Air Piano 를 위한 딥러닝 기반 객체 탐지

Deep Learning-based Object Detection for Air Piano

저자
(Authors)

고영진, 김태일, 김태영
Young-Jin Ko, Tae-Il Kim, Tae-Young Kim

출처
(Source)

한국HCI학회 학술대회 , 2020.2, 45-48 (4 pages)

발행처
(Publisher)

한국HCI학회
The HCI Society of Korea

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE10402704

APA Style

고영진, 김태일, 김태영 (2020). Air Piano 를 위한 딥러닝 기반 객체 탐지. 한국HCI학회 학술대회, 45-48.

Air Piano 를 위한 딥러닝 기반 객체 탐지

Deep Learning-based Object Detection for Air Piano

고영진
Young-Jin Ko
서경대학교
SeoKyeong Univ.
2014305004@skuniv.ac.kr

김태일
Tae-Il Kim
서경대학교
SeoKyeong Univ.
Xodlf8637@skuniv.ac.kr

김태영
Tae-Young Kim
서경대학교
SeoKyeong Univ.
tykim@skuniv.ac.kr

요약문

모바일 환경에서 높은 인식률과 실시간 추론 속도를 보장하는 딥러닝 기반 객체 탐지를 위해서는 네트워크 경량화가 필수적이다. 본 논문에서는 딥러닝 모델의 경량화를 통해 모바일 기기의 카메라로 실시간으로 손가락을 탐지해 공중이나 책상 위에서 피아노 연주를 하는 Air Piano 를 제안한다. 피아노 연주에 사용되는 손가락 탐지를 위해 MobileNet 을 특징 추출기로 사용하는 객체 탐지 모델인 SSD(Single Shot Detector)를 사용한다. 얻어지는 손끝 좌표와 모바일 디스플레이에 출력되는 Piano 건반 위치를 비교해 해당하는 음을 출력한다. 본 방법의 검증을 위해 GALAXY S10+ 기기에서 실험한 결과 평균 90.68%의 정확도로 손가락을 인식하여 모바일 환경에서 실시간 객체 탐지를 통한 Air Piano 연주의 활용 가능함을 알 수 있었다.

주제어

딥러닝, 객체 탐지, 에어 피아노

1. 서 론

1.1 연구 배경

최근 사용자 친화적 인터페이스를 제공하기 위하여 손 제스처 인식에 대한 연구가 활발히 이루어지고 있다[1]. 손 제스처 인식을 통한 인터페이스 조작은 기존의 하드웨어 장치보다 직관적이며 자유로워 사용자에게 몰입감을 높여준다. 손 제스처 인식에 관한 기존의 연구들은 깊이 정보를 파악하는 립모션, ToF 깊이 센서, 가속도 센서 등 별도의 장비가 필요하며 조명이나 거리 등 외부 환경에 제약이 있다[2]. 딥러닝 기술의 비약적인 발전으로 객체 탐지 분야를 이용한 많은 연구가 시도되고 있다[3]. 딥러닝 기술은 추가적인 장비의 도움 없이 외부환경에 강건한 객체 탐지가 가능하다는 장점을 가진다.

딥러닝 모델의 문제점은 인식률을 높이기 위해 깊게 쌓은 신경망의 가중치가 증가하고 추론 모델의 용량이 커짐에 따라 추론 시간 또한 길어진다는 점이다. 컴퓨터보다 상대적으로 연산 능력이 약한 모바일 환경에서 딥러닝 모델의 추론을 실시간으로 실행하기 위해서는 딥러닝 모델의 경량화는 필수적이다. ShuffleNet[4], MobileNet[5]과 같이 합성곱 필터의 변형을 통해 연산량을 줄이는 네트워크 경량화, 가중치의 양을 줄이는 네트워크 프루닝[6], 부동소수점을 가지는 가중치의 근사화인 양자화를 통한 모델 압축으로 모델의 경량화와 추론의 고속화를 구현할 수 있다.

1.2 연구 방법

본 연구에서는 딥러닝 모델 중 하나인 객체 탐지 모델의 경량화를 통해 모바일 기기에서 카메라를 통해 실시간으로 손가락을 인식하여 허공이나 바닥에 피아노를 치는 동작을 통해 피아노 연주가 가능한 Air Piano 를 제안한다. 모바일 기기에 표시되는 건반의 위치와 10 개의 손가락의 끝점이 상호작용하여 피아노 음을 출력하여 연주가 가능하도록 한다.

MobielNet 을 통해 입력 영상의 특징을 추출하고 객체 탐지 모델인 SSD(Single Shot Detector)[7]를 기반으로 각 손가락의 끝점 위치를 추론한다. Tensorflow Lite 변환기를 통해 양자화 및 최적화하여 모바일 기기에서 딥러닝 모델 추론이 가능하게 하였다. 또한 안드로이드 NNAPI(Neural Networks Application Programming Interface)를 활용해 하드웨어 가속 추론연산이 가능하도록 구현하였다.

본 방법의 검증을 위해 GALAXY S10+ 기기에서 실험한 결과 한 프레임에서 평균 90.68%의 정확도를 가지며 평균 25ms 로 손끝 객체를 탐지한다. 평균 초당 40 프레임의 영상이 입력된다는 점과 피아노 연주 속도를 감안할 때 열 개의 손끝 객체를 인식하여 정확한 피아노 연주에 무리가 없으며 피아노 화면 출력과 소리 출력 반응시간에 평균 12ms 가 소요되어 총 37ms 의 소요 시간을 통해 자연스러운 실시간 Air Piano 연주가 가능함을 알 수 있었다.

본 논문의 구성은 다음과 같다. 2 장에서 모바일 기반 손끝 객체 인식방법과 특징 추출기인 MobileNet,

객체 탐지 모델인 SSD 구현 방법에 대해 설명하고 3 장에서는 객체 인식 기반 Air Piano 의 실험과 본 연구에서 사용된 데이터 세트의 구성과 제작과정을 기술한다. 마지막으로 4 장에서 실험 결과와 함께 결론을 맺는다.

2. 모바일 기반 손끝 객체 인식 방법

본 연구에서 제안하는 모바일 기반 Air Piano 구현 과정은 (그림 1)과 같다. 사용자 앞에 세워진 모바일 기기의 카메라를 통해 열 손가락의 손끝 점이 모바일로 이식한 SSD 모델의 추론을 통해 열 개의 손끝 좌표로 변환된다. 다중 객체 탐지가 가능한 SSD 모델을 통해 손가락을 탐지하여 피아노 건반의 누름을 인식한다.

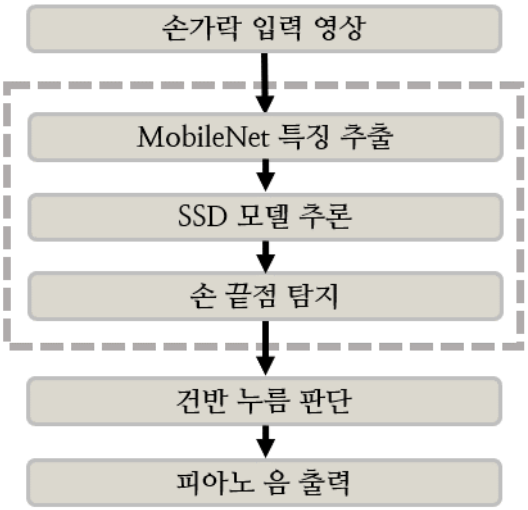

그림 1 Air Piano 처리 과정

2.1 MobileNet 기반 특징 추출

입력 영상 중 손가락 객체의 특징을 효과적으로 추출하기 위해 완전 연결 신경망에 합성곱 연산과 최대 풀링을 추가한 합성곱 신경망을 사용한다. MobileNet 은 기존의 합성곱 연산을 깊이필터를 이용한 깊이별 (Depthwise) 합성곱과 1x1 위치별 (Pointwise) 합성곱으로 분리해 정확도에 큰 손실없이 기존보다 연산량을 약 8~9 배 줄여 모바일 기기에서 사용할 수 있도록 추론의 고속화와 경량화를 위해 설계된 딥러닝 모델이다.

2.2 SSD 기반 객체 탐지

입력 영상을 기반으로 MobileNet 에서 추출된 특징을 사용해 SSD 모델로 손가락을 인식한다. SSD 는 회귀문제인 탐지된 객체의 경계 상자 추론의 오차 값과 분류 문제인 라벨 값 추론의 오차 값을 단일 네트워크로 구성하기 때문에 RPN (Region Proposal Network)과 같은 별도의 네트워크로 후보 영역을 계산하여 경계 상자를 예측하는 기존의 Faster-RCNN 과 같은 이중 네트워크 구조보다 빠른 객체 탐지가 가능하다.

300x300 크기의 3 채널 입력 영상을 MobileNet 특징 추출기와 합성곱 연산을 통해 6 개의 다양한 해상도를 가지는 특징맵을 생성한다. 각 특징맵에 대해 3x3 합성곱 연산을 통해 객체 유무 가능성과 경계 상자 (Bounding Box)의 위치에 대한 오프셋을 계산해 경계 상자를 추론한다. 예측된 수 많은 경계상자들의 겹치는 부분에 대해 NMS(Non-Maximum Suppresion) 알고리즘을 적용하여 경계 상자의 중복된 영역을 제거해 최종적으로 객체가 탐지된 경계상자를 구한다.

객체 탐지를 위해 특징맵을 일정 크기로 나눠 사전에 정의된 다양한 종횡비를 가진 기본 상자(Default Box)를 생성하고 참 경계 상자(Ground Truth Box)와의 IoU(Intersection over Union)의 값이 0.5 이상인 상자들을 선택해 객체로 설정하며 0.5 이하인 경우는 배경으로 설정한다. 일반적인 영상의 경우 객체보다 배경이 차지하는 부분이 많으므로 기본 상자의 대부분은 배경을 포함하고 일부 기본 상자만 객체를 포함하고 있어 경계 상자를 예측할 때 학습의 불균형이 발생한다. 이를 해결하기 위해 배경 중 객체라고 잘못 판단한 경계 상자를 부정 값(Negative Sample)으로 지정하고 객체를 긍정 값(Positive Sample)으로 하여 부정 값과 긍정 값의 비율을 3:1로 맞춰 학습한다. 이를 통해 빠른 수렴과 안정적인 학습이 가능하다. 추론 과정에서 다양한 해상도의 특징맵을 추출하는 SSD의 특성과 NMS 알고리즘을 통해 다양한 크기의 객체탐지가 가능하다.

2.3 건반 누름 판단

Air Piano 는 (그림 2)와 같이 모바일 기반 어플리케이션으로 스마트폰을 사용자의 앞에 세워 전면 카메라를 통해 사용자의 손 영상을 입력 받아 손가락의 움직임을 인식하여 사용자가 의도한 음을 출력한다. 사용자는 스마트폰 디스플레이에 출력되는 건반의 위치에 맞추어 허공이나 책상에 손가락으로 피아노를 치는 듯한 동작을 하여 연주가 가능하다.

그림 2 Air Piano 시연

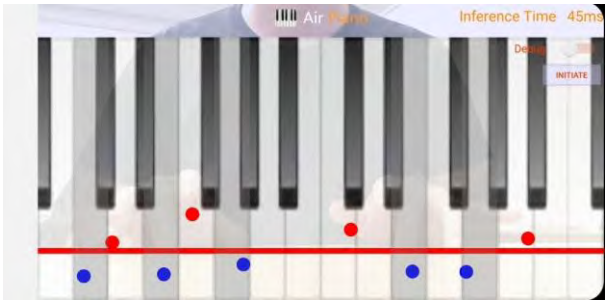

그림 3 Air Piano 실행 화면

그림 3 은 Air Piano 의 실행 화면 중 일부이며 디스플레이에 표시되는 반투명한 건반 이미지와 함께 사용자의 손 움직임을 함께 확인할 수 있는 입력 영상이 같이 출력된다. 붉은 실선이 실제 건반의 눌림을 판단하는 기준이 되어 손끝 점이 기준선 아래로 입력되는 순간 해당 건반의 음이 출력된다.

프레임마다 손가락 객체와 손끝 좌표를 계산하기 때문에 특정 하나의 건반을 계속 누르고 있다면 프레임 수만큼의 피아노 음이 반복해서 출력될 것이다. 이 문제를 해결하기 위해 각 건반은 5 개의 큐와 1 개의 건반 눌림 플래그로 구성되어 있다. 같은 음을 연달아 내기 위해서 누르고 있는 건반의 손가락을 뗀 후 다시 건반을 눌러야 하는 점은 실제 피아노의 동작과 유사하다. 또한 건반이 눌리게 되면 손끝을 인식하는 붉은색 점이 파란색으로 변하고 해당 건반의 색이 짙게 변하여 청각 뿐 아니라 시각적 효과도 줌으로써 사용자의 몰입도를 높인다.

3. 실험

3.1 학습 데이터세트

손가락 탐지를 위한 학습은 사람, 강아지, 자동차 등 80 가지의 클래스로 구성된 COCO 데이터 세트를 기반으로 학습한 SSD 모델의 가중치를 통해 전이 학습했다[8,9]. 전이 학습은 이전의 학습한 특징들을 활용해 재학습하기 때문에 무작위 값으로 초기화한 필터들을 학습하는 방법보다 적은 데이터 양으로 빠르게 학습시킬 수 있는 장점이 있다.

그림 4 는 학습 데이터세트의 예시 및 참 경계 상자이다. Air Piano 연주와 같이 연주하는 영상의 캡처를 통한 학습 데이터세트 생성과 손끝 객체의 정답 값과 위치 값을 xml 형식으로 저장해 지도학습을 수행한다. 학습 데이터세트로는 4 명의 학습자가 다양한 조명, 거리, 배경에서 수집한 2,000 장의 손가락 원본 영상 데이터세트에서 10%인 200 장을 테스트 데이터세트로 사용하고 나머지 데이터를 다양한 방법으로 증강시켰다. 이미지 데이터의 증강에 따라 참 경계상자도 함께 변환해 학습 데이터로 사용되었으며 명암 변화에 강인하도록 3 배 증강하였다. Air Piano 특성 상 영상을 통해 움직이는 객체의 영상을 입력으로 받기 때문에 7x7 크기의 모션 블러 필터로 블러 처리를 통해 손끝 움직임의 흐릿한 입력에도 높은 인식률을 유지하도록 증강하였다. 또한 다양한 각도에서도 올바르게 탐지되도록 증강된 손가락 영상을 좌우 최대 30 도씩 무작위로 회전시켜 3 배 증강하여 총 12 배의 증강을 통해 21,600 장의 학습 데이터세트를 구성하여 학습시켰다.

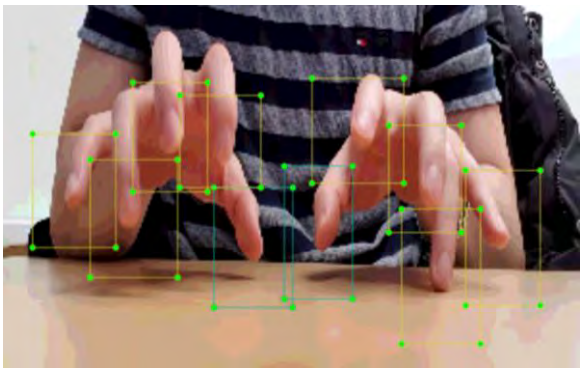

그림 4 학습 데이터 예시 및 참 경계 상자

3.2 실험 환경

본 실험은 프로세서 Intel Core i5 8400, 그래픽 카드 GeForce GTX1080Ti, 16GB RAM 으로 구성된 PC 에서 Python3.6 과 Tensorflow-gpu 1.13.1 을 개발도구로 사용한 환경에서 학습시킨 SSD 모델을 TensorFlow Lite Converter 를 통해 tflite 파일 형식으로 변환해 모바일 기기 GALAXY S10+에서 실험하였다.

3.3 성능 분석

본 실험에서는 4 명의 학습자가 다양한 장소와 조명에서 수집한 200 장의 테스트 데이터세트로 구성하여 손끝 객체 탐지의 성능을 분석하였다. 본 연구에서 제안한 Air Piano 의 모바일 객체 탐지 성능을 정량적으로 분석하기 위해 단일 객체 탐지의 성능 평가 지표로 사용되는 AP(Average Precision)를 사용한다. AP 는 검출된 객체들 중

정확한 검출의 비율을 의미하며 IoU 의 값이 0.5 이상인 객체를 정검출로 판단하였다. 원본 데이터에서 무작위로 추출한 손가락 영상 200 장을 테스트 데이터세트로 정확도를 계산한 결과 90.68%의 정확도를 보였다.

모바일 기기에서 학습된 객체 탐지 모델의 추론 속도는 실시간 적용을 위한 중요한 지표이다. 3 명의 학습자가 총 2 시간의 테스트를 통해 얻어진 추론 속도 데이터를 통해 평균 25ms 의 속도로 손가락 객체를 탐지하여 Air Piano 를 수행하는 것을 알 수 있었다.

4. 결론

본 논문에서는 모바일 기기에서 카메라를 통해 들어오는 입력 영상에 실시간으로 손가락을 탐지하여 피아노 연주를 하는 방법을 제안하였다. SAMSUNG GALAXY S10+기기에서 손가락 객체 탐지 모델을 활용한 결과 평균 25ms 의 추론 시간으로 손가락을 탐지해 실시간으로 실제 피아노를 연주하듯이 연주할 수 있음을 보였다. Air Piano 를 기반으로 피아노 연주 교육 및 게임 응용과 향후 연구로 깊이 정보 추론을 통한 흰 건반 뿐 아니라 검은 건반도 연주할 수 있는 연구를 수행할 예정이다.

사사의 글

이 논문은 2017 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업. (No.NRF-2017R1D1A1B03029834)

참고 문헌

1. 전찬규, 김민규, 이지원, 김진모. 손 인터페이스 기반 3 인칭 가상현실 콘텐츠 제작 공정에 관한 연구. 컴퓨터 그래픽스학회 논문지, 23(3) (2017) 9-17.
2. 이민호, 허환, 황영배. 홀로렌즈에서 시선 추적 및 손동작 인식을 이용한 통합 인터페이스 시스템. 한국HCI학회 학술대회, (2019), 1002-1005.
3. 오동한, 이병희, 김태영. 외부 환경에 강인한 딥러닝 기반 손 제스처 인식. 한국 차세대 컴퓨팅학회 논문지, 14(5) (2018), 31-39.
4. Zhang, X., Zhou, X., Lin, M., Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018), 6848-6856.
5. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint (2017), arXiv:1704.04861.
6. Han, S., Mao, H., Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint (2015), arXiv:1510.00149.
7. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., Berg, A. C. Ssd: Single shot multibox detector. In European conference on computer vision (2016), 21-37.
8. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L. Microsoft coco: Common objects in context. In European conference on computer vision (2014), 740-755.
9. Pan, S. J., & Yang, Q. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10) (2009), 22(10), 1345-1359.