시공간 움직임 정보를 활용한 다중 시점 카메라 기반의 3 차원 비디오 객체 검출 기술

이영우, 고준호, 최준원*
한양대학교

youngwoolee@spa.hanyang.ac.kr, jhkoh@spa.hanyang.ac.kr,

junwchoi@hanyang.ac.kr

Multi-view Camera-based 3D Video Object Detection using Spatio-temporal Motion Context

Youngwoo Lee, Junho Koh, Jun Won Choi*
Hanyang University

요 약

본 논문은 시공간 정보를 활용하여 다중 시점 3 차원 비디오 객체 검출을 수행하기 위한 새로운 시간 융합 구조를 제안하였다. 제안된 네트워크는 장기간 카메라 비디오 데이터가 가지는 시공간 정보를 효율적으로 활용하기 위하여 Recurrent temporal fusion 전략을 채택하였다. 연속된 다중 시점 비디오 프레임에서 얻은 특징지도를 어텐션 기반의 정렬 및 집계 과정을 거쳐 강인한 bird's eye view (BEV) 특징지도를 추출하는 것을 목표로 한다. 먼저 자차 및 주변 차량 등의 동적 객체의 움직임 맥락 정보를 기반으로 이전 BEV 특징지도를 현재 BEV 특징 지도 공간에 맞춰 정렬한다. 그 다음, 정렬된 BEV 특징지도는 Gated attention 모델을 활용하여 시공간으로 융합된 BEV 특징지도를 추출한다. 이는 최종 3 차원 객체 검출에 활용되며, 기존 Baseline 대비 높은 성능을 달성하였다. 실험은 nuScenes 데이터셋에서 진행되었으며, 기존 단일 프레임 기반 3 차원 객체 검출 알고리즘 대비 높은 성능을 달성하였다.

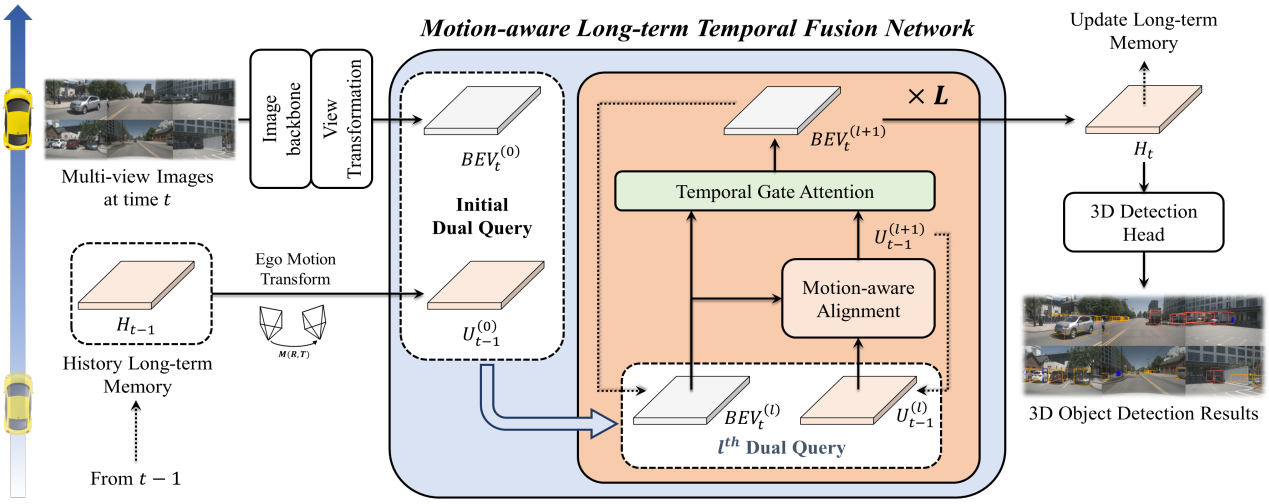

Figure 1. 제안하는 네트워크 전체 구조

Ⅰ. 서 론

본논문에서는 다중 시점 카메라 기반의 3 차원 비디오 객체 검출 수행을 위한 효율적인 시공간 융합 네트워크를 소개한다. 실제 다중 시점 카메라에서 취득되는 데이터는 연속적인 시계열 구조를 가지고 있기 때문에, 이 연속적인 데이터 내에 존재하는 시간정보를 장기간으로 활용하는 것이 중요하다. 이를 다루고자 과거 시공간적으로 융합된 특징지도를 모델 내 메모리에 저장하고, 현재 시점 프레임이 입력으로 들어오면 과거 융합된 BEV 특징지도를 현재 BEV 특징지도와 융합하여 연속적으로 장기간 시공간 정보를 활용하도록 네트워크를 설계하였다. 또한, 시계열 정보를 활용함에 있어 물체의 움직임으로 인한 오차를 정렬하기 위한 네트워크를 통해 효과적으로 장기간 시공간 정보를

활용하도록 네트워크를 설계하였다. 이를 통해 기존 기술 대비 높은 성능의 3 차원 객체 검출 결과를 얻을 수 있다.

Ⅱ. 본론

2.1) 제안하는 다중 시점 3 차원 객체 검출 기술

제안하는 3 차원 객체 검출기는 다중 시점 카메라 비디오 데이터를 장기간으로 시공간 융합을 하도록 모델을 구성하였으며, 전체 구조는 Fig. 1 과 같다.

과거 시점 (t-1)까지 시공간적으로 융합된 BEV 특징지도를 메모리에 저장하고 현재 시점에서 3 차원 객체 검출을 수행하는데 사용한다. 현재 다중 시점 RGB 이미지를 입력으로 BEV 특징지도를 추출한다. 메모리에 저장된 시공간 BEV 특징지도는 자차의 움직임을 반영하여 공간 변환을 수행하고, 현재 시점 BEV 특징지도와 함께 시공간 융합 네트워크의 입력으로 사용된다.

시공간 융합을 수행하기 위하여 Deformable attention 기법을 활용한다. 먼저 현재 시점 BEV 특징지도와 과거 시공간 BEV 특징지도 사이의 움직임 맥락정보를 생성한다. 이 움직임 맥락정보는 어텐션을 수행하기 위한 query 로서 활용되며, 과거 시점 BEV 특징지도를 업데이트 하는데 사용된다. 업데이트된 과거 시공간 BEV 특징지도는 현재 BEV 특징지도와 각각의 기여도를 기반으로 융합하기 위하여 Gated attention 기법을 활용하였다. 최종적으로 생성된 현재 시점 시공간 BEV 특징지도는 3 차원 객체 검출을 수행하는데 사용되며, 다음 시점 (t+1) 3 차원 객체 검출 수행을 위하여 메모리에 저장한다.

2.2) 듀얼 어텐션 기반 시공간 융합 네트워크

과거 시공간 BEV 특징지도와 현재 BEV 특징지도 사이의 움직임 맥락 정보를 활용하여 두 단계의 어텐션을 통해 시공간 융합을 수행한다.

먼저 과거 시공간 BEV 특징지도와 현재 BEV 특징지도 사이의 차이를 인코딩하여 움직임 맥락 정보를 추출한다. 추출된 움직임 맥락정보는 Deformable attention 의 query 로서 과거 시공간 BEV 특징지도를 정렬하는데 사용된다. 업데이트된 과거 시공간 BEV 특징지도는 현재 BEV 특징지도와 Gated attention 구조 기반의 시공간 융합 네크워크의 입력으로 활용된다. 각 시점의 기여도를 convolution network 를 통해 계산하여 두 BEV 특징지도를 합치는데 사용된다. 이러한 듀얼 어텐션 과정은 멀티 레이어로 수행되어 점진적으로 과거 BEV 특징지도와 현재 BEV 특징지도를 업데이트한다.

Table 1. nuScenes 데이터셋을 이용한 실험 결과

Method	NDS (%)	mAP (%)
BEVDepth[2]	32.6	30.4
SOLOFusion[3]	49.4	40.4
Our method	50.2	40.9

2.3) 실험 결과

제안하는 네트워크를 학습 및 성능 평가를 실시하기 위하여 nuScenes 공개 데이터셋을 활용하였다.

제안하는 기법의 성능 평가 결과는 Table 1 과 같이 나타났다. 대표적인 단일 프레임 기반의 다중 시점 3 차원 객체 검출기인 BEVDepth[2]에 비해 높은 성능을 나타냈으며, 대표적인 시간 정보를 활용한 다중 시점 3 차원 비디오 객체 검출 기술인 SOLOFusion[3]에 비해 높은 성능을 달성하였다.

Ⅲ. 결론

본논문에서는 다중 시점 카메라 기반의 3 차원 비디오 객체 검출기를 수행하기 위한 효율적인 시공간 융합 네트워크를 소개하였다. 효율적인 시공간 융합 네트워크 수행을 위하여 Recurrent fusion 방식을 채택하였다. 시공간 융합을 수행하기 위하여 Deformable attention 과 Gated attention 기법을 순차적으로 수행하는 듀얼 어텐션 네트워크를 제안하였다. 해당 알고리즘은 nuScenes 자율주행 공개 데이터셋에서 학습 및 성능 평가가 이루어졌으며, 기존 알고리즘 대비 매우 높은 성능 향상을 이루었다.

ACKNOWLEDGMENT

“이 논문은 2024 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2020R1A2C2012146)”

참 고 문 헌

[1] Zhu, Xizhou, et al. "Deformable DETR: Deformable Transformers for End-to-End Object Detection." In International Conference on Learning Representations. 2020.

[2] Li, Yinhao, et al. "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection." In Proceedings of the AAAI Conference on Artificial Intelligence. 2023.

[3] Park, Jinhyung, et al. "Time Will Tell: New Outlooks and A Baseline for Temporal Multi-View 3D Object Detection." In International Conference on Learning Representations. 2022.