Siamese 네트워크 기반 영상 객체 추적 기술 동향

Trends on Visual Object Tracking Using Siamese Network

오지용 (J. Oh, jiyongoh@etri.re.kr) 로봇IT융합연구실 선임연구원
이지은 (J. Lee, jieun.lee@etri.re.kr) 로봇IT융합연구실 박사후연수연구원

ABSTRACT

Visual object tracking can be utilized in various applications and has attracted considerable attention in the field of computer vision. Visual object tracking technology is classified in various ways based on the number of tracking objects and the methodologies employed for tracking algorithms. This report briefly introduces the visual object tracking challenge that contributes to the development of single object tracking technology. Furthermore, we review ten Siamese network-based algorithms that have attracted attention, owing to their high tracking speed (despite the use of neural networks). In addition, we discuss the prospects of the Siamese network-based object tracking algorithms.

KEYWORDS Siamese 네트워크, 심층 신경망, 영상 객체 추적

Ⅰ. 서론

영상 객체 추적은 사용자로부터 동영상의 특정 프레임에 나타난 임의의 객체를 포함하는 영역을 입력받아 이후의 동영상에서 선택된 객체를 자동으로 추적하는 기술을 의미한다[1]. 영상 객체 추적은 영상 보안, 로보틱스, 비디오 분석, 자율주행 자동차 등과 같이 동영상을 이용하는 다양한 응용 분야에서 활용될 수 있기 때문에 오래전부터 많은 연구가 수행되고 있다. 초기의 영상 객체 추적 기술은 얼굴이나 사람과 같이 특정한 단일 대상을 추적하는 것이 목표였다. 하지만 현재에는 사용자가 지정하는 임의의 객체를 추적할 수 있는 알고리즘이 개발되고 있으며, 추적 대상뿐만 아니라 추적하는 대상의 수도 확대되어 다중 객체들을 동시에 추적하기 위한 알고리즘에 대한 연구도 활발히 진행되고 있다. 특히 최근에는 컴퓨터 비전의 다른 분야와 마찬가지로 영상 객체 추적에도 심층 신경망을 활용하는 연구가 주를 이루고 있고, 그 결과로 객체 추적 기술도 빠르게 발전되고 있다.

영상 객체 추적 기술은 추적 대상의 수(단일, 다중) 및 추적을 위한 방법론에 따라 다양하게 분류된

* DOI: https://doi.org/10.22648/ETRI.2022.J.370108
* 본 연구 논문은 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음[21ZD1130, 지능제어기반 스마트 기계 및 로봇 기술 개발].

다. 본고에서는 심층 신경망을 활용하는 객체 추적 기술 중 주로 단일 객체 추적을 위해 활용되고 있는 Siamese 네트워크를 사용하는 알고리즘들을 소개하고자 한다. 심층 신경망을 분류기로 이용하는 온라인 추적기들과 비교해 Siamese 네트워크 기반 추적기들은 뛰어난 정확도 및 강인성과 함께 추적 속도 측면에서 매우 우수하다. 이런 이유로 로봇과 같이 실시간성이 요구되는 환경에서 특히 유용하게 활용될 수 있다. 독자들은 본고를 통해 Siamese 네트워크 기반 객체 추적 기술의 발전 과정을 파악할 수 있을 것으로 예상된다. Ⅱ장에서는 Siamese 네트워크 기반 추적 알고리즘들의 소개에 앞서 매년 영상 객체 추적 기술들의 도전 과제를 제시하여 기술 발전에 이바지하고 있는 VOT(Visual Object Tracking) 대회에 대해 간략히 살펴본다. Ⅲ장에서는 Siamese 네트워크를 영상 객체 추적에 적용한 효시적인 연구를 포함하여 CVPR, ICCV, ECCV와 같이 저명한 컴퓨터 비전 학회에서 발표된 10개의 알고리즘을 살펴본다. 마지막으로 Ⅳ장에서는 Siamese 네트워크 기반 객체 추적 기술의 발전 과정을 정리하고 향후 전망을 소개한다.

Ⅱ. Visual Object Tracking 대회

ICCV 2013 워크숍에서 처음 시작된 VOT 대회[2-10]는 영상 객체 추적 알고리즘들의 우수성을 겨루는 행사로 매년 ICCV나 ECCV에서 열리는 워크숍을 통해 개최되고 있다. VOT 대회는 객체 추적 알고리즘들을 객관적으로 평가할 수 있는 토대를 제공한다는 점에서 큰 의미가 있다. VOT 대회에서 공개하는 데이터셋 이외에도 영상 객체 추적 알고리즘 개발에 다른 데이터셋들이 활용되기도 한다. 예를 들면 VOT와 비슷한 시기에 공개된 OTB[11]도 여전히 많은 연구자가 활용하고 있고, 최근에 공개된 TrackingNet[12], LaSOT[13], GOT-10k[14]와 같은 데이터셋은 VOT에서 제공하는 데이터셋보다 규모도 크고 더욱 다양한 동영상을 포함하고 있다. 하지만 VOT 대회는 체계적으로 정리된 데이터셋과 성능 평가 방법을 공개하는 데에 그치지 않고, 매년 데이터셋을 도전적으로 업데이트함과 동시에 참가자들을 모집하고 순위를 공개함으로써 객체 추적 기술 연구자들에게 다른 형태의 동기부여를 제공하고 있다.

1. 역사 및 변화

VOT 대회는 해가 지남에 따라 눈에 띄는 변화들을 거쳐왔다(그림 1 참조). 2013년에는 추적 대상이 영상에서 가려지거나 사라지지 않는 시나리오

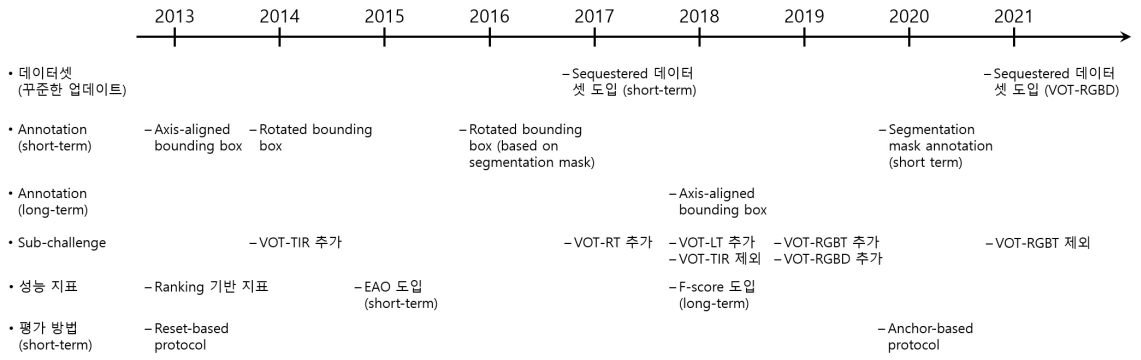


그림 1 VOT 대회의 연도별 주요 특징

에서 객체를 추적하는 문제로 시작되었지만, 2017년부터는 real-time 제약이 존재하는 환경에서 추적 기술의 성능을 평가하는 VOT-RT 부문이 추가되었고, 2018년부터 추적 대상이 가려지거나 영상에서 사라지는 조건에서 추적 알고리즘들의 우위를 가늠하기 위한 long-term 부분이 추가되었으며, 2019년부터는 RGBD 센서 기반 추적기의 성능을 평가하기 위한 VOT-RGBD 부문도 추가되었다. 한편 2015년부터 2017년까지는 열영상(Thermal Imagery)에 적용하는 객체 추적 알고리즘을 위한 별도의 VOT-TIR 대회가 개최되었는데, 이 대회는 2019년과 2020년에 VOT-RGBT 부문으로 VOT 대회에 포함되었다. 이러한 과정들을 거쳐 9회를 맞이한 2021년 VOT 대회[10]는 VOT-ST2021, VOT-RT2021, VOT-LT2021, VOT-RGBD2021과 같이 총 네 부문으로 나뉘어 개최되었다.

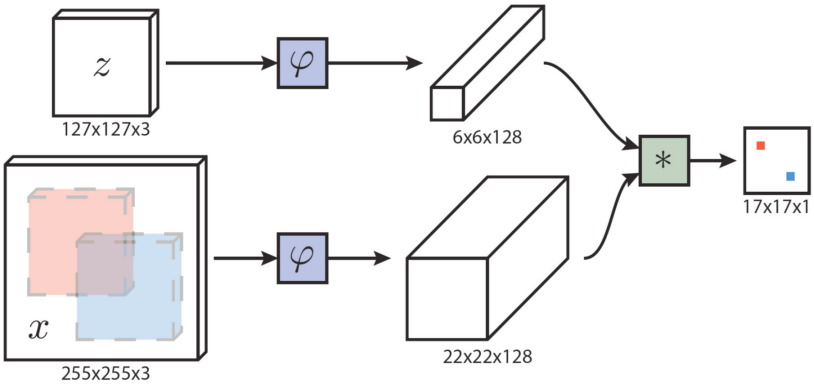
2. 객관적 평가를 위한 성능 측정

영상 추적 알고리즘의 성능을 측정하는 지표는 다양하다. VOT 대회에서는 추적 알고리즘의 객관적 평가를 위해 각 부문에 적합한 각각의 지표를 제시하고 있다. Short-term 시나리오에서는 추적 알고리즘을 평가하기 위해 기본적으로 정확성(Accuracy)과 강인성(Robustness)을 측정하는데, 정확성은 각 프레임마다 추적 알고리즘의 결과가 ground truth와 얼마나 많이 겹치는지를 의미하고, 강인성은 추적 알고리즘이 얼마나 많이 추적에 실패하지 않는지를 의미한다. 이를 바탕으로 VOT 대회의 short-term 부문은 정확성과 강인성을 함께 평가하기 위한 하나의 지표인 EAO(Expected Average Overlap)[4]를 통해 참가팀의 순위를 결정한다. Long-term 시나리오에서는 추적 대상의 사라짐 혹은 가려짐으로 인해 기본적으로 정밀도(Precision)와 재현율(Recall)을 성능 측정 지표로 활용하며 정밀도와 재현율을 하나의 지표로 결합한 F-Score를 통해 최종 순위를 결정한다.

III. 주요 알고리즘

1. SiamFC

ECCV 2016년에 발표된 SiamFC 알고리즘[1]은 영상 객체 추적을 위해 Siamese 네트워크를 활용한 효시적인 연구로 SiamFC 추적기의 구조는 그림 2와 같다. 그림 2에서 z는 사용자가 입력한 추적 대상을 포함하는 예시 영상(Exemplar Image)을 의미하



출처 Reprinted with permission from [1].

그림 2 영상 객체 추적을 위한 SiamFC의 구조

고, x는 추적기에 입력되어 추적 대상의 영역을 추론해야 하는 검색 영상(Search Image)을 의미한다. 그림 2에서 보인 바와 같이 예시 영상과 검색 영상은 같은 함수를 통과하는데, 이 함수는 동일한 구조의 CNN으로 구현되기 때문에 Siamese 네트워크 구조라 지칭된다. 동일한 네트워크를 통과한 예시 영상과 검색 영상의 특징 텐서(Tensor)들은 유사도 맵을 계산하기 위해 교차 상관(Cross Correlation)의 입력으로 사용되며, 계산된 유사도 맵의 각 성분은 검색 영상 내부에 대한 예시 영상과의 유사도에 해당한다. 예를 들면, 그림 2에서의 검색 영상 내부의 빨간색(혹은 파란색) 영역과 예시 영상 사이의 유사도는 유사도 맵의 빨간색(혹은 파란색) 성분에 해당한다. 마지막으로 검색 영상에 대한 추적 결과는 가장 높은 유사도를 갖는 검색 영상 내부 영역으로 결정된다.

SiamFC 추적 알고리즘의 강점은 추적 속도다. TLD[15], KCF[16]와 같이 SiamFC가 발표되기 전 영상 객체 추적기들은 추적 대상과 배경을 분류하기 위한 분류기를 포함하는데, 추적이 진행되는 프레임마다 그 분류기의 학습이 수행된다. 이러한 온라인 학습은 추적 속도가 느려지는 문제를 지니고 있다. MDNet[17]과 같이 심층 신경망을 활용하는 영상 객체 추적 알고리즘들도 CNN 기반의 분류기를 통해 추적기의 정확도를 크게 향상시켰지만, 온라인 학습 방식으로 인해 추적 속도가 느려 실시간 성이 요구되는 응용 분야에는 활용할 수 없는 문제를 안고 있었다. SiamFC는 이러한 추적 속도의 문제를 해결하기 위해 오프라인 학습을 채택하였다. SiamFC를 학습하기 위해 많은 양의 예시-검색 영상의 데이터 쌍을 준비해야 하는 어려움이 있지만 ImageNet Video[18]와 같은 대규모 데이터셋을 활용하여 학습시킨 결과, 표 1에서 확인할 수 있듯이 SiamFC의 추적 정확도는 온라인 방식의 추적기에 근접하면서도 압도적인 추적 속도(80fps 이상)를 보이는 것으로 발표되었다.

표 1 VOT2015 벤치마크[4]를 통한 추적기들의 성능 비교

Tracker	정확도	속도(fps)
MDNet	0.5620	1
EBT	0.4481	5
DeepSRDCF	0.5350	<1
SiamFC-3s	0.5335	86
SiamFC	0.5240	58

출처 Reproduced from [1].

2. Siamese-RPN

SiamFC 추적 알고리즘은 영상 객체 추적을 위해 오프라인으로 학습된 Siamese 네트워크를 활용하여 추적 속도를 크게 향상시켰지만, 정확성 측면에서는 최고 성능(State-of-the-Art)의 온라인 추적기와 격차를 보이는 것이 사실이었다. 또한 SiamFC는 추적 대상의 크기 변화를 위해 3 또는 5레벨을 갖는 영상 피라미드 기법을 활용하였는데, 해당 방법은 bounding box 형태로 수행되는 추적 알고리즘의 정확도 측면에서 불리하다. CVPR 2018에서 발표된 SiamRPN(Siamese-RPN)[19] 알고리즘은 SiamFC 구조에 영상 검출 문제에서 표준처럼 사용되던 RPN(Region Proposal Network)을 채택하여 추적 대상의 위치를 결정하기 위해 bounding box의 regression을 수행한다. 이로 인해 추적 대상의 크기를 기존보다 정확하게 추정할 수 있게 됨과 동시에 영상 피라미드의 채택으로 인한 반복적 계산을 회피함으로써 추정 속도도 향상시킬 수 있게 되었다. 실제로 참고문헌 [19]에 의하면 VOT2015 대회 및 VOT2016 대회에서 우수한 성능을 입증하였던 state-of-the-art 알고리즘들보다 더 높은 정확도를 보이며, VOT2017 대회 real-time 부문에서는 1위

를 차지하였고, 동시에 160fps의 속도로 동작하는 것으로 보고되었다.

3. DaSiameseRPN

ECCV 2018에서 발표된 DaSiamRPN(DaSiameseRPN)[20]을 제안한 연구진은 Siamese 네트워크를 활용하는 추적 알고리즘들이 추적에 실패하는 drift 현상에 주목하였다. 이런 현상에 대한 분석을 통해 학습 데이터를 수집하는 과정에서 무의미한 배경(Non-semantic Background)과 함께 잠재적으로 추적을 방해할 수 있는 다른 객체(Semantic Distractor)도 학습 데이터의 negative 데이터 쌍에 포함시키는 방법으로 추적기의 정확도를 향상시켰다. DaSiameseRPN 알고리즘은 학습 데이터를 수집하는 방법 이외에도 정확도를 높이기 위해 임의의 프레임의 추적 결과를 다음 프레임의 추적에 이용하는 방법도 포함하고 있다. 해당 방법에서는 추적 대상으로 선택될 확률이 높은 영역들 중에 실제 추적 대상으로 선택되지 않은 영역들을 distractor로 정의한다. 그리고 다음 프레임들에서의 추적 과정에서 distractor들과 유사도가 높은 영역이 추적 대상으로 선택될 확률을 감소시킨다. 또한, 추적 대상이 다른 물체나 대상에 의해 가려지거나 영상 밖으로 사라지는 경우가 포함되는 long-term 추적 문제를 위해 간단하면서도 효과적인 반복적 local-to-global 검색 전략도 함께 제안하였다.

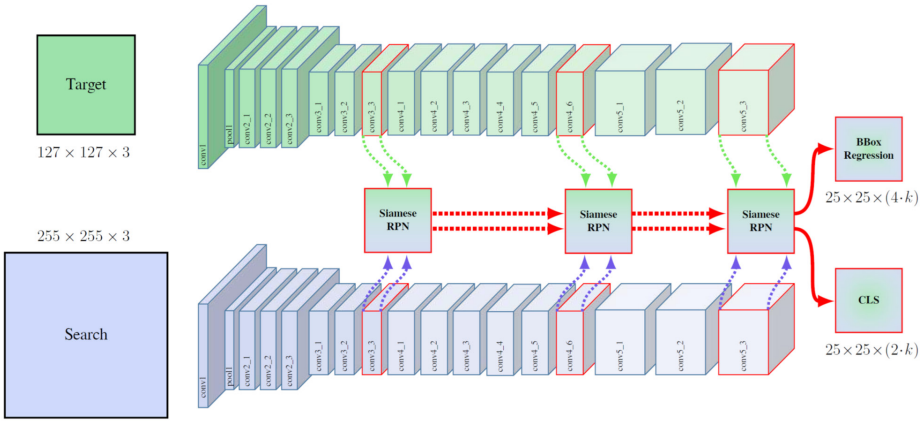
위와 같은 세 가지 방법으로 인해 추적 정확도가 향상된 DaSiameseRPN 알고리즘은 VOT2018 대회 long-term 부문에서 2위를 차지할 정도로 높은 정확도를 자랑함과 동시에 추적 속도는 short-term 시나리오에서는 160fps, long-term 시나리오에서는 110fps인 것으로 보고되었다.

4. SiamRPN++

컴퓨터 비전 분야에서는 AlexNet[21]보다 ResNet[22]과 같이 CNN을 구성하는 층(Layer) 수를 깊이 쌓을수록 네트워크의 성능이 향상되는 것으로 알려져 있다. 그럼에도 불구하고 SiamRPN 및 DaSiameseRPN의 경우 backbone으로 ResNet이 아닌 AlexNet을 사용하였다. 그 이유는 AlexNet을 ResNet으로 단순히 변경하게 되면 오히려 추적 정확도가 떨어지는 현상 때문이었다. CVPR 2019에서는 이 문제를 해결하고 Siamese 네트워크 기반 추적기의 정확도를 향상시킨 두 개의 알고리즘들이 발표되었는데, SiamRPN++[23]은 그 중 하나이다.

SiamRPN++의 연구진은 Siamese 계열의 추적기에 심층 신경망을 적용하는 것이 어려운 원인으로 심층 신경망을 구성하는 CNN 내부의 padding 연산과 RPN이 담당하는 분류(Classification) 문제와 회귀(Regression) 문제의 비대칭성을 꼽았다. 두 가지 원인 중 padding에 의해 정확성이 감소하는 문제를 해결하기 위한 방법으로 학습 데이터를 수집하는 과정에서 검색 영상(Search Image)에서 추적 대상들의 위치가 uniform 분포를 갖도록 만드는 spatial aware sampling 전략을 제안하였다. 실제로 연구진은 해당 전략에 의해 수집된 학습 데이터를 사용하여 ResNet-50을 backbone으로 채택한 SiamRPN 추적기를 학습시킨 결과 AlexNet보다 높은 정확도를 얻었다.

한편, SiamRPN에서 backbone 네트워크의 두 출력인 예시 영상과 검색 영상으로부터 계산된 두 특징 텐서들의 교차 상관분석 연산이 RPN에서 정의된 anchor 수만큼 반복된다. 이러한 반복으로 인해 RPN 내부의 분류 문제와 회귀 문제의 비대칭성이 발생하게 된다. 하지만 SiamRPN++에서는 그와 같은 up-channel 교차 상관을 깊이별(Depth-wise)



출처 Reprinted with permission from [23].

그림 3 SiamRPN++의 구조

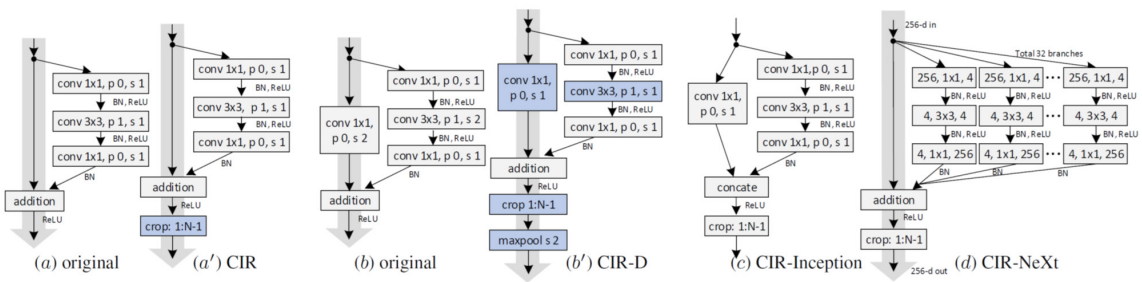
교차 상관으로 대체함으로써 RPN 내부의 비대칭성 문제를 피할 수 있게 되었다.

그림 3은 ResNet-50을 활용한 SiamRPN++의 구조를 보여준다. 앞서 언급한 두 가지 제안 이외에도 SiamRPN++에서는 예시 영상과 검색 영상이 backbone 네트워크의 다양한 계층의 영상 정보를 수집하기 위해 ResNet-50의 세 가지 블록의 특징 텐서들이 함께 RPN 단계로 전달되는 것도 확인할 수 있다.

5. SiamDW

CVPR 2019에서 함께 발표된 SiamRPN++과 SiamDW[24]은 Siamese 네트워크 기반 추적기에 ResNet과 같은 심층 신경망을 적용하기 위한 연구라는 점에서 두 연구의 출발점이 같다. 하지만 흥미롭게도 SiamDW는 같은 문제를 해결하기 위해 SiamRPN++과 다른 방법을 제안하였다.

SiamDW의 연구진도 마찬가지로 Siamese 계열의 추적기들에 심층 신경망을 적용하지 못하는 원인으로 CNN 내부의 padding 연산에 주목하였다. 하지만 SiamRPN++에서는 학습 데이터를 수집하는 과정을 변경한 것과 달리 SiamDW는 그림 4와 같이 cropping 연산을 residual 유닛 내부에 포함시키는 CIR(Cropping-Inside Residual) 유닛을 개발하였고, 그 개념을 CIR-D 유닛, CIR-Inception 및 CIR-



출처 Reprinted with permission from [24].

그림 4 다양한 cropping-inside residual 유닛들

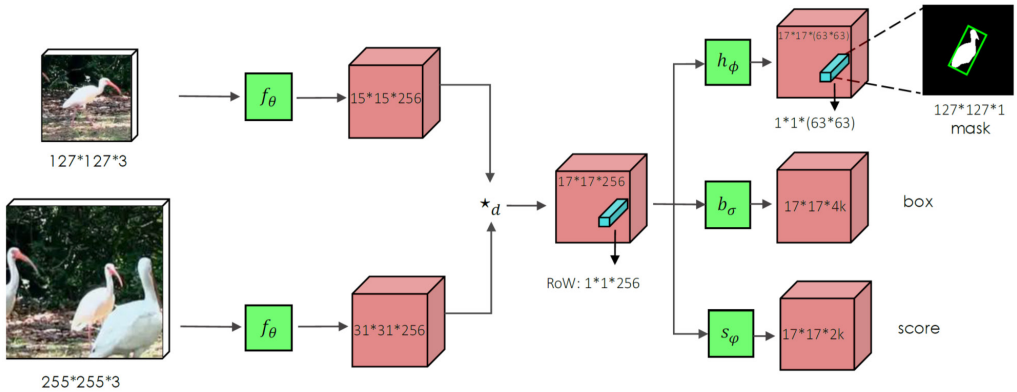
NeXt 유닛으로 더욱 확장하여 ResNet뿐만 아니라 Inception[25]과 ResNeXt[26]도 Siamese 네트워크 계열의 추적기에 활용하는 방법을 제시하였다. 이와 같은 방법으로 심층 신경망을 backbone으로 채택하게 된 SiamDW를 활용한 추적기들은 VOT2019-RGBD 1위와 VOT2019-RGBT 2위를 차지하여 그 우수성을 입증하였다.

6. SiamMask

CVPR 2019에서 발표된 SiamMask 알고리즘[27]은 Siamese 네트워크 기반 추적 기술을 semi-supervised VOS(Video Object Segmentation) 문제에 적용한 효시적인 연구라고 할 수 있다. 그림 5와 같이 SiamMask에서는 예시 영상과 검색 영상을 backbone에 해당하는 동일한 CNN을 통과시켜 얻어진 두 특징 텐서들의 깊이별 교차 상관이 수행되고, 이를 통해 다채널 응답 맵(Response Map)이 계산된다. 이 다채널 응답 맵으로부터 SiamFC 혹은 SiamRPN와 같은 출력과 함께 출력 영상에서의 추적 대상에 해당하는 세분화 마스크(Segmentation Mask) 정보가 출력된다. SiamMask는 간단한 구조로 인해 55fps라는 빠른 속도로 동작한다고 보고되었다. 기존의 VOS의 속도를 개선하기 위해 제안된 알고리즘들조차 real-time으로 동작하기 어려웠다는 점을 고려한다면 SiamMask 알고리즘으로 인해 VOS의 속도가 획기적으로 향상되었다고 판단할 수 있다.

7. UpdateNet

Siamese 계열 추적기에서 동영상의 첫 프레임에서 사용자에 의해 결정된 예시 영상은 이후 프레임에 대한 추적을 수행하는 동안 변화되지 않는다. 하지만 이로 인해 프레임들 사이에서 추적 대상의 생김새가 크게 변하게 되면 추적이 실패하는 경우가 발생한다. 이런 문제에 대응하기 위해 예시 영상을 선형적으로 업데이트하는 전략을 채택할 수 있다. 앞서 소개된 DaSiamRPN이 선형적 업데이트를 채택한 알고리즘이라고 할 수 있다. 하지만 간단한 선형 업데이트로는 추적 대상의 잠재적인 모든 변화에 적절히 대응하는 것이 쉽지 않다. 이러한 문제를 해결하기 위해 UpdateNet[28]의 연구진은 예시 영상의 업데이트를 데이터로부터 학습하는 방법을 제안하였다. UpdateNet은 사용자가 첫



출처 Reprinted with permission from [27].

그림 5 SiamMask의 구조

프레임에 대해 정의한 예시 영상과 이후 프레임들에 대한 추적 결과를 바탕으로 UpdateNet이라 명명된 CNN을 활용하여 예시 영상을 업데이트한다. 그런 다음 업데이트된 예시 영상이 다음 프레임에 대한 추적에 활용된다.

8. SiamAttn

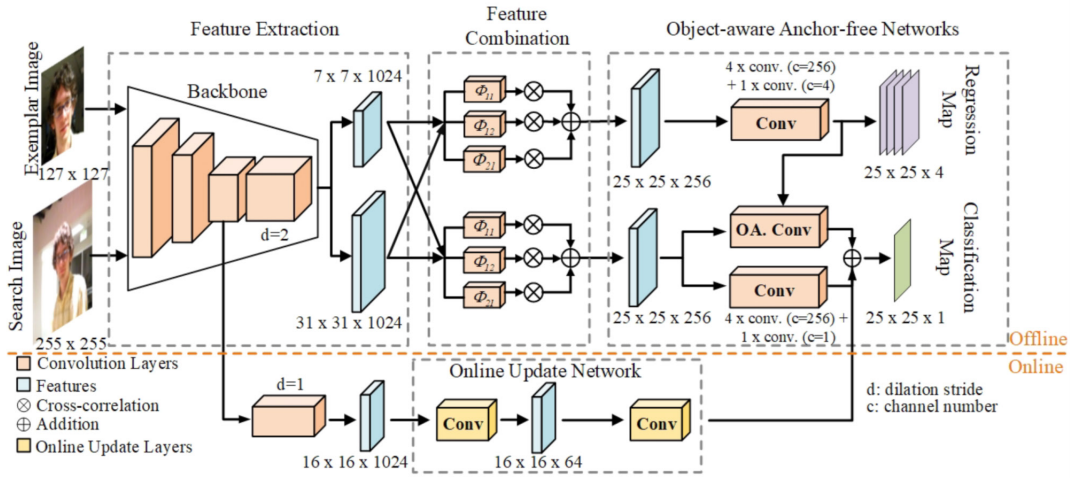
CVPR 2020에서 발표된 SiamAttn 알고리즘[29]도 기존 Siamese 네트워크 기반 추적 알고리즘들이 추적 대상의 변화를 학습하는 온라인 추적기와 달리 오프라인 학습에만 의존하는 문제를 해결하기 위해 제안된 방법이다. SiamAttn 연구진은 당시 컴퓨터 비전의 다른 분야에서 각광받던 attention 기법을 SiamRPN++ 구조에 적용하여 추적 과정에서 발생하는 추적 대상의 변화를 Siamese 네트워크에 반영하였다. SiamAttn 알고리즘의 핵심은 DSA(Deformable Siamese Attention) 모듈이다. DSA 모듈에는 Siamese 네트워크의 예시 영상과 검색 영상들로부터 계산된 각각의 특징 텐서들은 spatial과 channel 차원으로 self-attention이 수행되고 예시 특징 텐서와 검색 특징 텐서들 사이의 cross-attention도 함께 수행된다. SiamAttn 알고리즘은 VOT2016 데이터셋과 VOT2018 데이터셋을 이용한 실험을 통해 SiamRPN++[23]과 SiamMask[27]보다 높은 EAO 값을 보이는 것으로 확인되었으며, 그럼에도 불구하고 30fps 이상의 속도로 동작한다.

9. SiamBAN

SiamRPN이 발표된 이후로 Siamese 네트워크 기반 추적 알고리즘들은 거의 RPN 구조를 활용하였다. 하지만 객체 검출 분야에서 RPN의 중추적인 역할을 하는 anchor의 몇몇 문제점들이 보고되었고, 해당 문제점들을 해결하기 위해 anchor를 사용하지 않는 검출 알고리즘들이 발표되었다. 그 중 ICCV 2019에서 발표된 FCOS[30]는 검출 결과를 RPN 내부에 미리 정의된 anchor 기준으로 계산하지 않고 네트워크 출력 텐서의 각각의 요소를 기준으로 bounding box의 regression과 classification이 수행된다. FCOS 알고리즘은 이와 같이 간단한 방법을 통해 검출기의 구조를 간단하게 만들면서 동시에 검출기의 성능을 향상시켜 주목을 받았다. SiamBAN 알고리즘[31]은 SiamRPN++ 추적 알고리즘에 FCOS에서 제안한 기법을 적용한 방법으로 SiamBAN 연구진은 해당 방법을 box adaptive head라고 이름지었다. SiamBAN 알고리즘은 SiamRPN++ 구조에 간단한 아이디어만 추가했음에도 불구하고 VOT2019 데이터셋을 이용한 실험을 통해 EAO 측면에서 SiamRPN++, SiamMask, SiamDW보다 우수한 것으로 확인되었다.

10. Ocean

ECCV 2020에서 발표된 Ocean 알고리즘[32]은 anchor 기반 Siamese 추적기들이 추적 대상을 놓치기 시작하면 이후 프레임에서 추적 성능이 더욱 저하되는 현상에 주목하였다. 이 문제를 해결하기 위해 Ocean에서도 SiamBAN에서 제시했던 box adaptive head와 유사한 방법을 채택하였다. 이와 함께 Ocean 연구진은 객체 인지(Object-Aware) 분류 네트워크를 제안하였다(그림 6 참조). 객체 인지 분류 네트워크는 추적 대상 영역을 분류하는 데 적합한 특징을 계산하기 위해 convolution이 이루어지는 영역을 학습을 통해 결정하며 이러한 과정은 deformable convolution[33] 기법을 통해 구현되었다. 한편, SiamRPN++에서는 backbone 네트워크로부터 다양한 계층의 정보를 활용하기 위해 세 가



출처 Reprinted with permission from [32].

그림 6 Ocean의 네트워크 구조

지 블록에서의 특징들을 함께 사용한 반면, Ocean에서는 추적 대상의 크기 변화에 적절하게 대응하기 위해 세 가지 조합의 dilated convolution[34]을 채택하였다. 이러한 방법들로 구성된 offline Ocean 알고리즘은 VOT2018 데이터셋과 VOT2019 데이터셋을 이용한 비교 실험을 통해 SiamMask 및 Siam-RPN++보다 높은 EAO 값을 확인할 수 있었으며, online 모듈을 추가하여 추적 정확도가 더욱 향상되었다고 보고되었다.

Ⅳ. 결론

다양한 응용 분야에서 활용이 가능한 객체 추적 기술은 심층 신경망 기술과 함께 비약적으로 발전하고 있다. 특히 본고에서 살펴본 Siamese 네트워크 기반 추적 알고리즘들은 심층 신경망을 활용함에도 불구하고 빠른 속도로 동작하기 때문에 실시간성이 요구되거나 계산 능력(Computation Power)이 제한적인 응용 분야에 적합하다.

Siamese 네트워크 기반 객체 추적 알고리즘들은 기존 알고리즘의 문제점을 찾아내고 그 문제를 해결하기 위해 컴퓨터 비전의 다른 분야에서 제안되는 기법들을 적용하는 방식을 통해 빠른 속도로 발전하고 있다. 예를 들면, 단일 객체 추적 알고리즘의 또 다른 방법론인 심층 신경망 기반 온라인 추적 알고리즘들에서 제안된 방법들이 Siamese 네트워크 기반 추적 알고리즘에 활용되기도 한다. 최근에는 자연어 처리 분야에서 우수한 성능으로 각광을 받았던 transformer 기술을 단일 객체 추적 알고리즘에 활용하는 연구[35]가 발표되었으며, 네트워크의 구조를 학습을 통해 찾아내는 neural architecture search 기법을 활용하여 추적 정확도를 향상시킴과 동시에 추적 속도를 10배 이상 향상시킨 연구[36]도 발표되었다. 앞으로 수행될 후속 연구들에서도 추적 알고리즘을 더욱 발전시키기 위해 인공지능 관련 최신 기법들이 적극 활용될 것으로 전망된다. 이런 과정을 통해 기술적 완성도는 점차 높아져 일상생활의 다양한 분야에서 추적 기술을 더욱 자주 접하게 될 것이다.

약어 정리

CNN	Convolutional Neural Network
CVPR	IEEE/CVF Conference on Computer Vision and Pattern Recognition
ECCV	European Conference on Computer Vision
ICCV	IEEE/CVF International Conference on Computer Vision

참고문헌

[1] L. Bertinetto et al., "Fully-convolutional siamses networks for object tracking," in Proc. Eur. Conf. Comput. Vis. (ECCV), (Amsterdam, Netherlands), Oct. 2016, pp. 850-865.

[2] M. Kristan et al., "The visual object tracking VOT2013 challenge results," in Proc. IEEE Int. Conf. Comput. Vis. Workshops, (Sydney, Australia), Dec. 2013, pp. 93-111.

[3] M. Kristan et al., "The visual object tracking VOT2014 challenge results," in Proc. Eur. Conf. Comput. Vis. (ECCV), (Zurich, Switzerland), Sept. 2014, pp. 191-217.

[4] M. Kristan et al., "The visual object tracking VOT2015 challenge results," in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV), (Santiago, Chile), Dec. 2015, pp. 1-23.

[5] M. Kristan et al., "The visual object tracking VOT2016 challenge results," in Proc. Eur. Conf. Comput. Vis. (ECCV), (Amsterdam, Netherlands), Oct. 2016, pp. 777-823.

[6] M. Kristan et al., "The visual object tracking VOT2017 challenge results," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), (Venice, Italy), Oct. 2017, pp. 1949-1972.

[7] M. Kristan et al., "The sixth visual object tracking VOT2018 challenge results," in Proc. Eur. Conf. Comput. Vis. (ECCV), (Munich, Germany), Sept. 2018.

[8] M. Kristan et al., "The seventh visual object tracking VOT2019 challenge results," in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), (Seoul, Republic of Korea), Oct. 2019, pp. 2206-2241.

[9] M. Kristan et al., "The eighth visual object tracking VOT2020 challenge results," in Proc. Eur. Conf. Comput. Vis. (ECCV), (Glasgow, UK), Aug. 2020, pp. 547-601.

[10] M. Kristan et al., "The ninth visual object tracking VOT2021 challenge results," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 2711-2738.

[11] Y. Wu et al., "Object tracking benchmark," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, 2015, pp. 1834-1848.

[12] M. Müller et al., "TrackingNet: A large-scale dataset and benchmark for object tracking in the wild," in Proc. Eur. Conf. Comput. Vis. (ECCV), (Munich, Germany), Sept. 2018, pp. 300-317.

[13] H. Fan et al., "LaSOT: A high-quality benchmark for large-scale single object tracking," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), (Long Beach, CA, USA), June 2019, pp. 5374-5383.

[14] L. Huang et al., "GOT-10k: A large high-diversity benchmark for generic object tracking in the wild," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 5, 2021, pp. 1562-1577.

[15] Z. Kalal et al., "Tracking-learing-detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 7, 2011, pp. 1409-1422.

[16] J.F. Henriques et al., "High-speed tracking with kernelized correlation filters," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, 2015, pp. 583-596.

[17] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), (Las Vegas, NV, USA), June 2016, pp. 4293-4302.

[18] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," Int. J. Comput. Vis., vol. 115, 2015, pp. 211-252.

[19] B. Li et al., "High performance visual tracking with siamese region proposal network," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), (Salt Lake Sity, UT, USA), June 2018, pp. 8971-8980.

[20] Z. Zhu et al., "Distractor-aware siamese networks for visual object tracking," in Proc. Eur. Conf. Comput. Vis. (ECCV), (Munich, Germany), Sept. 2018.

[21] A. Krizhevsky et al., "ImageNet classification with deep convolutional neural networks," in Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS), (Lake Taho, NV, USA), Dec. 2012, pp. 1097-1105.

[22] K. He et al., "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), (Las Vegas, NV, USA), June 2016, pp. 770-778.

[23] B. Li et al., "SiamRPN++: Evolution of siamese visual tracking with very deep networks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), (Long Beach, CA, USA), June 2019, pp. 4282-4291.

[24] Z. Zhang et al., "Deeper and wider siamese networks for real-time visual tracking," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), (Long Beach, CA, USA), June 2019, pp. 4591-4600.

[25] C. Szegedy et al., "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), (Boston, MA, USA), June 2015, pp. 1-9.

[26] S. Xie et al., "Aggregated residual transformations for deep neural networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), (Honolulu, HI, USA), July 2017, pp. 1492-1500.

[27] Q. Wang et al., "Fast online object tracking and segmentation: A unifying approach," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), (Long Beach, CA, USA), June 2019, pp. 1328-1338.

[28] L. Zhang et al., "Learning the model update for siamese trackers," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), (Seoul, Republic of Korea), Oct. 2019, pp. 4010-4019.

[29] Y. Yu et al., "Deformable siamese attention networks for visual object tracking," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), June 2020, pp. 6728-6737.

[30] Z. Tian et al., "FCOS: Fully convolutional one-stage object detection," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), (Seoul, Republic of Korea), Oct. 2019, pp. 9627-9636.

[31] Z. Chen et al., "Siamese box adaptive network by visual tracking," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), June 2020, pp. 6668-6677.

[32] Z. Zhang et al., "Ocean: Object-aware anchor-free tracking," in Proc. Eur. Conf. Comput. Vis. (ECCV), (Glasgow, UK), Aug. 2020, pp. 771-787.

[33] J. Dai et al., "Deformable convolution networks," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), (Venice, Italy), Oct. 2017, pp. 764-773.

[34] H. Zhang et al., "Context encoding for semantic segmentation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), (Salt Lake Sity, UT, USA), June 2018, pp. 7151-7160.

[35] N. Wang et al., "Transformer meets tracker: Exploiting temporal context for robust visual tracking," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), June 2021, pp. 1571-1580.

[36] B. Yan et al., "LightTrack: Finding lightweight neural networks for object tracking via one-shot architecture search," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), June 2021, pp. 15180-15189.