라이다 포인트 클라우드 시계열 데이터를 이용한 시공간 특징지도 정렬 및 융합 기반 3 차원 물체 검출

고준호° 최준원°°

한양대학교 전기공학과° 한양대학교 전가공학과°°

3D Object Detection Based on Spatio-Temporal Feature Alignment and Aggregation using Sequential LiDAR Point Cloud Data

Junho Koh° Jun Won Choi°°

Dept. of Electrical Engineering° Dept. of Electrical Engineering°°

Hanyang University Hanyang University

jhkoh@spa.hanyang.ac.kr junwchoi@hanyang.ac.kr

요 약

최근 라이다 기반 3 차원 물체 검출 기법과 관련된 연구가 활발하게 진행되어 자율 주행 인지 분야에서 중요한 역할을 하고 있다. 하지만, 기존 연구들은 포인트 클라우드 시계열 데이터를 이용하여 시간 정보를 활용하지 않아 실제 주행 환경에서 들어온 시계열 데이터의 시간 정보를 활용하지 못한다. 본 논문에서는 포인트 클라우드 시계열 데이터를 입력으로 시공간 특징지도를 재배열하고 융합하는 3 차원 물체 검출 기술을 소개한다.

1. 서론

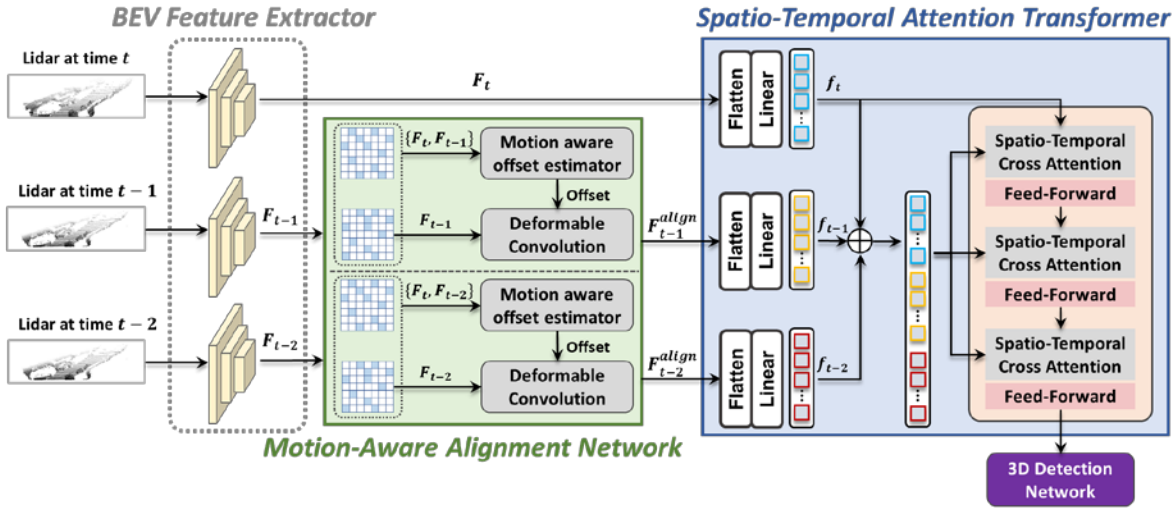
최근 딥러닝의 발전으로 여러 컴퓨터 비전 분야가 상당한 발전을 이루었다. 또한, 이러한 컴퓨터 비전 분야의 발달과 함께 자율주행에서의 인지 성능이 비약적으로 향상하였다. 특히, 3 차원 물체 검출 기술 [1]-[4]은 자율주행 인지 분야에서 매우 중요한 역할을 하기 때문에 연구가 활발히 진행되고 있다. 이러한 3 차원 물체 검출 기술은 단일 시간에 얻은 포인트 클라우드 정보만을 사용하여 물체 검출을 수행한다. 실제 자율주행 환경에서는 연속된 포인트 정보가 들어와 시계열 데이터를 구성하지만, 기존 3 차원 물체 검출 기술은 시계열 데이터에서 추출가능한 시간 정보를 활용하지 못한다.

본 논문에서는 포인트 클라우드 시계열 데이터를 이용하여 이전 포인트 클라우드에서 얻은 특징지도를 현재 특징지도에 정렬하고, 정렬된 특징지도와 현재 특징지도를 융합하는 방식으로 시간정보를 활용하는 3 차원 물체 검출기인 3D AlignNet 을 제안한다. 이전 특징지도를 현재 특징지도에 정렬하기 위하여 물체의 움직임이 반영된 모션 특징지도를 추출하고 이를 입력으로 Deformable Convolution [5]에서 사용될 offset 과 attention weight 를 계산한다. 이렇게 계산된 offset 과 attention weight 는 이전 특징지도를 입력으로 Deformable Convolution 을 수행하는데 사용되어 이접 특징지도를 정렬한다. 정렬된 이전 특징지도와 현재 특징지도는 최근 시계열 데이터를 처리하여 융합하는데 널리 사용되는 Transformer 구조를 활용한다. 기존 Transformer 방식보다 효율적이고 선택적으로 융합하는데 용이한 Deformable DETR [6]에서 영감을 얻어 시공간 특징지도 융합을 수행한다. 이렇게 제안된 3 차원 물체 검출기는 베이스라인 3 차원 물체 검출기인 PointPillar [2]와 비교하여 nuScenes [7] 공개 데이터셋에서 높은 성능을 나타내고 있다.

2. 제안하는 3 차원 물체 검출기

3D AlignNet 은 시간적으로 특징지도를 현재 특징지도에 재배열하고 융합하여 시간정보를 활용하는 3 차원 물체 검출 기술이다. 특히, 시간적으로 특징지도를 융합하기 이전에 재배열하는 작업을 진행하여 시공간 특징지도 융합을 더욱 정교하게 수행한다.

우선 입력으로 들어온 시계열 포인트 클라우드 데이터를 각각 복셀화를 진행하고 Convolutional Neural network (CNN)을 거쳐 bird’s eye view (BEV) 특징 지도를 추출한다. 위 작업은 시간적으로 가중치 (weight)를 공유하여 이루어진다. 이렇게 얻은 BEV 특징지도들은 Deformable convolution [5] 기반의 Motion-Aware Alignment Network (MA2Net)를 이용하여 현재 특징지도의 위치에 맞도록 정렬



(Alignment) 된다. 과정은 다음과 같다. 우선 이전 BEV 특징지도와 현재 BEV 특징지도의 차이를 계산하고, 이를 입력으로 여러 Non-local block [8]을 통과하여 물체의 움직임 정보에 집중된 특징지도를 추출한다. 이를 이전 BEV 특징지도와 다시 합쳐주어 움직임 정보가 포함된 모션 BEV 특징지도를 추출한다. 이렇게 얻은 모션 BEV 특징지도는 Deformable Convolution 을 연산하는데 사용되는 offset 과 attention weight 를 CNN 을 통해 각 픽셀마다 얻는데 사용된다. 이렇게 얻어진 offset 과 attention weight 를 통해 이전 특징지도를 Deformable Convolution 을 통과하여 정렬된 이전 BEV 특징지도를 추출한다. 위 작업은 각 시간마다 얻어진 특징지도에서 모두 이루어진다.

추출된 정렬된 BEV 특징지도와 현재 특징지도는 시공간 특징지도 융합을 수행하기 위하여 Spatio-Temporal Attention Transformer (STAT)의 입력으로 사용된다. 과정은 다음과 같다. 우선, 현재 BEV 특징지도를 Query 로 지정하고, 이전 정렬된 BEV 특징지도와 현재 BEV 특징지도를 모아 Key 로 지정한다. Query 특징지도에서 reference point 로 지정된 해당 픽셀의 특징 벡터를 입력으로 각 시간의 특징지도에서 어떤 위치의 특징 벡터를 합쳐주고 얼마나 가중치를 부여할지 Fully Connected 레이어를 통해 계산한다. 각 시간마다 선택된 특징벡터들은 계산된 가중치가 곱해지고 모두 더해져 융합된 특징벡터를 추출한다. 위 작업을 현재 특징지도의 각 픽셀마다 모두 작업하여 시공간적으로 융합된 BEV 특징지도를 추출하고, 이를 입력으로 최종 3 차원 물체 검출 결과를 얻는다.

3.데이터셋과 실험결과

본 논문에서 실험은 nuScenes [7] 공개 데이터셋을 사용하여 학습과 성능평가를 수행하였다. 표 1 은 본 논문에서 베이스라인으로 사용한 PointPillar[2] 3 차원 물체 검출기를 기준으로 각 제안한 네트워크를 순차적으로 추가할 때마다 성능이 얼마나 향상되는지를 보여주는 표이다. STAT 을 추가한 경우 1.19%의 성능 향상을 보였다. 또한 MA2Net 을 추가하였을 때 4.18%의 성능 향상을 보였다. 이는 시공간 특징지도 융합을 수행하기 전에 특징지도를 정렬하는 것이 얼마나 중요한 작업인가를 증명해준다.

Network	mAP (%)
PointPillar	41.18
PointPillar + STAT	42.37
PointPillar + STAT+MA2Net	45.36

4. Acknowledgement

이 성과는 정부 (과학기술정보통신부)의 재원으로 한국 연구재단의 지원을 받아 수행된 연구임 (2020R1A2C2012146)

5. 참고 문헌

[1] Y. Yan, Y. Mao and B. Li, “Second: Sparsely embedded Convolutional Detection,” *Sensors* (2018), .
[2] A. H. Lang, S.Vora, H. Caesar, L.Zhou, J. Yang and O. Beijbom, “PointPillars: Fast Encoders for Object Detection from Point Clouds,” in CVPR (2019).
[3] Y. Zhou and O. Tuzel, “VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection,” in CVPR (2018).
[4] S. Shi, C. Guo, L. Jiang, Z. Wang, Z. Shi, X. Wang and H. Li, “PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection,” in CVPR (2020).
[5] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu and Y. Wei, “Deformable Convolutional Networks,” in ICCV (2017).
[6] X. Zhu, W. Su, L. Lu, B. Li, X. Wang and J. Dai, “Deformable DETR: Deformable Transformers for End-to-End Object Detection,” in ICLR (2020).
[7] H. Caesar, V. Bankiti, A. H. Lang, S.Vora, V. E. Liong, Q. Xu, A.Krichnan, Y. Pan, G. Baldan and O. Beijbom, “nuScenes: A multimodal dataset for autonomous driving,” in CVPR (2020).