동영상 기반 산불 감지 학습 모델 분석

장인수, 김광주, 임길택
한국전자통신연구원
jef1015@etri.re.kr, kwangju@etri.re.kr, ktl@etri.re.kr

Analysis of video-based fire detection learning model

In-su Jang, Kwang-Ju Kim, Kil Taek Lim
Electronics and Telecommunications Research Institute

요 약

정형적인 형태가 있는 객체 검출 기법과는 달리 화재 인식은 시간에 따란 형태가 변하는 비정형적인 객체를 검출해야 하기 때문에, 일반적인 이미지 한 장을 이용하는 딥러닝 네트워크 기반 검출 기법을 적용하는데 한계가 있다. 시간에 따른 변화를 인지하기 위해 동영상 데이터를 사용하는 경우, 촬영된 카메라와 네트워크의 상태에 따라 모션 정보가 다르게 저장될 수 있기 때문에 정확한 인식을 위해서는 다양한 환경에 대한 대량의 데이터가 요구되며, 확보된 데이터에 최적화된 딥러닝 모델이 필요하다. 이에 본 논문에서는 실제 화재 동영상에서 불꽃과 연기에 해당하는 부분을 추출하여 동영상 데이터 셋을 구축하였다. 또한, 다양한 동영상 기반 인식 모델에 대해 구축된 데이터 셋을 활용하여 테스트를 진행하였으며 결과를 분석하였다.

Ⅰ. 서 론

일반적으로 많이 사용되는 센서 기반 화재 감지 시스템의 경우, 센서 성능의 한계와 주변 환경 요소로 인해 오 탐지가 발생하더라도 직접 가지 않고서는 확인이 어려운 실정이다. 또한, 센서의 감응 영역에 따라 센서의 가격이 크게 상승하기 때문에 최근 딥러닝 기반의 카메라를 활용한 화재 감지 시스템이 등장하였다. 일반 CCTV 카메라로 획득한 영상을 딥러닝 기반의 분석을 통해 화재를 인식한다. 하지만, 일반적인 객체 검출 네트워크의 경우 형태가 명확한 사물에 대한 검출률도 학습한 데이터 셋의 객체가 정확하게 정의 되어 있어야 높게 나타난다. 화재 인식의 경우 불꽃과 연기의 경계가 모호하기 때문에 학습용 데이터 셋을 구축하는 과정에서 최종 화재 인식 시스템의 정확도가 일반 사물 검출보다 낮아지게 된다[1]. 따라서 이러한 결과를 한 번 더 걸러주는 추가적인 시스템이 필요로 한다. 본 논문에서는 딥러능을 활용한 이미지 기반 화재 검출 결과를 추가적으로 분석하기 위해 동영상 기반의 인식 모델을 사용하였다. 기존의 동영상 기반 인식 모델의 경우 정형화된 사물의 움직임 정보 추정을 위해 개발 되었기 때문에 실제 화재를 인식하는데 사용하기 위해서는 모델의 수정이 불가피하다. 따라서 기존의 다양한 모델들에 대한 학습 결과를 토대로 불꽃 및 연기 인식을 위한 최적의 학습 환경을 분석하고자 한다.

그림 1. 불꽃 및 연기 데이터 레이블링

Ⅱ. 본론

불꽃과 연기 같은 비 정형 객체의 경우 일반적인 객체 검출 학습을 위해 사용하는 사각형의 레이블링 작업에는 한계가 있다. 그림 1과 같이 불꽃이나 연기의 경우 그 형태가 주변 환경 요소들에 의해 시간에 따라 변화하며 그 경계 또한 불분명한 경우가 대부분이다. 특히 연기의 경우 길게 뻗어져 나가면 그 크기가 영상 전체 크기와 유사하게 커지지만 연기에 해당하지 않는 영역도 포함하게 된다. 이는 결과적으로 학습 과정의 오차로 반영되어 최종 성능 저하의 원인이 된다. 따라서 이미지 기반의 객체 검출 네트워크를 불꽃, 연기 감지에 사용하는 것에는 한계가 있다. 이를 보완하기 위해서 본 논문에서는 동영상 기반의 딥러닝 네트워크를 활용하여 추가적인 검증을 진행하였다.

일반적으로 동영상 기반의 행동 인식 네트워크는 시간에 따른 사물의 변화 데이터를 사용하기 때문에 동영상에서 특정 프레임을 추출하여 사용한다. 추출하고자 하는 움직임의 특성에 따라 빠른 움직임일 경우 프레임 간격을 넓히고, 느리게 움직일 경우, 프레임 간격을 좁혀 그 특징을 효율적으로 추출하는 것이 핵심이다. 동영상 기반 데이터셋은 표 1과 같이 화재 동영상에서 256(w) × 256(h) 크기로 100개의 프레임을 추출하여 하나의 데이터로 생성하였다. 클래스는 'Fire', 'Smoke', 'None'의 3개로 구분하였으며, 영상 내에서 불과 연기의 크기 및 위치는 다양하게 분포되어 있다.

표 1 화재 감지 동영상 데이터셋

Class	Size	Frames	Num.	Images
Fire	256	100	3,270	
Smoke	256	100	2,392	
None	256	100	819	

정형화된 사물의 위치 및 움직임을 인지하는 딥러닝 기반 모델들은 많은 연구가 되고 있으나, 비 정형적인 불꽃, 연기, 물 등의 대한 인지 관련 딥러닝 모델들에 대한 연구는 많지 않으며, 정형화된 데이터셋이 없기 때문에 서로에 대한 절대적 평가가 어렵다. 이에 본 논문에서는 기존의 다양한 동영상 기반 인지 모델들을 비교 분석하였다.

표 2에서는 학습에 사용된 모델들과 그 구조 및 결과를 보여준다. 동영상 인식 네트워크에서 일반적인 방법인 C3D[2]와 시간 축을 분리하여 사용한 R2plus1D[3], 공간 축과 시간 축의 프레임 수를 달리 사용하여 빠르게 변화하는 모션을 인지하는 네트워크와 느리게 변하는 모션을 인지하는 네트워크로 구분하여 학습한 Slowfast[4] 와 이 중 느린 변화를 인지하는 네트워크만 사용한 Slowonly 모델을 사용하였다. 학습을 위해 사용되는 입력 데이터는 100개의 프레임 중 각 모델에서 실제 사용하는 프레임 수는 각 모델에 따라 다르며 표 2에서와 같이 특정 간격을 두고 샘플링된 프레임들을 사용하였다. 예를 들어 4 × 16 의 경우 16 프레임 간격으로 4 프레임을 사용하며, 16 × 1 의 경우 프레임 간격 없이 16프레임을 사용하였다. 빠른 변화 감지 네트워크에는 4 프레임에 8배하여 32 프레임을 사용하였다.

실험은 Intel(R) i7-6800K CPU 와, Titan RTX 24GB 2개가 탑재된 컴퓨터로 진행하였다. 실험 결과는 표2에서와 같이 데이터셋 중 학습에 사용하지 않은 20% 테스트 데이터를 사용하여 검증하였다. 전반적으로 단일 네트워크를 사용한 C3D 나 R2plus1D 모델 보다 이중 네트워크를 사용한 Slowfast 모델의 정확도가 높게 나타났다. 또한 프레임 간격 없이 16 프레임을 사용한 C3D 모델의 경우 정확도가 가장 낮게 나왔으며, 16 프레임 간격으로 4 프레임을 사용한 네트워크보다는 8 프레임 간격으로 8 프레임을 사용한 경우의 성능이 높게 나타났으나, 2배로 늘어난 계산량 대비 효율이 낮다.

표 2 각 모델별 성능 비교

Models	Backbone	Frames× Interval	Acc.
C3D	-	16 × 1	0.9398
R2plus1D	res34	4 × 16	0.9533
Slowonly	res50	4 × 16	0.9616
Slowonly	res50	8 × 8	0.9625
Slowfast	res50	4 × 16 (32)	0.9691
Slowfast	res50	8 × 8 (64)	0.9797

Ⅲ. 결론

본 논문에서는 동영상 기반 불꽃 및 연기 인식을 위해 동영상 데이터 셋을 구축하였으며, 기존의 동영상 기반 인식 모델들에 대한 분석을 진행 하였다. 빠른 모션과 느린 모션 변화를 구분하여 추정하는 기법이 성능이 전반적으로 높게 나왔다. 하지만 프레임 수를 더 늘여 시간 축으로 모델을 확장하는 경우에는 소요된 계산량 대비 효율이 좋지 않았다. 이는 불꽃이나 연기의 변화가 빠르지 않아 프레임에 대한 샘플링을 증가하여도 모션에 대한 특징을 추출하는데는 영향을 미치지 않는다. 따라서, 성능의 개선을 위해서는 공간적인 네트워크의 최적화 및 시간적인 경량화가 필요하다.

ACKNOWLEDGMENT

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [21ZD1120, Development of ICT Convergence Technology for Daegu-GyeongBuk Regional Industry].

참 고 문 헌

[1] 김광주, 장인수, 임길택, "산불 연기 데이터셋 구축 및 심층 신경망 기반 검출 기술 비교 분석," 한국통신학회 동계종합학술대회, 2021.

[2] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri1, “Learning Spatiotemporal Features with 3D Convolutional Networks,” IEEE International Conference on Computer Vision, pp. 4489-4497, 2015

[3] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, Manohar Paluri, “A closer look at spatiotemporal convolutions for action recognition,”IEEE conference on Computer Vision and Pattern Recognition, pp. 6450-6459, 2018

[4] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, Kaiming He, “Slowfast networks for video recognition,” IEEE international conference on computer vision, pp. 6202-6211, 2019.

[5] Joseph Redom, Ali Farhadi, "YOLOv3: An Incremental Improvement," arXiv preprint arXiv:1804.02767, 2018.

[6] Ren, S, Girshick, R. Girshick, R, Sun, J, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Transcations on Pattern Analysis & Mechine Intelligence 2017, 23539, 1137- 1149.

[7] R.Girshick, "Fast r-cnn," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440-1448.