다중 객체 비디오에서의 어텐션 기반 단일 객체 추적 모델 연구

안성현[O1] 조영완[2] 박상현[†2]
[1]가톨릭대학교 컴퓨터정보공학부
[2]연세대학교 컴퓨터과학과
skd@catholic.ac.kr, jyy1551@yonsei.ac.kr, sanghyun@yonsei.ac.kr

Attention based Single Object Tracking Model In Multiple Object Video

Sunghyun Ahn[O1] Youngwan Jo[2] Sanghyun Park[†2]
[1]School of Computer Science and Information Engineering, Catholic University
[2]School of Computer Science, Yonsei University

요 약

단일 객체 추적(Single Object Tracking)이란 비디오에서 선택된 객체를 추적하는 기술이다. 단일 객체 추적 연구는 Fully Convolutional Siamese Network (Siam FC)를 기반으로 하며, 모델은 크게 세 가지 네트워크로 구성된다. 먼저 특징 추출기는 추적 대상을 포함하는 템플릿 영상과 추적 대상을 찾을 검색 영상의 특징을 각각 추출한다. 이후 특징 융합 네트워크는 추출된 특징에 대한 유사도를 계산하고, 예측-헤드 네트워크는 가장 높은 유사도를 통해 검색 영상에서 추적 대상의 위치를 예측한다. 그러나 유사도를 이용한 객체 추적 방식은 여러 객체들이 동시에 움직이는 비디오에서 잘못된 객체를 추적할 수 있다는 문제점이 있다. 따라서 우리는 정확한 추적 대상 탐지를 위해 특징 추출기와 예측-헤드 네트워크에서 어텐션 모듈을 적용하는 방식을 제안한다. 제안하는 방법론을 검증하기 위해 GOT-10k 데이터셋에 대해 기존의 모델과 비교 실험을 진행하였으며, 제안하는 방법론의 우수성을 검증하였다.

1. 서 론

단일 객체 추적(Single Object Tracking)이란 비디오의 특정 프레임에서 객체를 포함하는 영역이 입력될 때, 이후의 프레임들에서 해당 객체를 추적하는 기술이다[1]. 이 기술은 스마트시티에서 치안 문제를 해결하기 위해, 드론이나 CCTV를 이용하여 도주하는 차량을 추적하는 형태로 활용되고 있다[2].

단일 객체 추적 연구는 Fully Convolutional Siamese Network (Siam FC)[1]를 기반으로 하며, 모델은 특징 추출기, 특징 융합 네트워크, 예측-헤드 네트워크 세 가지로 구성된다. 먼저 특징 추출기(feature extractor)는 추적 대상을 포함하는 템플릿(template) 영상과 추적 대상을 찾을 검색(search) 영상을 입력으로 받아 각 영상에 대한 특징맵을 추출한다. 이후 특징 융합 네트워크(feature fusion network)는 교차상관관계를 통해 두 특징맵 간 유사도를 계산한다. 이 때 가장 높은 유사도를 통해 예측-헤드 네트워크(prediction head network)는 검색 영상에서 추적 대상의 위치를 예측한다.

그러나 교차상관관계는 전역적인 정보를 충분히 활용할 수 없어 Local Optimum에 빠질 수 있고, 연산 과정에서 의미있는 정보가 소실될 수 있다는 한계점이 존재한다. 따라서 최근에는 교차상관관계를 Transformer[3] 구조로 대체하는 TransT[4] 모델이 대두되었다. TransT에서는 Transformer의 교차 어텐션(cross attention)을 사용해 두 특징맵 간의 유사도를 측정한다. 이를 통해 전역적인 정보를 활용할 수 있고, 유의미한 정보의 손실을 막는다. 이러한 구조는 교차상관관계를 사용했을 때보다 높은 성능을 보여주었다.

하지만 Siam FC, TransT와 같이 유사도를 이용해서 추적 대상을 파악하는 방식은 여러 객체들이 동시에 움직이는 비디오에서 낮은 정확도를 보였다. 여러 객체들이 동시에 움직이면 템플릿 영상 속에 다른 객체의 일부가 포함되어 노이즈로 작용하기 때문이다.

따라서 우리는 특징 추출기와 예측-헤드 네트워크에서 어텐션 모듈을 적용하는 방식을 제안한다. 특징 추출기에 적용된 어텐션 모듈은 템플릿 영상에 다른 객체의 일부가 포함되어 있더라도 추적 대상에만 집중하도록 한다. 예측-헤드 네트워크에 적용된 어텐션 모듈은 유사도가 계산된 이후 추적 대상의 위치와 크기를 정확하게 예측할 수 있도록 한다. 우리 모델을 GOT-10k[5] 데이터셋에 대해 학습한 결과, 기존 모델과 비교하여 더 좋은 성능을 달성하였다.

---

* 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(IITP-2017-0-00477, (SW 스타랩) IoT 환경을 위한 고성능 플래시 메모리 스토리지 기반 인메모리 분산 DBMS 연구개발)과 국토교통부의 스마트시티 혁신인재육성사업으로 지원을 받아 수행된 연구임.

† 교신 저자: sanghyun@yonsei.ac.kr

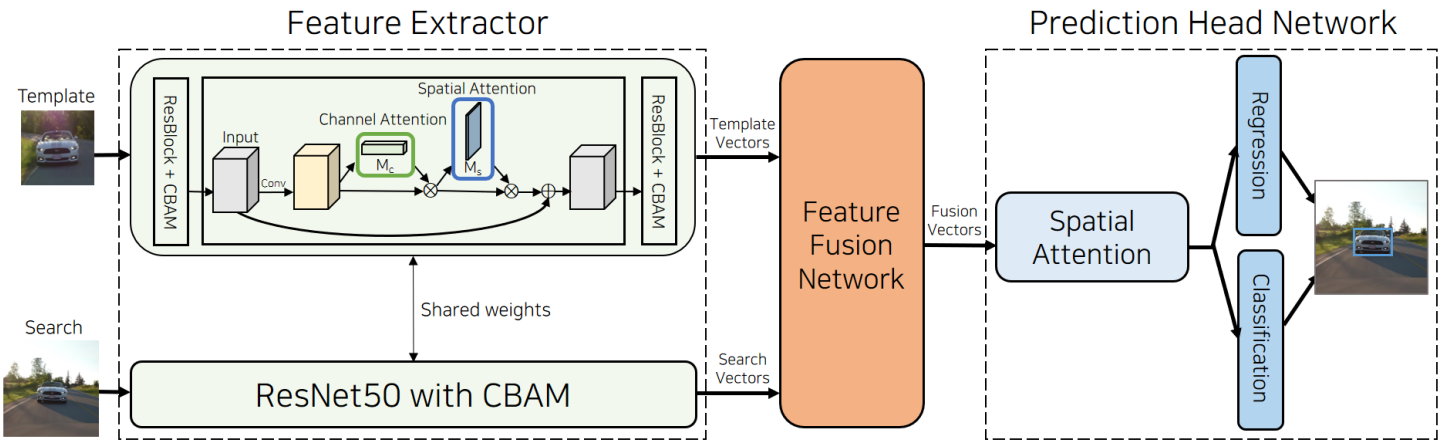

그림 1 특징 추출기와 예측-헤드 네트워크에 어텐션이 추가된 객체 추적 모델 구조

2. 본 론

우리가 제안한 모델의 구조는 그림1과 같으며, TransT의 구조를 따르고 있다. TransT는 Transformer 구조를 사용하며 크게 세 가지 네트워크로 구성되어 있다. 먼저 특징 추출기는 가중치를 공유하는 두 ResNet[6]이 템플릿 영상과 검색 영상의 특징을 각각 추출하고, 템플릿 벡터들(template vectors)과 검색 벡터들(search vectors)을 반환한다. 이후 특징 융합 네트워크는 Transformer의 교차 어텐션을 통해 템플릿 벡터들과 검색 벡터들의 유사도를 계산하고, 유사도가 저장된 융합 벡터들(fusion vectors)을 반환한다. 예측-헤드 네트워크는 융합 벡터들을 통해 검색 영상에서 추적 대상의 존재 여부와 위치 및 크기를 예측한다.

그러나 TransT처럼 유사도를 이용한 객체 추적 모델은 여러 객체들이 동시에 움직이는 비디오에서 문제점을 갖는다. 템플릿 영상에 다른 객체의 일부가 포함될 경우, 다른 객체에 대한 유사도가 커지게 된다. 따라서 다른 객체를 추적 대상으로 판단하거나, 여러 객체를 추적 대상으로 판단하는 문제가 발생할 수 있다.

이 문제를 해결하기 위해 특징 추출기와 예측-헤드 네트워크에 어텐션을 적용하였다. 먼저 특징 추출기는 ResNet의 모든 Residual Block에 CBAM[7] 모듈을 추가하였다. CBAM은 특징맵에서 어떤 채널이 중요한지를 결정하는 채널 어텐션(channel attention)과, 특징맵에서 어떤 위치가 중요한지를 결정하는 공간 어텐션(spatial attention)이 연속하는 구조를 지닌다. 이는 특징맵에서 중요한 부분을 더 강조하므로, 템플릿 영상에서 배경보다는 객체에 더 집중하고, 객체가 다수인 경우에는 추적 대상에 더 집중하게 한다.

또한 예측-헤드 네트워크에는 공간 어텐션 모듈을 추가하였다. 그림 2와 같이 융합 벡터들은 융합 특징맵(fusion feature map)으로 변환되고, 공간 어텐션을 통해 중요한 위치가 강조된다. 이후 평탄화(flatten)를 거쳐 어떤 융합 벡터가 중요한지를 알 수 있다. 따라서 추적 대상의 위치와 크기는 강조된 융합 벡터를 통해 더 정확하게 예측될 수 있다.

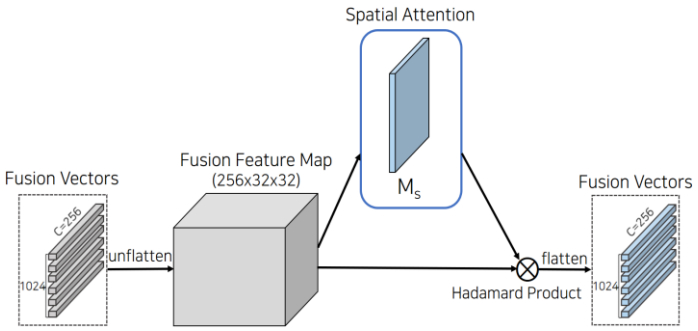

그림 2 예측-헤드 네트워크에서의 공간 어텐션 과정

3. 실 험

GOT-10k 데이터셋을 이용하여 모델을 학습시켰다. GOT-10k는 현실 세계에서 사람이나 동물 등이 움직이는 비디오가 약 만 개 저장된 데이터셋이다. 모델의 특징 추출기는 ImageNet[8] 데이터셋에 대해 사전 학습된 ResNet50 with CBAM을 사용하였다. Epoch, iteration, batch size는 각각 500, 1000, 32로 구성하였고, 학습률(learning rate)은 epoch 수가 250을 넘으면 10배 감소시켰다. 이 실험에서는 NVIDIA GeForce RTX 3090을 두 대 사용하였다.

4. 결 과

표 1 GOT-10k 데이터셋 실험 결과

Model	AO	$SR_{0.50}$	$SR_{0.75}$
TransT (Baseline)	65.3%	75.0%	58.7%
TransT (Prediction Head + Attention)	65.3%	75.4%	58.9%
TransT (Feature Extractor + Attention)	66.0%	76.1%	58.7%
TransT (All Network + Attention)	**66.0%**	**76.3%**	**58.9%**

표 1에서 Average Overlap(AO)은 추적 결과와 정답 사이의 중복 비율 평균이고, Success Rate(SR_T)는 중복 비율이 임계값(T)을 초과하는 프레임의 비율이다. 특징 추출기에 CBAM을 추가하고, 예측-헤드 네트워크에

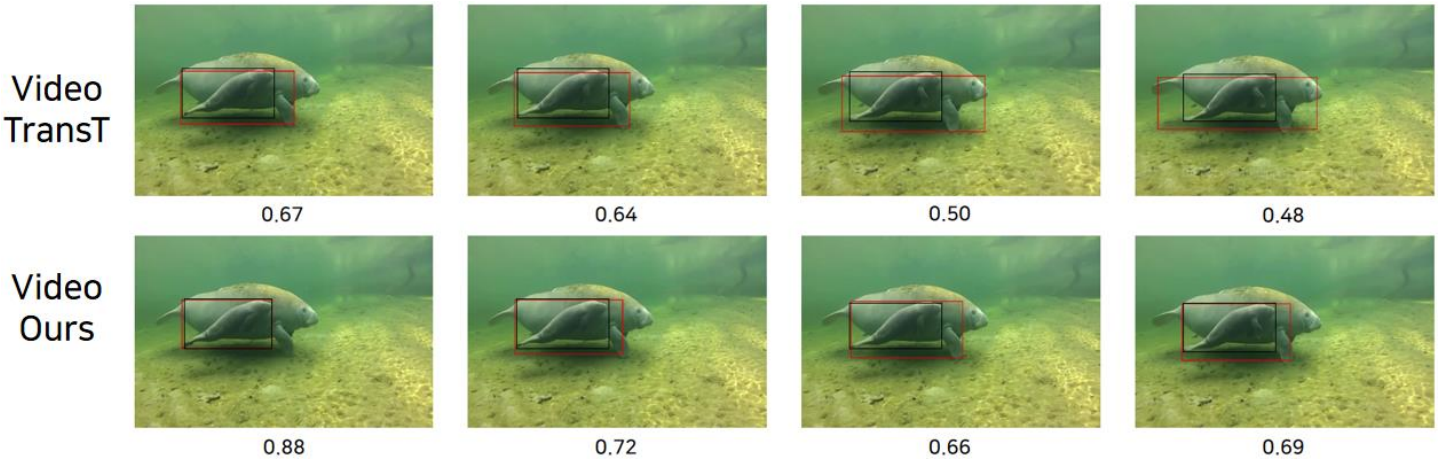

그림 3 각 모델의 프레임별 IoU 변화. IoU 값이 높을수록 추적 대상의 위치를 정확하게 예측한 것이다.

공간 어텐션 모듈을 추가한 결과, TransT 모델과 비교하여 AO, $SR_{0.50}$, $SR_{0.75}$가 각각 0.7%, 1.3%, 0.2% 상승하였다.

그림 3은 여러 객체들이 동시에 움직이는 비디오에서, 우리가 제안한 모델과 TransT를 비교한 결과이다. 빨간색 사각형은 각 모델이 예측한 추적 대상의 영역이고, 검은색 사각형은 정답에 해당하는 영역이다. 우리는 각 프레임별로 Intersection of Union(IoU)을 계산하였다. IoU는 두 영역 사이의 교집합의 넓이를 합집합의 넓이로 나눈 값이다. 우리의 모델은 위와 같은 비디오에서 추적 대상을 더 정확하게 예측하는 것을 확인할 수 있다.

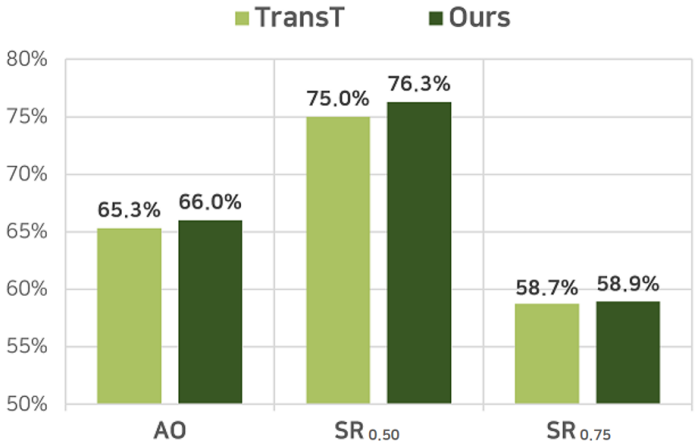

그림 4 객체 추적 성능 비교

5. 결 론

우리는 특징 추출기와 예측-헤드 네트워크에 어텐션을 추가하였으며, 기존의 TransT보다 AO와 $SR_{0.50}$이 각각 0.7%, 1.3% 상승하였다. 또한 여러 객체들이 동시에 움직이는 비디오에서 추적 대상에 대한 정확한 탐지를 보였다. 하지만 비슷한 객체가 갑자기 나타나거나, 유사한 객체들이 많은 비디오에서는 여전히 낮은 정확도를 보인다. 향후에는 이러한 문제점에 대한 추가적인 연구를 수행할 예정이다.

참 고 문 헌

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip H S Torr. “Fully-convolutional Siamese networks for object tracking”, European Conference on Computer Vision (ECCV), pp.850-856, 2016.

[2] SongWon Lim, SungMan Cho, and GooMan Park. “Integrated Video Analytics for Drone Captured Video”, Korean Institute of Broadcast and Media Engineers, pp.1-4, 2019.

[3] Vaswani, Ashish et al. “Attention is all you need”, Avances in neural information processing systems (NIPS), 2017.

[4] Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H. “Transformer tracking”, Computer Vision and Pattern Recognition (CVPR), pp.8126-8129, 2021.

[5] Lianghua Huang, Xin Zhao, Kaiqi Huang. “Got-10k: A large high-diversity benchmark for generic object tracking in the wild”, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2019.

[6] He, K.; Zhang, X.; Ren, S.; Sun, J. “Deep residual learning for image recognition”. Computer Vision and Pattern Recognition (CVPR), 2016.

[7] Sanghyun Woo, Jongchan Park, Joon Young Lee, In So Kweon. “Cbam: Convolutional block attention module”, European Conference on Computer Vision (ECCV), pp.4-6, 2018.

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy,Aditya Khosla, and Michael Bernstein. “ImageNet Large scale visual recognition challenge”. International Journal of Computer Vision (IJCV), 2015.