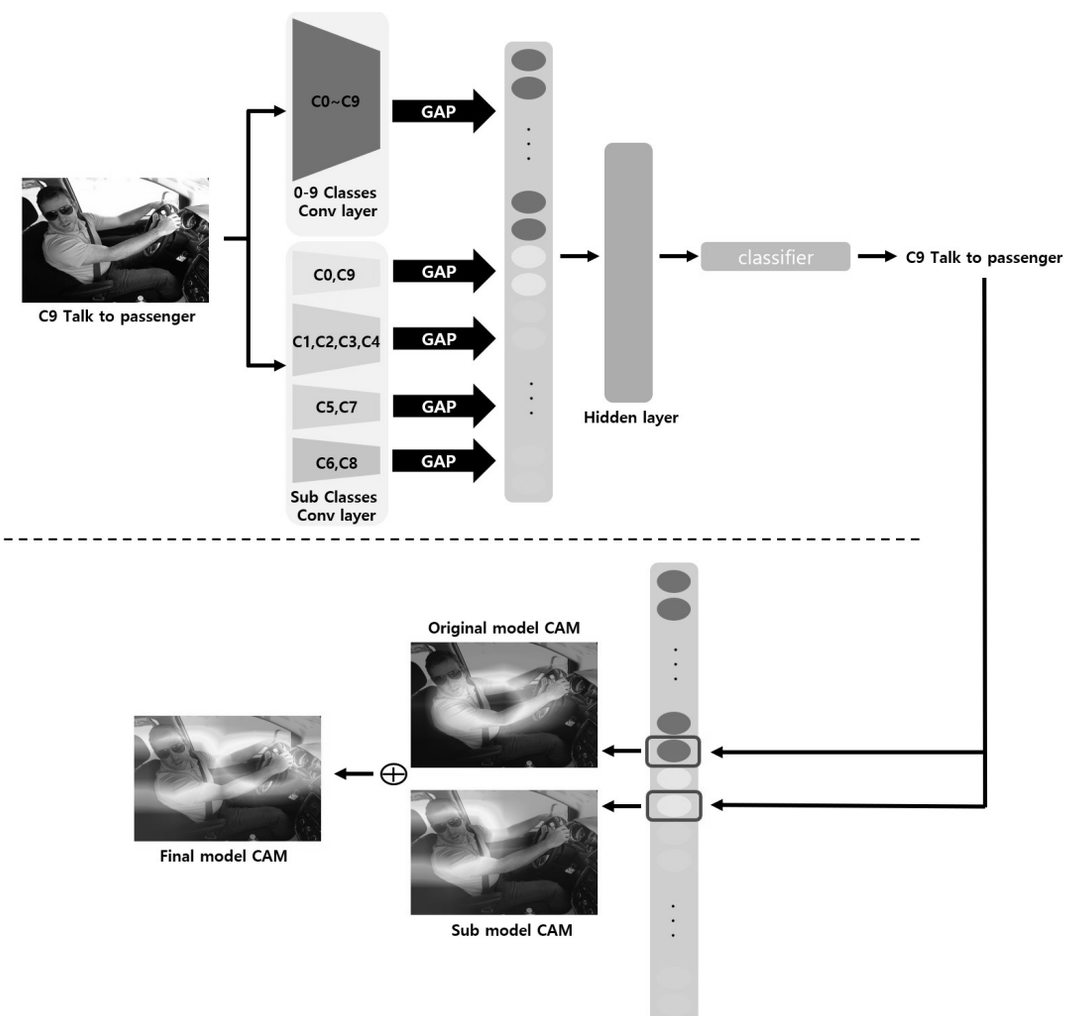


Fig. 1. Overview of the Proposed Method

는 모습과 손으로 핸들을 잡고 있는 모습이 특징적인 주의 영역으로 찾아진 것을 보여주고 있다. 두 개의 히트맵을 결합하는 방법을 포함한 제안하는 방법의 각 과정에 대한 구체적인 설명은 다음과 같다.

3.1 CAM을 이용한 전체 분류 및 상세 분류

각 클래스의 학습 이미지로 CNN 모델을 학습시키며, 여기서 사용한 CNN 모델은 특징추출의 마지막 단계인 합성곱(Convolution) 층(layer)과 연결된 완전 연결 층을 제거하고 대신 GAP 층으로 대체하여 CAM으로 구성한 모델이다[6]. CAM은 이미지 분류와 함께 객체의 위치 정보를 파악할 수 있는 장점이 있다. 서브 클래스 분류기는 클래스 상호 간에 혼동되는 클래스들을 선정하여 구성하며, 상세 분류를 통해서 분류 정확성을 높이고, 분류된 클래스의 내용에 맞도록 특징 영역을 정확하게 주의(Attention)하도록 돕는 역할을 한다. 10개의 전체 클래스 분류 및 상세 분류의 서브 클래스 분류기 모두를 CAM으로 구성한다.

Fig. 2는 검증 데이터에 대한 전체 클래스 모델의 혼동 행렬(Confusion Matrix)을 보여준다. 10개 클래스의 분류 결과에서 클래스간의 상호 오류가 많이 발생하는 C0, C9 클래스와 C6, C8 클래스를 각각 구별하기 위한 서브 클래스 분류 모델을 만든다. 또한 Fig. 3에서 확인할 수 있듯이 C0 Drive safe 클래스와 C9 Talk to passenger 클래스는 얼굴의 방향은 다르지만 상체의 움직임이 유사하며, 혼동 행렬의 결과에서도 이 두 클래스의 상호 오류가 많이 발생함을 보여준다. 또한 C6 Drink와 C8 Hair&Make up 클래스는 특징적인 오른쪽 팔이 비슷한 형태여서 여기서도 분류 결과에 대한 상호 오류가 빈번하게 발생하여 서브 클래스 모델을 만든다. 다음으로 C1 Text right, C2 Talk right, C3 Text left, C4 Talk left 네 개의 클래스는 모두 휴대폰을 사용하기 때문에 이에 대한 주의 영역을 정확히 추출하고자 서브 클래스 모델을 추가하고, C5 Adjust radio와 C7 Reach behind클래스는 왼쪽 팔과 얼굴 방향이 비슷하여 각각 서브 클래스 모델을 만든다. 따라서 C0/C9, C1/C2/C3/C4, C5/C7, C6/C8 클래스를 각각 상세하게 분류하는 4개의 서브 클래스 모델을 구성한다. 이들도 전체 클래스 모델과 같이 완전 연결 층 대신 GAP 층으로 대체하여 CAM을 구성한다.

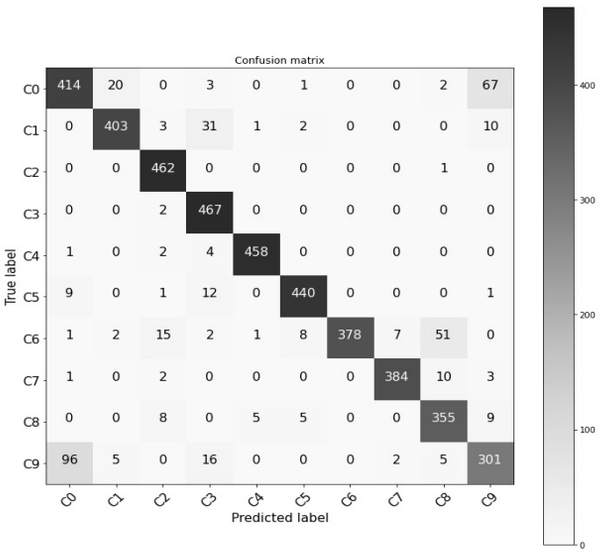


Fig. 2. Confusion Matrix in 10-class Base Model

3.2 특징맵 스코어를 이용한 분류

10개 클래스 분류기와 4개의 서브 클래스 분류기의 CAM 분류 결과는 주어진 이미지에 대해서 컨벌루션 층의 특징맵들과의 매칭 정도를 표현하는 것으로서 이를 임베딩된 20차원의 특징 벡터로 생각할 수 있다. 따라서 이 값들을 특징으로 하여 하나의 은닉층을 갖는 간단한 완전 연결 신경망을 구성하여 학습하고 이를 최종 분류 결과로 정한다.

Fig. 4는 각 모델에 훈련 데이터를 넣고 GAP 층을 거친 각 클래스의 결과 값을 보여준다. Fig 4(a)에서 C3 Text left 클래스의 이미지를 전체 모델과 서브 모델에 각각 넣었을 때 얻은 점수들을 확인할 수 있다. 전체 모델(Original Conv layer)에 넣어준 후의 값은 C3에서 36.01로 다른 점수보다 높으며, 서브 모델(Sub Conv layer)에 넣어준 후의 값에서

Fig. 3. Example Images According to the Sub-class Models

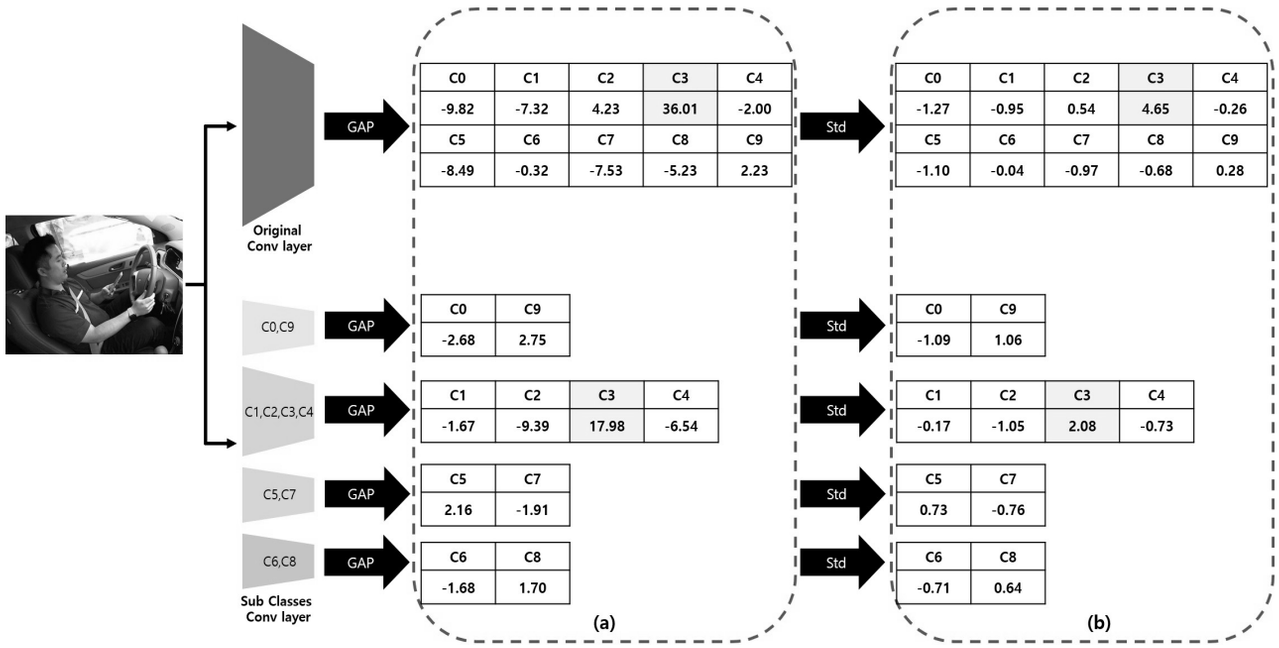


Fig. 4. Example Values After the GAP Layer When C3 Image Input to the Classification Models and Their Standardized Feature Values

는 C1, C2, C3, C4 서브 클래스 모델에서 C3가 17.98로 점수가 높은 것을 알 수 있다. 이와 같이 각각의 훈련 데이터를 넣어 얻은 20개의 분류 값을 Fig 4(b)와 같이 각 모델별로 표준화한 후 입력으로 사용하고, 그 데이터에 대한 타겟 클래스를 출력으로 하는 은닉층 1개로 구성된 간단한 완전 연결 신경망을 구성하여 학습한다. 그 결과 전체 모델과 서브 클래스 모델들을 산술적으로 결합한 결과보다 훨씬 높은 정확성의 분류 결과를 얻을 수 있었다.

3.3 분류 결과를 이용한 특징 영역 추출

최종 분류 클래스가 결정되면 전체 모델에서의 해당 클래스 히트맵과 서브 클래스 모델의 해당 클래스 히트맵을 결합하여 하나의 히트맵으로 만들고, 이로부터 특징적인 영역을 주시한다. 히트맵의 결합은 전체 모델과 서브 클래스 모델에서 나온 두 개의 GAP 층 이후의 값들을 각 픽셀마다 서로 비교하여 더 큰 값을 선택하여 하나의 최종 히트맵을 만든다.

Fig. 5는 두 개의 히트맵을 결합하는 방법을 보여준다. Fig. 5(a)의 첫 번째 히트맵은 0부터 145의 값 분포로서 가장 큰 값인 145와 그 근처의 값들이 빨간색으로 강하게 주시된다. 하지만 Fig. 5(b)의 히트맵은 14부터 255의 값 분포로서 Fig. 5(a)에서 가장 큰 값이었던 145는 비교적 약한 값으로 히트맵에 표현된다. Fig. 5(c)는 Fig. 5(a)와 Fig. 5(b)를 각 픽셀마다 비교하여 큰 값을 선택해서 새롭게 구성한 히트맵을 보여준다. 새롭게 구성된 히트맵은 하나의 이미지에 대해서 더욱 세밀하게 특징적인 영역이 주시 되기도 하며, 서로 다른 위치의 중요한 특징 영역들이 주시되어 전반적으로 주의 영역이 확장되어 찾아지기도 한다.

Fig. 6(a)는 C9 Talk to passenger 클래스 이미지를 넣었

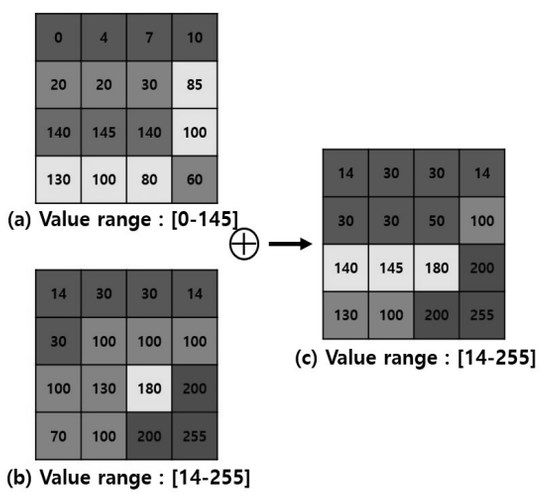


Fig. 5. Example for Combining Two Heatmaps

을 때의 히트맵으로서 Fig. 6(a)는 전체 클래스 모델에서의 히트맵, Fig. 6(b)는 서브 클래스 모델에서의 히트맵이며, Fig. 6(c)는 이 둘을 결합한 최종 히트맵을 보여준다. Fig. 6(a)에서는 C9클래스의 중요한 특징인 얼굴 영역을 주시하지 못했지만 서브 클래스 모델을 통해서 얼굴 영역을 주시할 수 있게 되었고, 결과적으로 Fig. 6(c)에서 얼굴과 팔의 동작이 모두 특징적인 주의 영역으로 찾아진 것을 확인할 수 있다. Fig. 6(d)는 C5 Adjust radio 클래스 이미지를 넣었을 때의 전체 클래스 모델에서의 히트맵이며, Fig. 6(e)는 서브 클래스 모델에서의 히트맵이고, Fig. 6(f)는 이 둘을 결합한 최종 히트맵 결과이다. 전체 클래스 모델로 훈련한 결과인 Fig. 6(d)에서는 C5의 특징인 뻗은 손을 주시하지 못했지만, Fig. 6(e)의

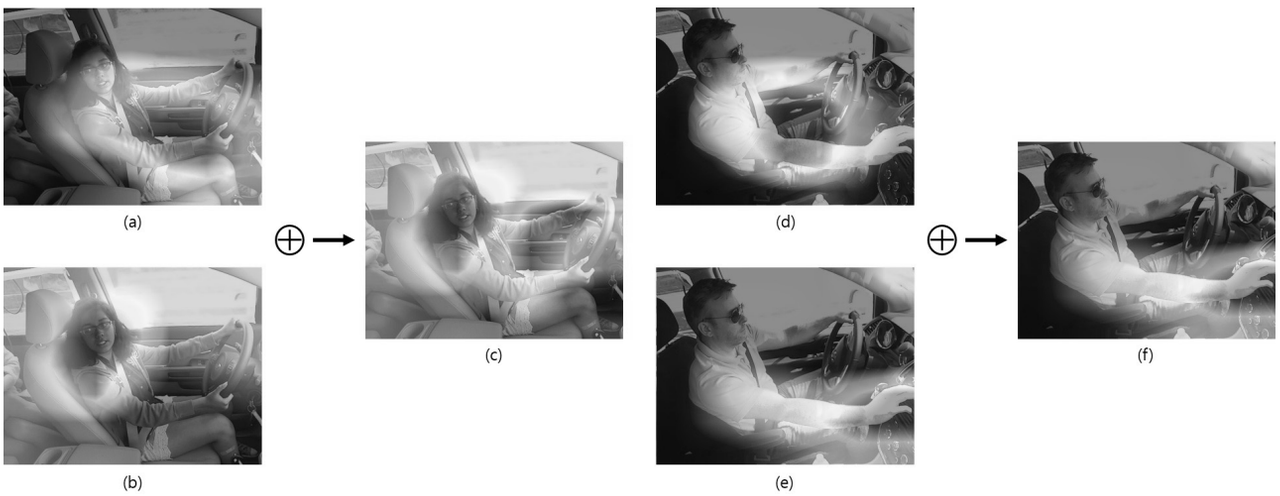


Fig. 6. Heatmaps for Base and Sub-class Models and Their Final Combined Heatmap

서브 클래스 모델에서는 뻗은 손을 특징 영역으로 주시하여 최종적으로 Fig. 6(f)와 같이 뻗은 손 영역을 찾게 되었다.

4. 실험 및 결과

4.1 실험 환경

실험 데이터는 2016년 Kaggle 경연대회에서 사용한 State Farm 데이터 셋이며[4], 이 셋은 총 22,424개의 학습 데이터와 클래스 별로 분류되지 않은 79,728개의 테스트 데이터로 구성되어 있다.

학습 데이터는 아시아, 미국, 유럽 등 다양한 국적의 남자 13명과 여자 13명의 이미지로 구성되어 있고, 전체 26명의 22,424개의 학습용 데이터로 구성되어 있어서 동일한 인물의 유사한 데이터들이 많이 존재한다. Table 1과 같이 전체 10개의 클래스로 구분되어 있으며, 본 논문에서는 정확성 평가에 대한 신뢰도를 높이기 위해서 학습 데이터 참가자들과 다른 참가자들로 구성된 2개의 테스트 데이터 셋을 만들었으며 각각 클래스별로 약 100개의 이미지로 구성하였다. Table 1은 실험에서 사용한 학습 데이터, 학습 데이터의 일부로 구성한 검증(Validation) 데이터 및 2개의 별도 테스트 데이터 셋의 클래스별 데이터 개수를 보여준다.

CAM을 구축하기 위한 CNN 모델은 ILSVRC[21]로 사전 학습한 Resnet50 모델이며, Cross Entropy 손실 함수와 Adam 알고리즘으로 CNN 모델을 최적화시켰다. 전체 Epoch은 10개 클래스 모델에서는 10회, 각 서브 클래스 모델에서는 7회로 설정하고, 초기 학습률은 모두 동일하게 0.001, Epoch은 5회마다 학습률이 1/10이 줄어들도록 설정하였다.

CNN 모델에서 GAP 층 이후 값들을 훈련하기 위한 완전 연결 신경망은 실험을 통해 18개의 은닉 노드를 갖는 한 개의 은닉층으로 구성하였다. 시그모이드 활성화 함수를 사용했으며, 20차원의 입력 특징값들은 모두 표준화(Standardization)하여 입력하였다.

하드웨어는 NVIDIA의 GeForce RTX 2080 Ti GPU를 사용했다. 총 훈련 시간은 약 2시간 1분 16초가 소요되었으며, 하나의 이미지 데이터를 모델에 넣어 분류 결과를 얻고 이를 바탕으로 특징적주의 영역을 찾는데 걸리는 시간은 평균 2.93초이다. 각 인식 모델별로 GPU를 병렬로 담당시켜 처리한다면 처리시간이 훨씬 단축될 것이다.

Table 1. Train, Validation and Test Data Sets

	Train	Validation	Test	
			Test set 1	Test set 2
C0 Drive safe	1,992	497	121	114
C1 Text right	1,814	453	130	100
C2 Talk right	1,854	463	147	104
C3 Text left	1,877	469	127	103
C4 Talk left	1,861	465	112	104
C5 Adjust radio	1,850	462	140	107
C6 Drink	1,860	465	118	101
C7 Reach behind	1,602	400	119	106
C8 Hair&Make up	1,529	382	102	100
C9 Talk to passenger	1,704	425	120	106
Total	17,943	4,481	1,236	1,045

Table 2. Compared Model Accuracies

	Original Model Accuracy		Sub Model Accuracy		Ours	
	Test set 1	Test set 2	Test set 1	Test set 2	Test set 1	Test set 2
C0 Drive safe	90.1%	88.6%	93.3%	94.7%	96.7%	92.2%
C9 Talk to passenger	60.0%	63.2%	86.6%	89.6%	76.6%	85.8%
C1 Text right	84.6%	81.0%	100.0%	100.0%	96.9%	98.0%
C2 Talk right	98.0%	98.0%	98.6%	99.0%	98.0%	100.0%
C3 Text left	100.0%	100.0%	99.2%	97.1%	100.0%	100.0%
C4 Talk left	96.4%	92.3%	94.6%	93.2%	98.2%	92.3%
C5 Adjust radio	90.0%	91.5%	100.0%	99.1%	97.1%	97.2%
C7 Reach behind	90.8%	96.2%	100.0%	100.0%	98.3%	97.2%
C6 Drink	68.6%	73.2%	92.3%	93.1%	90.6%	83.2%
C8 Hair&Make up	75.5%	89.0%	91.1%	96.0%	94.1%	97.0%
Total	86.00%	87.36%	95.87%	96.17%	95.15%	95.12%

4.2 분류 정확성 평가

Table 2는 CAM 기반의 10개 클래스 CNN 모델과 제안한 서브 클래스 모델, 이 둘을 학습하여 결합한 방법의 최종 정확도를 클래스별로 비교한 결과이다. 서브 클래스 정확도는 C0, C9의 두 클래스 상세 분류 정확성, C1, C2, C3, C4의 네 클래스 분류 정확성, C5, C7의 분류 정확성 및 C6, C8의 두 클래스 정확성을 측정한 것으로서 분류하는 클래스 수가 줄어들어 정확성이 향상되는 것은 예상할 수 있는 결과이다. 최종 모델의 클래스 정확도는 테스트 셋 1에서는 86.00%에서 95.15%로 9.15%의 큰 향상을 보였고, 테스트 셋 2에서도 87.36%에서 95.12%로 7.76%의 큰 향상을 보였으며 모든 클래스에서 정확성이 향상되었다. 서브 클래스 모델의 정확도에 비해 최종 모델의 정확도가 일부 클래스에서 낮아졌지만, 이는 서브 클래스 모델이 적은 수의 클래스만을 분류하는 것으로서 개별적인 정확도가 높을 수 있기 때문이다.

Table 3은 CNN 모델에서 GAP 레이어 이후 값들을 훈련하기 위해서 만든 모델의 은닉층의 은닉노드 수와 활성화 함수를 변화시키며 측정한 정확도 결과이다. CAM으로 표현된 20차원의 입력은 잘 정제된 특징과 같아서 한 개의 은닉층만으로 구성한 완전 연결 신경망 모델로도 정확성을 높일 수 있었다. Table 4는 State Farm 데이터 셋을 사용한 방법들의 정확도를 비교한 것으로서 정확도가 우수한 DD-RCNN [10]의 92.2%보다 제안한 방법의 테스트 셋 1의 정확도는 2.95%, 테스트 셋 2의 정확도는 2.92%가 높아진 것을 확인할 수 있다.

4.3 주의 영역 검출 결과

주의 영역 검출이 어느 정도 개선되었는지를 정량적으로 평가하기 위해서 각 클래스 별로 이미지의 어느 부분이 집중되어야 하는지를 설정하고 이를 기반으로 평가한다. Fig. 7의 각 클래스에 설정된 영역은 CNN 및 CAM 기반으로 10개의 클래스를 학습하고 분류할 때 분류기의 클래스 선정 영역을 기반으로 한 클래스 상호간에 구별되는 특징적인 영역이다.

Table 3. Accuracies According to the Number of Hidden Units and Activation Functions

Hidden layers(Nodes)	Activation Function	Accuracy
2(15, 10)	ReLU	93.61%
	Sigmoid	93.53%
1(25)	ReLU	93.28%
1(20)	ReLU	94.90%
	Sigmoid	94.98%
1(18)(Ours)	Sigmoid	95.15%

Table 4. Compared with other methods using the same State Farm Dataset

Method		Accuracy
Optical flow(Single Frame)[11]		82.38%
3D Convolutional Neural Networks(Single Frame)[12]		85.00%
DD-RCNN[10]		92.20%
Ours	Test set 1	95.15%
	Test set 2	95.12%

C0 Drive safe 클래스는 운전대를 잡고 있는 양손 영역, C1 Text right, C2 Talk right, C3 Text left와 C4 Talk left 클래스들은 핸드폰을 잡고 있는 운전자의 손 영역이 특징 영역이다. C5 Adjust radio 클래스는 라디오를 조작하고 있는 오른손 영역, C6 Drink 클래스는 음료수를 쥐고 있는 손 영역, C7 Reach behind 클래스는뒤를 돌아보고 있는 운전자의 머리 영역, C9 Talk to passenger 클래스는 조수석으로 향한 운전자의 얼굴 영역과 운전대를 잡고 있는 두 영역을 특징 영역으로 설정한다. C8 Hair&Make up 클래스는 운전자 앞에 보이는 거울 영역이 설정되는데, 이는 C2 Talk right, C6 Drink 클래스들의 오른손 영역과 구분하기 위해서 CNN이 학습하면서 이들과 구분된 거울 영역을 이 클래스의 특징 영역으로 삼은 것이다.

Table 5는 테스트 셋 1에 대해서 기존 CAM 결과와 이 논

Table 5. Accuracy Comparison of Characteristic Atention Area

	Original CAM Accuracy	Ours
C0 Drive safe	84.3%	90.1%
C1 Text right	83.8%	97.7%
C2 Talk right	94.6%	96.6%
C3 Text left	99.2%	100.0%
C4 talk left	76.8%	93.8%
C5 Adjust radio	74.3%	95.7%
C6 Drink	61.9%	69.5%
C7 Reach behind	82.4%	96.6%
C8 Hair&Make up	48.0%	97.1%
C9 Talk to passenger	13.3%	61.7%
Total	73.0%	89.9%

Fig. 7. Areas to Focus on for Class Distinction

문에서 제안한 CAM 결과를 비교한 결과이다. 이는 Fig. 7의 각 클래스별로 집중해야 하는 영역이 대부분 찾아질 때 맞는 영역으로 하여 정확도를 평가한 표이다. 모든 클래스에서 정확도가 향상되었으며, 특히 C9 Talk to passenger 클래스는 기존 CAM에서는 C0 Drive safe 클래스와 마찬가지로 운전대를 잡고 있는 두 팔에 집중했던 것이 제안한 방법에서는 조수석을 보고 있는 운전자의 얼굴 영역도 함께 주의되면서 정확도 및 의미적으로 큰 향상이 있었다. 여기서 기존의 방법은 두 영역 중 한 부분만 집중해도 찾은 것으로 하였는데 이는 두 영역 모두를 찾는 경우가 거의 없기 때문이다. 반면 제안한 방법은 두 영역 모두를 집중할 때 찾은 것으로 정확성을 평가하였다. 또한 C8 Hair&Make up 클래스는 기존의 CAM에서는 C2, C9 클래스들과 구분이 어려워서 정확성이 낮았지만 제안한 방법으로 구분된 영역이 주의 되면서 정확도가 크게 향상된 것을 확인할 수 있다.

Fig. 8은 기존의 CAM 기본 모델로 분류한 결과의 특징맵에 대한 히트맵과 제안한 방법으로 분류한 결과의 특징맵에 대한 히트맵을 10개의 모든 클래스에서 비교한 예를 보여준다. 기존 CAM은 클래스를 구별할 수 있는 신체의 일부분만을 집중하거나, 사람이 느끼기에 별로 중요하지 않은 영역에도 집중했지만, 제안한 방법에서의 CAM은 클래스를 구별하고 의미있는 중요한 부분에 주의하는 경향으로 개선된 것을 확인할 수 있다.

5. 결 론

이 연구에서는 운전자 부주의 유형을 정확히 분류하고, 분류된 운전자의 모습에서 특징적인 이미지 영역을 주시하는 방법을 제안하였다. 제안한 방법인 10개 클래스 분류 모델과 상세 분류의 서브클래스 분류 모델의 결과를 학습하여 분류 정확성을 크게 높였으며, 각 분류 모델의 주의 영역을 결합하여 운전자의 특징 영역을 더욱 정확하고 의미있게 찾아낼 수 있었다. 10개 클래스 분류 모델만을 이용했을 때의 86.00% 정확도에서 제안한 방법으로 95.14%의 정확도로 9.14%의 큰 성능 향상을 만들었다. 이는 State Farm 데이터 셋의 학습 및 테스트 데이터를 구분해서 이용한 다른 연구 결과들과 비교하여 우수한 결과이다.

향후 연구에서는 운전자의 부주의 검출을 바탕으로 운전자 위험도를 평가하는 체계를 완성하고, 운전자의 표정과 시선을 함께 검출하여 위급한 상황을 판단하고자 한다. 또한 운전

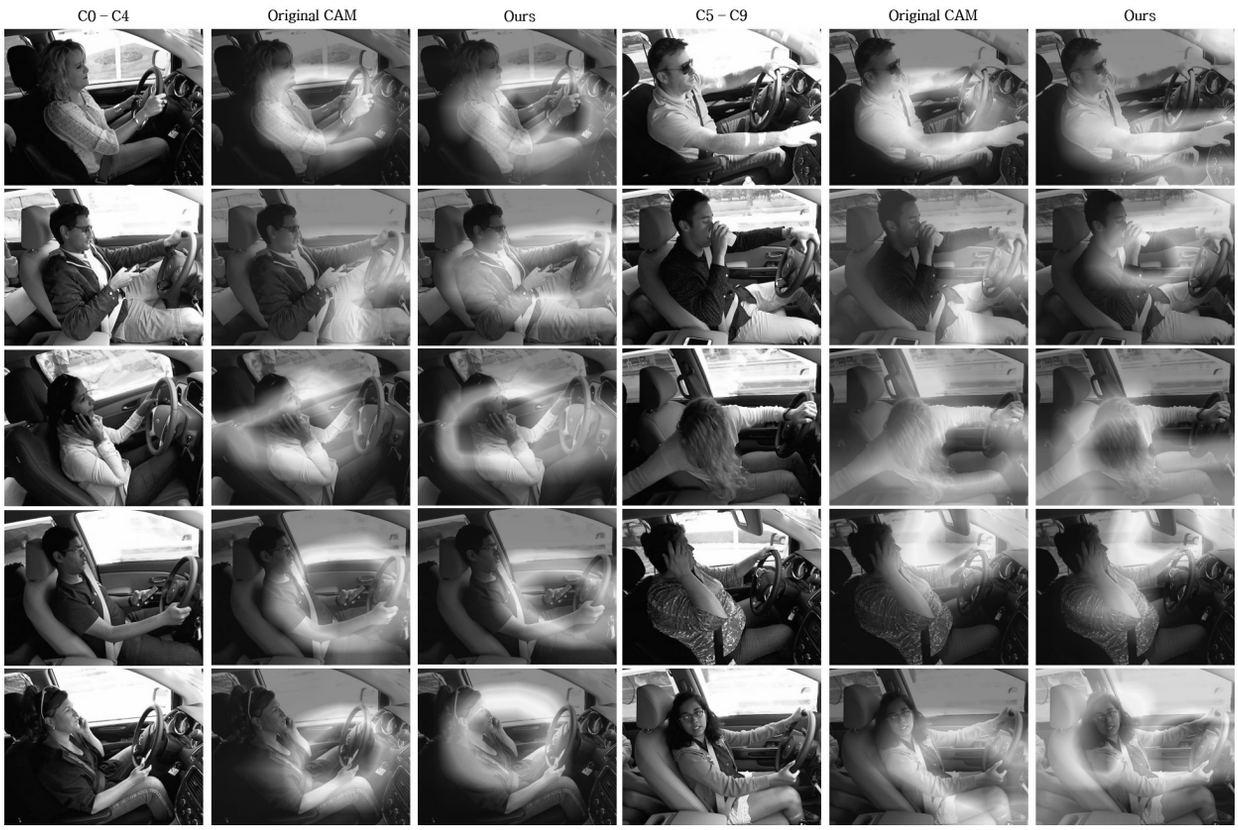


Fig. 8. Attention Areas Compared Original CAM with Our Approach

자 부주의 데이터 셋에서 운전자의 행동 검출에 방해가 되는 그림자를 제거하는 방법을 추가하여 더욱 정확한 운전자 부주의 및 행동을 탐지하고자 한다.

References

[1] World Health Organization, *Global Status Report on Road Safety 2018: Summary*, World Health Organization, Geneva, Switzerland, 2018.

[2] National Highway Traffic Safety Administration, "2015 motor vehicle crashes: Overview," *Traffic safety facts: research note*, U.S. Department of Transportation, August, 2016.

[3] C. H. Zhao, B. L. Zhang, J. He, and J. Lian, "Recognition of driving postures by contourlet transform and random forests," *Intelligent Transport Systems*, Vol.6, pp.161-168, 2012.

[4] I. Sultan. Academic purposes, [Internet], https://www.kaggle.com/c/ state-farm-distracted-driver-detection/discussion/20043#117982. 2016.

[5] H. M. Eraqi, Y. Abouelnaga, M. H. Saad, and M. N. Moustafa, "Driver distraction identification with an ensemble of convolutional neural networks," *Journal of Advanced Transportation*, Vol.2019, 2019.

[6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp.2921-2929, 2016.

[7] M. Alotaibi and B. Alotaibi, "Distracted driver classification using deep learning," *Signal Image Video Process*, pp.617-624, 2019.

[8] M. Q. Lu, Y. C. Hu, and X. B. Lu, "Driveraction recognition using deformable and dilated fasterR-CNN with optimized region proposals," *Applied Intelligence*, Vol.50, pp.1100-1111, 2020.

[9] L. C. Valeriano, P. Napoletano, and R. Schettini, "Recognition of driver distractions using deep learning," In *Proceedings of the 2018 IEEE 8th International Conferenceon Consumer Electronics*, pp.1-6, 2018.

[10] N. Moslemi, R. Azmi, and M. Soryani, "Driver distraction recognition using 3D convolutional neural networks," In *Proceedings of the 2019 4th International Conferenceon Pattern Recognitionand Image Analysis*, pp.145-151, 2019.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, Vol.1, pp.770-778, 2016.

[12] J. Chung, S. Ahn, and Y. Bengio, "Hierarchical multiscale recurrent neural networks," *IInternational Conference on Learning Representations*, 2017.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, and A. Rabinovich, "Going deeper with convolutions," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1-9, 2015.

[14] S. Masood, A. Rai, A. Aggarwal, and M. N. Doja, "Detecting distraction of drivers using Convolutional Neural Network," *Pattern Recognition Letters*, Vol.139, 2018.

[15] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, 2015.

[16] 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, pp.1137-1149, 2015.

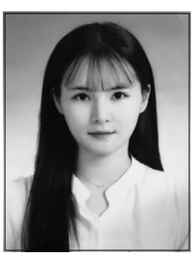
[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *IEEE International Conference on Computer Vision*, pp.618-626, 2017.

[18] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly- supervised object and action localization," *IEEE International Conference on Computer Vision*, pp.3544-3553, 2017.

[19] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," *IEEE/CVF International Conference on Computer Vision*, pp.6022-6031, 2019.

[20] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang, "Adversarial complementary learning for weakly supervised object localization," in *IEEE Computer Vision and Pattern Recognition*, pp.1325-1334, 2018.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *Conference on Computer Vision and Pattern Recognition*, pp. 248-255, 2009.



고 수 연

https://orcid.org/0000-0002-2373-8695
e-mail : sy1122@sookmyung.ac.kr
2020년 숙명여자대학교 컴퓨터과학과(학사)
2020년 ~ 현 재 숙명여자대학교 컴퓨터과학과 석사과정
관심분야 : 머신러닝, WSOL

최 영 우

https://orcid.org/0000-0003-0364-236X
e-mail : ywchoi@sookmyung.ac.kr
1985년 연세대학교 전자공학과(학사)
1986년 Univ. of Southern California 컴퓨터공학과(석사)
1994년 Univ. of Southern California 컴퓨터공학과(박사)
1994년 ~ 1997년 LG전자기술원 선임연구원
1997년 ~ 현 재 숙명여자대학교 컴퓨터과학과 교수
관심분야 : 시각정보처리, 머신러닝