Journal of Korea Multimedia Society Vol. 20, No. 8, August 2017(pp. 1271-1281)
https://doi.org/10.9717/kmms.2017.20.8.1271

다중 센서를 사용한 주행 환경에서의 객체 검출 및 분류 방법

김정언†, 강행봉††

A New Object Region Detection and Classification Method using Multiple Sensors on the Driving Environment

Jung-Un Kim†, Hang-Bong Kang††

ABSTRACT

It is essential to collect and analyze target information around the vehicle for autonomous driving of the vehicle. Based on the analysis, environmental information such as location and direction should be analyzed in real time to control the vehicle. In particular, obstruction or cutting of objects in the image must be handled to provide accurate information about the vehicle environment and to facilitate safe operation. In this paper, we propose a method to simultaneously generate 2D and 3D bounding box proposals using LiDAR Edge generated by filtering LiDAR sensor information. We classify the classes of each proposal by connecting them with Region-based Fully-Covolutional Networks (R-FCN), which is an object classifier based on Deep Learning, which uses two-dimensional images as inputs. Each 3D box is rearranged by using the class label and the subcategory information of each class to finally complete the 3D bounding box corresponding to the object. Because 3D bounding boxes are created in 3D space, object information such as space coordinates and object size can be obtained at once, and 2D bounding boxes associated with 3D boxes do not have problems such as occlusion.

Key words: Deep Learning. Sensor Fusion. Proposal. Object Classification. Autonomous Driving

1. 서 론

최근 객체 인식 및 분류에 대한 연구는 딥러닝(Deep learning), 특히 CNN(Convolutional neural networks)[1]을 통해 비약적으로 발전하였다. 특히, 자동차 시장에서 자율 주행 기술이 주목을 받고 있고, 많은 상용차 업체들이 2020년 실차 출시를 목표로 관련 기술을 개발하고 있기 때문에 주행 환경에서의 물체 인식과 객체의 행동 분석을 위한 심층적인 네트워크 연구가 활발히 진행되고 있다. 이 분야의 가장 최신의 연구들은 크게 두 가지 방향으로 발전하고 있다. 이는 객체의 영역 검출과 객체 분류를 한번에 처리하는 단일 처리 방법(single stage method)과 객체의 영역을 선검출하고, 영역 내의 객체를 분류하는 두 번의 과정을 거치는 분할 처리 방법(two stage method)이다. 단일 처리 방법의 경우 객체의 영역 상자(Bounding box) 검출과 영역 내의 객체 특징(object feature)에 대한 클래스별 확률 계산이 함께 수행된다. YOLO[6]의 경우 격자무늬(Grid)를 기반으로 영역 상자를 생성하며, 상자 회귀(box regression) 과정에서 클래스 레이블(class label)을 결정하기 때문에 속도면에서 매우 우수한 성능을 보인

※ Corresponding Author : Hang-Bong Kang, Address: (14662) 43, Jibong-ro, Bucheon-si, Gyeonggi-do, Republic of Korea, TEL : +82-2-2164-4598, FAX : +82-2-2164-4945, E-mail : hbkang@catholic.ac.kr
Receipt date : Jul. 17, 2017, Approval date : Jul. 31, 2017

† Dept. of Media Eng., The Catholic University of Korea (E-mail : amysh@catholic.ac.kr)
†† Dept. of Media Eng., The Catholic University of Korea
※ This research was supported by a grant from Agency for Defense Development, under contract #UD150016ID.

다. 이와 유사한 단일 처리 방법 중 하나인 Single shot detector(SSD)[7]는 영역 상자(Bounding box)와 클래스별 신뢰 지도(class confidence map)을 각 단계의 콘벌루션 특징 지도(convolution feature map)을 사용하여 평가하였다. 이 접근 방법은 하이퍼넷(HyperNet)[4], SDP[23] 등에서도 제안된 형태로서, 이는 관심영역(ROI)의 크기에 따라 각기 다른 단계의 콘벌루션 특징을 사용하여 분류하게 된다. 이러한 접근 방법은 콘벌루션(convolution)과 풀링(pooling)을 반복하는 CNN의 특성상 하위 레이어(Layer)의 객체 특징의 크기(feature size)가 점점 작아져 작은 객체를 판단하기에 정보가 충분하지 않게 되는 단점을 보완하게 된다.

두 방법 모두 가려지거나 잘린 객체의 보이는 영역의 특성을 기반으로 영역 상자가 만들어지고, 객체는 각 상자에 포함된 객체의 일부분에 대한 정보를 통해 분류된다. 이는 객체의 클래스는 식별할 수 있지만, 가려진 영역의 실제 객체 크기를 예측하기 어렵다. 또한 객체의 3차원 정보를 파악할 수 없기 때문에 주변 객체의 상대적 움직임을 예측하고 대응하는데 어려움이 생긴다. 따라서 안전한 자율 주행을 위해서는 물체의 실제 크기나 3차원 위치, 방향과 같은 3차원 공간 정보를 획득할 필요가 있다.

이러한 3차원 공간 정보를 처리하기 위해 김대년 등[24]은 소실점과 다중 특징을 이용하여 외부 환경에서의 물체를 분석하고 사물의 높이와 크기를 추정하는 연구를 수행하였다. 또한 3DOP [8]는 스테레오 영상(Stereo image)을 이용하여 각 영역별 깊이를 추정하고 2차원 영상의 각 픽셀(pixel)을 추정된 깊이를 통해 3차원 공간에 투영하여 객체의 3차원 포즈(3D pose)를 추정하였다. 각 픽셀 간의 좌표, 색상, 변화도(gradient) 등의 여러 속성정보를 사용하여 픽셀 주변의 관계를 MRF(Markov random field)[9] 에너지 함수로 정의하고 SVM (Support Vector Machine)[10]을 사용하여 객체를 분류하게 된다. 하지만 3DOP는 3차원 공간 투영을 위해 사용하는 스테레오 깊이 영상에 대한 의존도가 높아 깊이 영상이 매우 정교해야 좋은 결과를 얻을 수 있고, 정교한 스테레오 영상을 획득하기 위해서는 많은 시간 및 자원이 소요되는 단점이 있다. 반면 라이다(LiDAR, light detection and ranging) 등 레이저 센서를 사용하면 별도의 과정 없이 차량 주변의 3차원 공간정보를 실시간으로 획득할 수 있기 때문에 스테레오 영상을 효과적으로 대체할 수 있을 것으로 기대되어 이를 활용하기 위한 연구가 수행되었다. 예를 들어 vote3D[11]는 3차원 복셀 격자 영상(voxel grid)을 통해 3차원 공간에서의 라이다 구조를 직접 학습하는 방법을 사용했다. 이는 객체에 해당하는 3차원 복셀을 2차원의 특징 영상처럼 사용하였지만, 라이다 센서의 특성상 동일 객체에 대한 복셀 구조의 변화폭이 크기 때문에 좋은 성능을 기대하기는 어렵다. 이런 문제 때문에 라이다 데이터를 직접적으로 사용하는 연구보다는 CAD 모델 등을 이용하여 가상의 3차원 정보를 생성하는 연구가 수행되었다. 3DVP [12]의 경우 3차원 복셀로 변환된 CAD 모델을 수작업으로 2D 영상위에 정렬하는 방법으로 3차원 복셀을 생성하고 이를 학습하는 방법을 사용하였다. 이를 응용한 SubCNN(sub-category aware CNN)[13]은 CNN에 의한 1차적인 객체 분류 후 해당 클래스 카테고리(class category)에 대한 추가적인 분류기로 3DVP를 연결하여 차량의 가려짐, 잘림, 차량의 방향 등 세부 정보를 분류하였다. 하지만 이러한 방법은 분류하고자 하는 클래스에 대한 세부 분류기가 별도로 연결되어야하며 각 네트워크에 속한 자원을 공유하는 것도 어렵기 때문에 많은 시간과 자원을 요구한다.

따라서 본 논문에서는 라이다 센서 데이터의 약점을 단안 카메라 도메인에서의 필터링(filtering)을 통해 보완하고, 실시간으로 획득되는 3차원 공간 정보를 통해 객체의 3차원 포즈를 복원하는 방법을 제안한다.

제안된 방법은 3차원 복셀 대신 일정 높이로 z축을 통합한 투영 평면(projection plane)을 생성하고 이를 카메라 공간에서 정리하는 간단한 방법을 통해 각 객체의 외곽 형태(shape) 정보를 추출하고 각 객체의 영역을 분할한다. 분할된 영역과 깊이 정보, 그리고 분류하고자 하는 객체의 평균 크기를 이용해 2차원 객체 프로포절(proposal)을 생성하고 이를 관심 영역(ROI, region of interest)으로 사용하는 객체 분류 방법을 제안한다. 이는 라이다 기반의 프로포절로 얻어진 2차원에서의 객체 분류 결과를 3차원 공간상의 라이다 점 집단, 라이다 에지(LiDAR point group, 라이다 에지)으로 전파하여 2차원과 3차원의 객체 영역 상자를 동시에 생성하고 각 클래스의 평균 크기와 종횡비(aspect ratio)를 바탕으로 영역 상자를 확장하기 때문에 가려짐 등의 문제와 3차원 포즈 복원

의 문제를 한번에 해결할 수 있다. 우리는 이 방법을 통해 객체 검출에 있어 기존 최신 연구에 비해 분류의 정확도를 높이고, 보다 실측 데이터에 가까운 영역 상자를 생성하였다.

요약하면 본 연구의 기여도(contribution)는 다음과 같다.

1. 객체 분류에 사용하기 어려운 라이다 포인트 클라우드(LiDAR Point Cloud)의 단점을 라이다 에지(라이다 에지)를 통해 보완

2. 라이다를 통한 의미있는 2차원 프로포절 생성과 객체의 3차원 좌표 획득

3. 실측 자료(ground truth)와의 높은 중첩율(IOU rate) 확보

2. 제안 방법

본 연구는 이 장에서 Fig. 1과 같은 과정을 통해 불균일하고 희소하게 분포된 라이다 포인트 클라우드를 정리하여 객체의 외곽 영역을 표현하는 라이다 에지를 만들고, 이를 통해 카메라 공간 기반 분류기를 위한 프로포절을 생성하는 방법을 보인다. 또한 이 과정에서 프로포절 생성에 사용된 라이다 구성 점들의 3차원 좌표와 방향성을 통해 2차원에서 분류된 클래스 레이블에 적합한 평균 크기를 통해 3차원 영역 상자를 동시에 생성한다. 생성된 3차원 영역 상자는 3차원 공간 정보를 가지고 있을 뿐 아니라 카메라 공간에서 가려짐 등에 의해 보이지 않는 영역으로 확장되므로 일반적인 2차원 프로포절과 상자 영역 회귀를 통해 얻기 힘든 실측 자료에 가까운 영역 상자를 획득할 수 있다. 또한 라이다 에지를 통해 분할된 영역에 따라 초기 프로포절을 생성하기 때문에 적은 수의 의미 있는 프로포절을 생성하여 분류 시간을 줄이게 된다.

2.1 2차원 및 3차원 객체 외곽선(object shape edges) 생성

이 장에서 촬영 차량을 중심으로 방사형으로 퍼져 있는 분산된 라이다 포인트 클라우드 집합 P를 이용하여 2차원 객체를 분류하기 위한 프로포절을 생성하는 프로세스를 설명한다. 3차원 공간에 퍼져 있는 라이다 포인트 집합 P의 각 원소 p는 센서의 레이어 수, 회전속도, 주사율, 각 객체와의 거리 등에 따라 각기 일정하지 않은 낮은 밀도의 분포를 지니고 있다. 이러한 특성 때문에 동일한 객체에 대해서도 촬영 시점에 따라 일정하지 않은 점 분포를 보이며 이는 객체의 학습과 분류에 있어 정확도를 떨어뜨리는 원인이 된다. 따라서 구성 점들의 밀도를 증가시키고 신뢰할 수 있는 형태의 특징으로 변환할 필요가 있다. 이를 위해 3차원 격자형 복셀의 높이를 기반으로 지면(Ground)에 속한 점들을 제거하고, 남은 점 집합 P의 높이를 통일시켜 단위 면적당 점들의 밀도를 증가시킨다. 다음으로 같은 높이에 모인 점들 중 객체를 표현하는 외곽선을 찾기 위해 최 외곽 성분을 제외한 잡음 성분(noise points)들을 카메라 영역에서 제거하는 방법을 소개한다. 이 과정을 통해 불균일하게 넓게 퍼져있던 라이다 점들을 정리하여 균일한 직선성분의 객체 외곽선을 생성하게 된다.

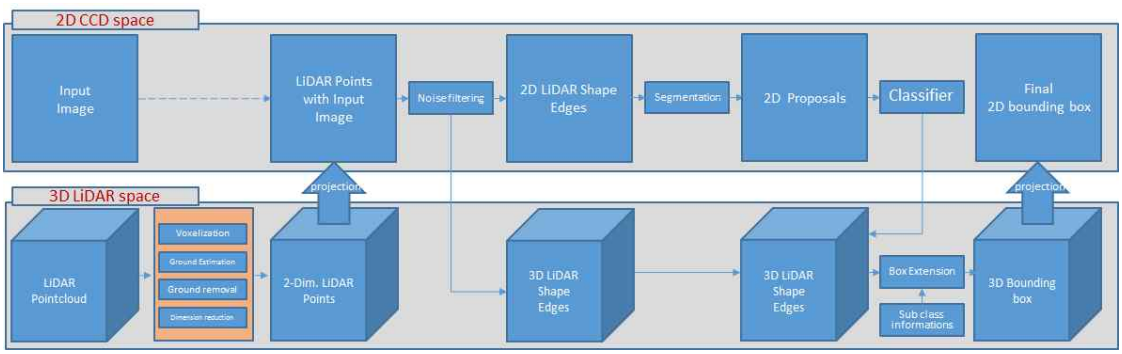

Fig. 1. System overview. The upper row shows the process in the two-dimensional CCD space, and the lower row shows the process in the three-dimensional LiDAR space. The LiDAR Point cloud is filtered through the two-dimensional and three-dimensional space to create an 'object proposal' that represents the area of the object.

2.1.1 지면 검출 및 제거

주행 환경에서의 객체 검출 및 분류를 효과적으로 수행하기 위해 먼저 3차원 공간의 라이다 점들 가운데 지면을 나타내는 점들을 검출하고 제거하여 객체의 가능성이 높은 점들을 분류해야 한다. 라이다 센서는 특정 높이에 설치되어 64개 레이어의 광선이 방사형으로 회전하면서 사출되어 사물 표면까지의 거리 정보를 수집하게 된다. 때문에 3차원 격자로 영역을 분할하였을 때 지면을 포함하는 격자 셀(grid cell)의 경우 사물과의 경계선을 제외하면 z축에 대해 하나의 셀만 존재하게 된다. 따라서 본 연구에서는 라이다 포인트가 분포된 전체 3차원 공간을 20cm x 20cm x 10cm의 복셀 격자로 나누고 각 격자 셀이 포함하는 점들이 임계값(threshold, 본 연구에서는 3)을 초과하는 경우 사물이 존재한다고 판단하고 분류하여 사물 존재 여부에 대한 복셀 격자 공간 V_{xy}를 생성한다.

다음으로 지면에 속한 점들과 객체에 속한 점들을 구분하기 위해 복셀 V의 높이가 1인 셀의 모든 p를 그룹 O_p로 묶는다. 이 중 객체일 가능성이 높은 그룹 O_p의 소속 점 p의 밀도를 높이기 위해 모든 p의 z축 크기를 통일하여 XY 평면에 평행한 집합 P_{plane}으로 변환한다. 이 과정에서 라이다 점들의 3차원 형태 정보에 손실이 발생하지만, 점들의 밀도가 높아지고 인접 점과의 연속성을 판단할 수 있게 된다. 본 연구에서는 P_{plane}의 소속 점들을 이용하여 객체 영역의 분할(segmentation)을 수행한다. 하지만 P_{plane}의 소속 점들은 KITTI 데이터셋(dataset)을 기준으로 1.4m 높이에서 방사형으로 촬영된 자료이기 때문에 객체의 형태에 따라 불균일한 군집 형태로 존재하게 된다. 이는 나. 의 과정을 통해 필터링하여 각 객체를 표현하는 라이다 외곽선(LiDAR shape edge, LiDAR Edge) 형태로 정리한다.

2.1.2 카메라 공간에서의 객체 외곽선 생성

위 과정에서 생성된 P_{plane}의 소속 점 p 가운데 객체의 외곽 형태를 표현하는 점들을 제외한 내부 점들을 제거하여 객체 외곽의 형태를 나타내는 선의 형태로 표현하고, 이웃하는 선과의 연속성을 기준으로 영역을 나누어 객체를 분할한다. 라이다 센서의 수집 정보들은 차량에 설치된 센서로부터 방사형으로 발사되는 레이저를 통해 수집되기 때문에 P_{plane}내의 각 객체의 외곽선은 회전방향으로 촬영 차량과 가장 가까운 점에 대한 집합이 된다. 하지만 실수 좌표계인 3차원 공간에서 각도에 따른 최소거리 점을 구하기 어렵기 때문에 정수 좌표계인 카메라 공간으로 투영하여 외곽선을 구한다. 이를 위해 먼저 3차원 공간에 있는 P_{plane}의 z축 크기를 주행환경의 객체 외곽 형태가 가장 잘 나타나는 차량의 라디에이터 그릴의 높이인 40cm 높이로 맞추어 2차원 공간에 투영한다. 일반적인 차량들에 대해 카메라 공간에 투영된 P_{plane}은 Fig. 2의 (b)처럼 표현된다. Fig. 2-(b)에서 투영된 점들 중 객체의 최외각 형태 성분은 차량과 가장 가까운 점인 카메라 공간에서의 최 하단의 점들이 된다. 카메라 공간은 픽셀 단위의 정수 공간이기 때문에 x축 방향으로 각 열의 최하단 픽셀에 투영된 점들을 남기고 나머지 열의 점들을 모두 제거하게 된다. 이후 중간값 필터(median filter)[14]를 통해 남은 잡음을 제거한다. Fig. 2-(c)는 이 결과를 나타낸다. 이 때 카메라 공간에서의 중간값 필터 결과는 이웃 열에 위치한 점의 1차원적 높이에 따라 결정되지만, 3차원 공간의 각 점들의 좌표는 2차원의 값이 필요하다. 따라서 카메라 공간에서 중간값 필터에 의해 높이 변동이 일어나는 점 p_t에 대응하는 3차원 공간상의 p_t는 중간값을 갖는 점에 통합(merge)시켜 필터링한다. 이 과정에 의해 카메라 공간과 3차원 라이다 공간에서 각 점들을 정리한 결과는 각각 Fig. 3-(a)와 3-(b)와 같다. 이 과정을 통해 재구성된 라이다 점들은 객체의 외곽 표면을 따라 연결된 선의 형태를 띄기 때문에 본 논문에서는 이를 라이다 에지로 부른다.

2.2 경계선 기반의 2차원 프로포절 생성

생성된 라이다 에지들은 카메라 공간에서의 연속성을 통해 각기 그룹으로 분할할 수 있다. 분할된 에지의 경계선과 분류하고자 하는 클래스의 평균 크기 정보를 통해 대략적인 2차원 객체 프로포절을 생성한다.

2.2.1 라이다 에지 분할

위 2.1.절의 외곽선 생성 과정에서 생성된 2차원 라이다 에지를 아래 식 (1)과 같이 거리에 대한 LoG(Laplacian of Gaussian)를 통해 이웃하는 점들 간의 거리 대비 기울기 변화량을 기준으로 분할한다.

$$LoG^{*}f(x)=[\triangle[G_{\sigma}(x)]^{*}f(x) \quad (1)$$

Fig. 2. (a) is the original input image, (b) is the projected image of the P_{plane} in the two-dimensional CCD space, and (c) is the result of generating the two-dimensional LiDAR Edge through filtering.

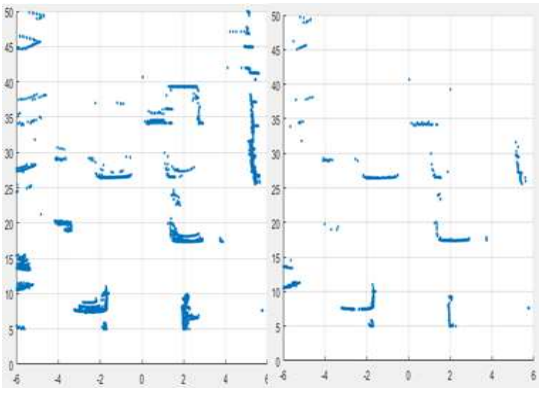

Fig. 3. 3D LiDAR Point Cloud, P_{plane} filtering result. The left side is Pplane before filtering and the right side is Pplane after filtering. You can see that most of the noise points inside the object are arranged.

Fig. 4의 주황색 선은 위 과정을 통해 분할된 라이다 에지와 그 경계선을 나타낸다. 여기에 센서와의 거리 정보와 분류하고자 하는 클래스의 평균 높이 정보를 더해 객체의 프로포절을 생성한다.

2.2.2 프로포절 생성

위 과정에서 분할된 라이다 에지는 하나의 면을 따라 흐르는 객체의 외곽선을 의미한다. 따라서 에지의 양 모서리는 촬영 차량의 시점에서 관측 가능한 객체의 좌우 범위를 나타낸다. 이 에지의 양 끝점 p_1과 p_2에 대해 각 점과 센서와의 거리와 분류하고자 하는 클래스들의 평균 높이 정보를 더해 $376 * 1241$ 픽셀의 해상도를 갖는 KITTI Dataset을 기준으로 아래 식과 같이 프로포절 상자의 높이를 결정한다.

Fig. 4. A box proposal based on the boundary of a filtered 2D LiDAR Edge.

$$C_{Px}^{h} = \frac{C_{real}^{h} \times 750}{d} \quad (2)$$

위 식 (2)의 C^{h}_{Px}는2차원 공간에서의 클래스의 변환 높이로, 구하고자하는 영역 상자의 높이를 의미하고, C^{h}_{real}은 각 클래스의 실제 평균 높이를 나타낸다. d는 센서로부터 각 라이다 점까지의 거리를 의미한다.

Fig. 4의 초록색 영역 상자는 식 3을 통해 생성된 프로포절의 예를 나타낸다. RPN(region proposal network)[3]과 같이 위 과정을 통해 생성된 각 상자를 앵커(anchor)를 통해 확장하여 객체 분류기를 위한 프로포절 집합을 생성한다. 3개의 종횡비와 3개의 비율(scale)을 통해 각 프로포절 영역 상자를 9개로 확장하여 다양한 객체 형태에 대응시킨다.

2.3 2차원 객체 분류와 3차원 확장

2.3.1 2차원 분류기를 통한 객체 분류

2.2에서 생성된 프로포절은 R-CNN(region based convolutional networks)[2] 분류기의 관심영역으로 사용된다. 분류기는 어떤 것을 사용해도 무관하지만, 본 연구는 분류 단계에서 객체의 실제 크기가 아닌 생성된 프로포절의 레이블이 필요하기 때문에 학습 과정에서 객체 전체 영역이 아닌 구획별로 객체의 특징을 각각 학습하는 R-FCN(region based fully-convolutional networks)[5]을 사용하였다. 본 연구에서 생성하는 프로포절들은 객체에 반사되는 센서를 기반으로 생성하기 때문에 내부에 객체를 포함할 가능성이 매우 높지만, 잡음 등으로 인해 객체의 일부분만 포함하는 작은 프로포절도 생성되는데 이러한 경우 R-FCN은 관심 영역을 n x n개의 격자로 나누어 각 부분별로 분류를 수행하기 때문에 프로포절이 객체의 일부분만을 포함하는 경우에도 높은 분류율을 보이는 특성이 있다. 또한 이러한 특성으로 인해 관심 영역이 배경을 포함하여 생성되는 경우 배경 영역에 대해 낮은 점수가 매겨진다. 이 때문에 프로포절이 객체의 일부를 포함하지만, 실측 영역과의 중첩율이 낮은 경우 자체적으로 필터링되는 효과를 기대할 수 있다.

R-FCN을 사용하여 각각 프로포절 상자의 레이블이 결정되면 NMS(non maximum suppression)를 통해 상자를 정리한 뒤 각 상자를 생성한 2차원 라이다 에지와 동일한 점 구성을 갖는 3차원 공간의 라이다 에지에도 각 레이블을 매핑한다.

2.4 3차원 객체 포즈 복원

2.4.1 하위 클래스 정보 사용(sub-class information)

3차원 객체의 포즈 복원을 위해 본 연구는 분류하

고자 하는 클래스의 평균 크기 정보를 사용한다. 평균 크기 정보는 강체 모델(rigid body model)과 유체 모델(soft body model)에 따라 다르게 구성된다. 차량 등의 강체 모델은 입방체(cube) 형태로 표현하며 클래스의 평균적인 종횡비와 각 방향별 평균 크기를 가진다. 또한 사람과 같은 유체 모델의 경우 원통 형태로 표현하며 평균 높이와 지름 정보를 가진다.

2.4.2 3차원 영역 상자 생성

앞서 2.3에서 3차원 공간에 존재하는 라이다 에지에 각각 클래스 레이블을 매핑하였기 때문에 각 레이블에 해당하는 하위 클래스 정보를 사용하여 클래스에 적합한 3차원 상자를 생성한다. 3차원 공간에 상자를 생성하기 위해서는 상자의 모서리 좌표와 방향, 크기에 대한 정보가 필요한데, 상자의 모서리 좌표는 2차원 라이다 에지의 인접 픽셀 관계를 통해 획득할 수 있다. 이웃한 두 2차원 라이다 에지가 인접해 있는 경우 경계선의 에지의 높이가 낮은 쪽이 전경 객체(foreground object)가 되고, 이는 가려지지 않은 객체의 모서리를 의미한다. 따라서 이 점을 기준으로 아래 식 (4)에 의해 얻어진 에지 방향으로 상자를 생성할 수 있다.

$$\theta_i = \frac{p_i^y - \min(L_i^y)}{p_i^x - \min(L_i^x)} \times \frac{180}{\pi} \tag{3}$$

위 식 (4)의 L_i는 i번째 라이다 에지를 구성하는 점들의 그룹을 나타내고 p_i는 L_i의 모서리 점들을 의미한다. 하나의 모서리 점과 한쪽의 방향만으로 3차원 상자를 생성하기 때문에 90도로 틀어진 2개의 영역 상자를 생성할 수 있는데, 이는 이웃 상자와의 중첩 여부와 측정된 라이다 에지의 길이와의 관계를 통해 하나를 선택하게 된다.

2.4.3 2차원 영역 상자 재구성

초기 2차원 라이다 에지를 통해 생성한 2차원 영역 상자는 눈에 보여지는 객체의 형태만을 커버하도록 설계되었기 때문에 실제 객체 영역 전부를 의미하지 않는다. 따라서 하위 클래스 정보를 통해 확장된 3차원 영역 상자를 2차원으로 투영하여 2차원 분류기를 위한 영역 상자를 확장한다. Fig. 5는 확장 전후의 2차원 영역 상자를 나타낸다.

3. 실험 결과 및 결 론

3.1 실험 환경

본 연구는 NVIDIA TITAN X GPU 환경에서 수행되었고 caffe framework[17] 상에서 구현하였다. R-FCN은 원래 Pascal VOC Dataset[18]에서 훈련되었지만, 본 연구에서는 라이다 데이터를 제공하는 KITTI Dataset[19]에서 재학습하여 사용하였다. 본 연구는 2차원 프로포절 생성을 통한 객체 분류를 수행하기 때문에 비교를 위해 널리 사용되고 있는 프로포절 생성기인 selective search[20]와 Faster R-CNN의 RPN, 그리고 edge boxes[15]와 비교하였다. 비교에는 standard mean average precision(mAP)와 최종 분류된 영역 상자와 실측 자료와의 Intersection over uniin(IOU) 비율을 사용하여 평가하였다.

3.2 KITTI Dataset에서의 분류 정확도 측정

본 연구는 라이다 정보와 카메라 정보를 함께 사

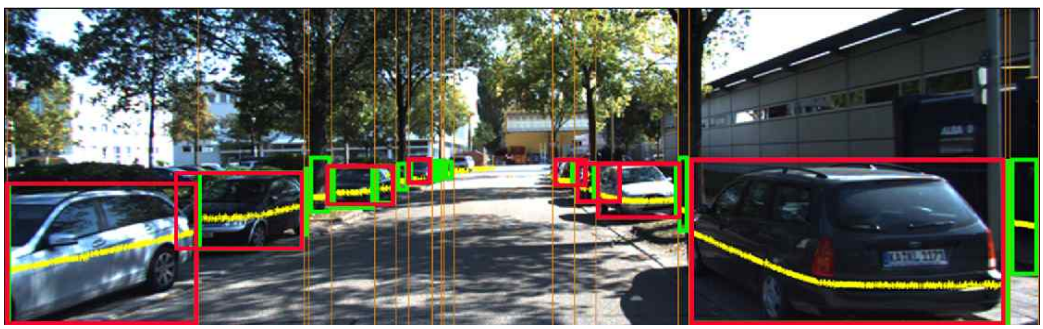

Fig. 5. Comparison of 2D box before and after expansion using projection of 3D bounding box. The red box is after expansion.

용하기 때문에 공개 데이터 세트 가운데 이를 모두 제공하는 KITTI Dataset을 통해 학습을 수행하였다. 학습에 사용된 카메라 데이터는 KITTI 'Object' 항목과 'Tracking' 항목의 학습 시퀀스를 사용하였다. 'Object' 카테고리는 총 7481장의 레이블된 학습용 데이터가 제공되어 이 중 6,000 장을 학습에 사용하고, 평가를 위해 1,481장을 사용하였다. 또한 'Tracking' 시퀀스의 8,008장의 학습용 데이터를 함께 사용하였다.

또한 본 연구에서 제안하는 방법은 라이다를 위한 별도의 학습을 요구하지 않기 때문에 카메라 기반의 R-FCN 구조를 그대로 사용하여 카메라 영상에 대해서만 학습을 수행하였다. 다만, KITTI dataset에서 주어지는 라이다 데이터의 경우 실제 유효 측정 범위가 약 50m정도이기 때문에 라이다 데이터를 중심으로 객체의 프로포절을 생성하는 연구의 특성상 50m 범위 밖의 물체에 대해서는 평가하지 않았다.

아래 Table 1은 '차량', '보행자', '자전거'의 3가지 클래스의 평균 정확도를 측정하여 기존의 연구들과 비교한 것이다. 실측 자료와의 중첩 비율이 50% 이상인 상자에 대한 정밀도를 측정한 결과 특히 '차량'과 '보행자'에 대해 좋은 결과를 나타내었다. '자전거'의 경우 자전거 바퀴의 스포크 형태에 따라 라이다 점들이 충분히 찍히지 않아 상자 생성을 위한 기준을 충족하지 못하는 경우가 발생하여 상대적으로 평균 정확도 점수가 낮게 나타났다.

3.3 Pascal VOC Dataset에서의 분류 정확도 측정

본 연구는 라이다와 카메라 영상의 융합을 통해 구현되었지만, 객체를 분류하는 과정에는 카메라 영상만을 사용하기 때문에 기존의 R-FCN이 별도의 수정 없이 사용되었다. 따라서 R-FCN이 기존 벤치마크에서 평가받은 Pascal VOC 환경의 R-FCN의 caffe 모델을 사용하여 동일한 분류기에서의 selective search와 RPN과의 비교를 수행하였다. Table 2는 본 연구 결과를 정확도와 계산 시간을 통해 다른 두 프로포절 생성기와의 성능을 비교한 것이다. 또한 Table 3은 실측 자료와의 중첩 비율에 따른 평균 정확도의 변화를 나타낸다. Table 3에서 나타난 것처럼, 본 연구 결과의 영역 상자가 다른 두 방법에 비해 보다 정확하게 실제 객체의 영역을 포함하고 있는 것을 알 수 있다. 이는 일반적으로 영상에 나타나는 특징을 통해 객체 영역을 판단하는 기존 프로포절 생성기와는 달리, 하위 클래스 정보를 통해 생성한 3차원 영역 상자를 확장하여 실제 객체 크기에 근접하게 상자를 재조정하기 때문이다.

5. 결 론

본 연구는 자율 주행을 위한 객체 검출과 분류에 있어 효과적으로 위치와 크기를 추정할 수 있는 방법을 제안하였다. 우리는 2차원 및 3차원 공간에서 라이다 데이터를 필터링하여 라이다 점들로 구성된 라이다 에지를 만들었고 이를 기반으로 2차원 프로포

Table 1. Experimental results of the KITTI Dataset

Method	Car		Pedestrian		Cyclist	
	E	M	E	M	E	M
3DVP	87.46	75.77				
SubCat	84.14	75.46				
Regionlet	84.75	76.45	73.14	64.19	74.08	61.31
SDP	90.33	83.53	77.74	61.15	70.41	58.72
Ours	95.42	88.53	81.78	65.72	72.11	60.85

Table 2. Experimental results of the Pascal VOC

Method	Training data	Test data	mAP (%)	Test time (s/image)
RPN	07+12	KITTI	75.7	0.37
SS	07+12	KITTI	77.4	2.21
Edgebox	07+12	KITTI	78.2	0.35
Ours	07+12	KITTI	**82.4**	**0.17**

Table 3. The AP change according to the IOU rate

Method	Training data	Test data	AP@ 0.5	AP@ 0.7	AP@ 0.9
RPN	07+12	KITTI	84.8	77.4	55.2
SS	07+12	KITTI	86.3	80.4	58.4
Edgebox	07+12	KITTI	87.1	81.1	59.7
Ours	07+12	KITTI	**89.7**	**82.5**	**77.2**

절을 생성하였다. 이를 CNN 기반의 분류기(본문에서는 R-FCN)를 통해 분류하고, 3차원 공간의 라이다 에지의 모서리와 방향을 기준으로 분류된 클래스 레이블의 평균 크기의 3차원 영역 상자를 생성하였다. 또한 이를 2차원에 재투영하여 2차원과 3차원 공간 모두에 적합한 객체 검출을 완성하였다. 필터링 과정에서 객체를 판단하는데 사용되는 라이다 점들의 수가 크게 줄었으며 이를 기반으로 생성되는 프로포절 역시 기존 연구 대비 더 적은 숫자로 객체가 존재할 가능성이 있는 위치에 집중적으로 생성되어 객체 검출의 효율성을 높이고 전체 프로세스의 소요 시간을 크게 줄였다. 또한 3차원 공간에서 실제 객체의 크기만큼 상자를 확장하는 과정을 통해 2차원 공간에서 가려져있던 객체의 부분에 대해서도 효과적으로 복원하여 실측 자료 대비 검출 상자의 중첩율을 크게 높였다. 다만, 라이다 센서의 탐지 범위를 벗어나는 객체나, 카메라 영상에서는 보이지만 가까운 차량에 의해 라이다 센서가 가로막힌 차량의 경우 검출에 어려움이 있다. 우리는 향후 이 문제를 카메라 영상에서의 특징과의 융합 프로세스를 통해 풀어갈 것이다.

REFERENCE

[1] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems*, pp. 1097-1105, 2012. (NIPS)

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Setection Detection and Semantic Segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580-587, 2014.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks," *Advances in Neural Information Processing Systems*, pp. 91-99, 2015.

[4] T. Kong, A. Yao, Y. Chen, and F. Sun, "Hypernet: Towards Accurate Region Proposal Generation and Joint Object Detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 845-853, 2016.

[5] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object Detection Via Region-based Fully Convolutional Networks," *Advances in Neural Information Processing Systems*, pp. 379-387, 2016.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-time Object Detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779-788, 2016.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, and C.Y. Fu, et al., "SSD: Single Shot Multibox Detector," *Proceeding of European Conference on Computer Vision,* pp. 21-37, 2016.

[8] X. Chen, K. Kundu, Y. Zhu, A.G. Berneshawi, H. Ma, and S. Fidler, et al., "3D Object Proposals for Accurate Object Class Detection," *Advances in Neural Information Processing Systems*, pp. 424-432, 2015.

[9] G.R. Cross and A.K. Jain, "Markov Random Field Texture Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, No. 1, pp. 25-39, 1983

[10] C. Cortes and V. Vapnik, "Support-vector Networks," *Machine Learning*, Vol. 20, No. 3, pp. 273-297, 1995.

[11] D.Z. Wang and I. Posner, "Voting for Voting in Online Point Cloud Object Detection," *Proceeding of Robotics: Science and Systems*, 2015.

[12] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Data-driven 3D Voxel Patterns for Object Category Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1903-1911, 2015.

[13] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategory-aware Convolutional Neural Networks for Object Proposals and Detection," *Proceeding of IEEE Winter Conference on Applications of Computer Vision*, pp. 924-933, 2016.

[14] T. Huang, G. Yang, and G. Tang, "A Fast Two-dimensional Median Filtering Algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 27, No. 1, pp. 13-18, 1979.

[15] C.L. Zitnick and P. Dollar, "Edge Boxes: Locating Object Proposals from Edges," *Proceeding of European Conference on Computer Vision,* pp. 391-405, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, and R. Girshick, et al., "Caffe: Convolutional Architecture for Fast Feature Embedding," *Proceeding of the 22nd Association for Computing Machinery International Conference on Multimedia,* pp. 675-678, 2014.

[18] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (voc) Challenge," *International Journal of Computer Vision*, Vol. 88, No. 2, pp. 303-338, 2010.

[19] A. Geiger, P. Lenz, and R. Urtasun, "Are We Ready for Autonomous Driving-the Kitti Vision Benchmark Suite," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354-3361, 2012.

[20] J.R.R. Uijlings, K.E.A. Van De Sande, T. Gevers, and A.W.M. Smeulders, "Selective Search for Object Recognition," *International Journal of Computer Vision*, Vol. 104, No. 2, pp. 154-171, 2013.

[21] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for Generic Object Detection," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 17-24, 2013.

[22] E. Ohn-Bar and M.M. Trivedi, "Learning to Detect Vehicles by Clustering Appearance Patterns," *IEEE Transactions on Intelligent Transportation Systems*, Vol. 16, No. 5, pp. 2511-2521, 2015.

[23] F. Yang, W. Choi, and Y. Lin, "Exploit All the Layers: Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classifiers," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2129-2137, 2016.

[24] D. Kim and G. Hyun, "Object Analysis on Outdoor Environment Using Multiple Features for Autonomous Navigation Robot". *Journal of Korea Multimedia Society*, Vol. 13, No. 5, pp. 651-662. 2010.

김 정 언

2007년 가톨릭대학교 컴퓨터공학과(학사)
2009년 가톨릭대학교 컴퓨터공학과(석사)
2010년~현재 가톨릭대학교 미디어공학과 박사과정

관심분야 : 기계학습, 영상처리, 자율주행

강 행 봉

1980년 한양대학교 전자공학과(학사)
1986년 한양대학교 전자공학과(석사)
1989년 Ohio State Univ. 컴퓨터공학(석사)
1993년 Rensselaer Polytechnic Institute 컴퓨터공학(박사)
1993년~1997년 삼성종합기술원 수석연구원
1997년~현재 가톨릭대학교 디지털미디어학과 교수
2005년UC Santa Barbara, Visiting Professor
관심분야 : 컴퓨터비전, 기계학습, HCI, 컴퓨터그래픽스, 인공지능, 빅데이터