객체 추적 성능향상을 위한 Heatmap Detection 및 Transformer 기반의 MOT 모델 설계

양현성[1], 심춘보[2], 정세훈[3]
[1]순천대학교 IT-Bio융합시스템전공 석사과정
[2]순천대학교 인공지능학부 교수
[3]순천대학교 컴퓨터공학과 교수
naiu8165@naver.com, shjung@scnu.ac.kr

Design of a MOT model based on Heatmap Detection and Transformer to improve object tracking performance

Hyun-Sung Yang[1], Chun-Bo Sim[2], Se-Hoon Jung[3]
[1]Interdisciplinary Program in IT-Bio Convergence System, Sunchon National University
[2]Dept. of Artificial Intelligence Engineering, Sunchon National University
[3]Dept. of Computer Engineering, Sunchon National University

요 약

본 연구는 실시간 MOT(Multiple-Object-Tracking)의 성능을 향상시키기 위해 다양한 기법을 적용한 MOT 모델을 설계한다. 연구에서 사용하는 Backbone 모델은 TBD(Tracking-by-Detection) 기반의 Tracking 모델을 사용한다. Heatmap Detection을 통해 객체를 검출하고 Transformer 기반의 Feature를 연결하여 Tracking 한다. 제안하는 방법은 Anchor 기반의 Detection의 장시간 문제와 추적 객체 정보 전달손실을 감소하여 실시간 객체 추적에 도움이 될 것으로 사료된다.

1. 서론

MOT는 컴퓨터 비전 분야의 영상 분석 기술에 있어서 중요한 주제 중 하나이다. MOT는 영상에 포함된 여러 객체의 궤적을 추정하고, 각 Frame과 연결하는 것으로 수행되며 다양한 응용 분야에서 활용되고 있다. 교통 분야에서는 자동차나 보행자와 같은 객체를 추적하고, 자율주행 차량에 적용하여 주변 환경을 실시간으로 인식함으로써 안전한 주행을 보장한다. 또한, 보안 분야에서는 CCTV 모니터링 및 이상행동 및 사건 예방을 위해 사용된다.

Tracking 작업은 실시간으로 이루어지기 때문에, 매우 높은 성능이 요구된다. TBD는 주로 사용되는 Tracking 접근방법 중 하나이다. Anchor 기반으로 진행되는 Detection을 통해 객체의 위치를 탐지하고 연결된 Frame의 객체 정보를 연결하여 Trajectory 내 객체를 추적한다. Tracking의 성능은 Object Detection 정확도와 탐지된 객체 간 Association 방법에 영향받기 때문에 Detection과 Association 모두 중요하다고 볼 수 있다[1]. 하지만, Detection 과정에서 사용되는 다수의 Anchor는 장시간 Detection을 지속시키기 때문에, 자원 소모가 크게 발생할 수 있다. 또한, Association 과정에서 생기는 객체 정보 손실은 Tracking의 정확도를 저하시킨다.

본 연구는, Anchor 기반의 Detection에서 걸리는 시간을 줄이기 위해 Key Point Estimation 기반의 Heatmap Detection을 사용한다. 또한, 객체의 Appearance, Motion 정보의 손실을 방지하기 위해 Transformer 기반의 객체 모델을 사용하고자 한다. 이를 기반으로 실시간 Tracking에 적합한 MOT를 설계하여 성능향상을 기대한다.

2. 관련 연구

[2] Heatmap Detection에 관한 연구에서는 객체의 중심점을 예측하고, 이를 기반으로 객체의 경계상자와 객체 분류를 수행했다. 이전 연구들과 달리 Keypoint Detection, Bounding Box Regression, Object Classification을 하나의 네트워크로 통합하여 효율적인 학습을 진행했다.

[3] 객체 추적을 위한 객체 정보 추출에 관한 연구에서는 탐지된 객체의 외형 정보라고 볼 수 있는 Appearance 정보를 CNN(Convolutional Neural Network)을 활용하여 추출했다. Motion 정보는 객체의 움직임에 대한 정보로 매 Frame에서 객체 간의 유사도를 측정하여 추출한다. 이러한 방법은 현재까

지도 Tracking 연구에 이바지하고 있다.

[4] 객체 탐지와 추적을 동시에 수행하는 연구에서는 Transformer 기반의 MOTR(Multiple-Object Tracking with TRansformer)을 제안했다. MOTR은 시간적인 정보를 처리하는 Transformer의 Attention 기법을 이용하여 물체를 인식하고 추적한다. 이를 위해 모든 Frame 속 객체의 Proposals, Feature 및 Queries를 생성한다. Transformer 기반의 MOTR은 객체 정보의 손실을 최소화하기 때문에 높은 성능을 보였다.

3. 제안하는 방법

Transformer 기반의 MOTR 모델은 객체 탐지와 추적을 하나의 End-to-End 학습 방법으로 진행한다. 이러한 학습 방법은 복잡한 모델 구조 대신 단순한 모델 구조를 갖고, 높은 성능과 데이터 사용 효율성이 높다는 장점이 있다. 그러나, End-to-End 학습 방법은 높은 성능을 내기 위해 많은 양의 학습 데이터를 요구하고, 모델 구조상 입력과 출력 사이에 파이프라인이 없어 네트워크 구조가 복잡하여 실시간 처리에 문제가 될 수 있다.

본 연구는 Heatmap Detection과 Transformer 기반의 MOTR 모델을 결합하여 실시간 다중 객체 추적을 위한 MOT 모델을 설계한다.

3.1 객체 추적을 위한 Heatmap Detection 모델 설계

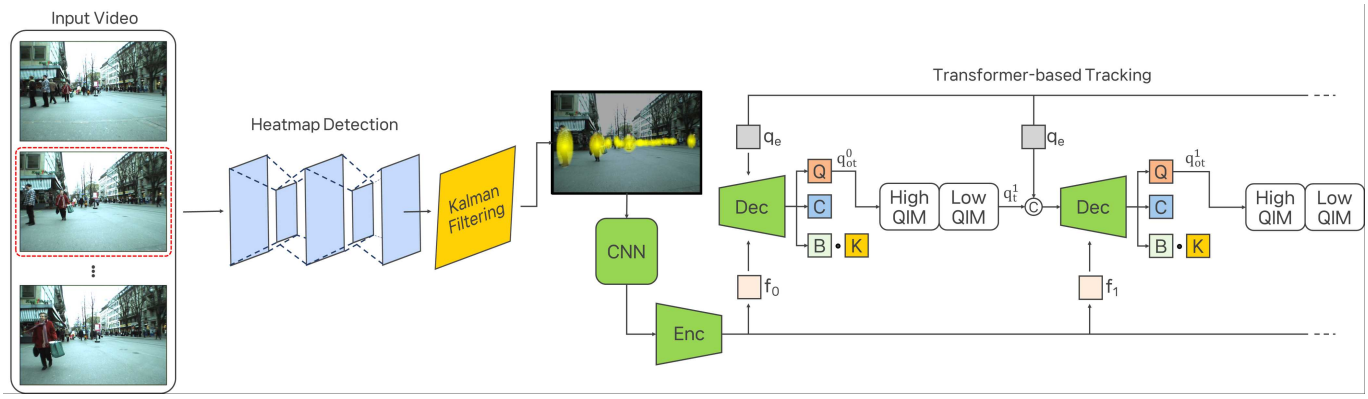
Anchor 기반의 모델은 CNN을 통해 입력 영상의 Feature Map을 추출하여 각 Frame의 객체를 탐지한다. 이를 위해 수천 개의 Anchor를 사용하는 객체 탐지 과정은, 훈련에 장시간 소요된다는 문제가 있다. 반면, Heatmap 기반의 Detection은 객체마다 단 하나의 Keypoint인 Center Point를 사용하기 때문에, 수천 개의 Anchor Processing을 요하지 않아 단시간의 훈련을 보장한다.

Heatmap Detection은 동일 객체가 여러 Heatmap 픽셀에 걸쳐 검출될 경우 한 객체를 N개의 객체로 인식하여 더 많은 계산량을 요구할 수 있다. 따라서, 하나의 객체로 인식하기 위해 후처리 과정으로 NMS(Non-Maximum Suppressions)와 KF(Kalman Filtering)를 적용한다. KF는 주로 Tracking에서 Motion Cost를 다루기 위해 사용되지만, Detection의 후처리로 사용될 경우 객체의 위치 정보를 평활화하여 분산된 Heatmap을 하나의 객체로 인식할 수 있다.

3.2 객체 추적 알고리즘

실시간 객체 추적은 객체 고유정보를 사용하여 진행한다. 이를 위해 Trajectory에 존재하는 객체 특유 Appearance와 Motion 정보를 사용한다. 하지만, 매 Frame에 등장하는 다양한 객체나 한 번 등장하는 객체와 같은 불필요한 정보는 Trajectory의 저장공간 낭비와 정보 손실을 유발한다.

객체의 정보 전달손실을 줄이기 위해 Transformer 기반의 Tracking 모델을 설계한다. 모델의 전체 구조는 Deformable DETR[5]의 구조를 갖는다. CNN을 통해 입력 영상 속 객체의 고유 Feature Map을 추출하고 Transformer 기반의 Encoder로 입력한다. 처음 Decoder는 각 Encoder 출력과 비어있는 Query를 입력으로 받아 C(Class-token), B(Bounding Box-token)와 Q(Query)를 출력한다. C, B, Q는 다음 Frame의 Encoder 입력으로 사용된다. 신경망의 깊이가 깊어질수록 Feature가 손실되는 문제[6]처럼, Feature는 시간상 인접한 Frame에 비해 멀리 떨어진 Frame의 정보 전달손실이 크다. 객체 정보 손실을 줄이기 위해 Detection 단계에서 얻은 KF 정보를 사용한다. KF 정보는 객체의 위치 정보뿐만 아니라 다음 Frame의 객체 존재 확률 정보를 포함한다. Explicit 정보만 사용하는 기존 Tracking 방법들과 달리, 확률 정보를 포함하



(그림 1) 제안하는 Object Tracking 전체 흐름 설계도

는 Implicit 정보는 객체 위치 예측에 도움을 줄 수 있다. 따라서, Decoder의 출력으로 나온 B와 Detection 단계에서 출력한 정보 K를 내적 하여 매니폴드 차원 및 정보 전달손실을 감소시키고자 한다. Feed-Forward 방식은 MOTR과 같은 방식을 사용하고, Decoder의 출력을 다음 Decoder의 입력으로 보내는 과정에서 객체 연결 정보를 위한 QIM(Query Interaction Module)을 사용한다.

Tracking의 필수적인 과정은 다양한 객체에 대한 추적 처리다. 추적 중인 객체는 지속해서 진행해야 하고, 새롭게 등장하거나 사라지는 객체는 Trajectory에서 제거해야 한다. 새로운 객체와 사라진 객체 처리는 임계치와 이전 Frame 정보를 확인하여 진행한다. 낮은 임계치를 고려하는 방법[7]은 Low Score 객체를 한 번 더 사용하여 높은 정확도를 보였다. 따라서, 높은 정확도를 위해 두 번의 QIM을 진행한다.

4. 결론

MOT에 관한 다양한 방법론은 성능 향상과 실시간 추적을 위해 연구되고 있다. 본 연구는 실시간 추적을 위해 기존 TBD 방식 기반의 Tracking 모델을 설계했다. 제안하는 Tracking 방법은 Anchor 기반의 객체 Detection이 아닌 Heatmap Detection을 사용하여 검출 시간을 줄이고, 정확한 객체 추적을 위해 Transformer를 사용하여 Feature 정보 전달손실을 감소하고자 한다. Detection 과정에서 사용된 KF 정보는 객체의 정보 손실 감소를 위해 재사용된다. 또한, 다양한 객체 처리를 위해 두 번의 QIM을 진행한다. 추후 이를 구현하여 Detection 과정에서 발생하는 속도 문제를 완화하고, 정보 전달손실을 감소하여 실시간 Tracking을 하고자 한다.

사사문구

이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2021R1I1A3050843). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2021R1I1A3050843).

참고문헌

[1] Alex Bewley and et al, “Simple online and realtime tracking”, *In IEEE international conference on image processing*, pp. 3464-3468, 2016.

[2] Kaiwen Duan and et al, “Centernet: Keypoint triplets for object detection”, *In Proceedings of the IEEE/CVF international conference on computer vision,* pp. 6569-6578, 2019.

[3] Nicolai Wojke, Alex Bewley and Dietrich Paulus, “Simple online and realtime tracking with a deep association metric”, *In IEEE international conference on image processing*, pp. 3645-3649, 2017.

[4] Zeng, Fangao and et al, “Motr: End-to-end multiple-object tracking with transformer”, In Computer Vision – ECCV 2022: 17th European Conference, Cham: Springer Nature Switzerland, pp. 659-675, October 2022.

[5] Carion, Nicolas and et al, "End-to-end object detection with transformers", Computer Vision – ECCV 2020: 16th European Conference, Springer International Publishing, August 2020.

[6] Kaiming HE and et al, “Deep residual learning for image recognition”, *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.

[7] Zhang Yifu and et al, “Bytetrack: Multi-object tracking by associating every detection box”, *In Computer Vision – ECCV 17th European Conference*, pp. 1-21, 2022.