딥 러닝 기반 감정인식 시스템 개발

이민규, 김대하, 최동윤, *송병철
인하대학교
*bcsong@inha.ac.kr

Emotion Recognition System based Deep Learning

Min Kyu Lee, Dae Ha Kim, Dong Yoon Choi, *Byung Cheol Song
Inha University

요 약

최근 딥 러닝의 발전으로 얼굴인식뿐만 아니라 더 세부적인 기술인 ID식별, 감정인식 등을 분류할 수 있는 알고리즘이 많이 제안되었다. 하지만 딥 러닝은 방대한 연산량을 처리해야 하기 때문에 실시간으로 영상을 구현하는 것은 한계가 있다. 본 논문은 위와 같은 문제를 개선하기 위하여 얼굴인식은 연산량이 비교적 적은 HOG알고리즘을 적용하여 전처리를 진행한다. 그 이후 ID식별 네트워크인 FaceNet과 EmotiW 2017 Challenge의 논문의 감정인식 네트워크를 Multi-Thread 기술을 적용하여 스레드를 분할 연산을 통하여 실시간으로 영상을 출력하는 알고리즘을 제안한다.

1. 서론

최근 강력한 병렬 처리 성능을 제공하는 GPU의 도입과 빅데이터의 발전으로 인공지능이 급속도로 발전하고 있다. 인공지능의 세부 분야인 딥 러닝은 이러한 방대한 데이터를 바탕으로 인간이 가르치지 않아도 스스로 학습하는 기술이다. 딥 러닝은 컴퓨터 비전 알고리즘의 성능을 향상시켜주는 머신 러닝(기계학습)의 일종으로, 이를 이용하여 얼굴인식, ID식별, 감정인식 등을 구현한 다양한 알고리즘이 제안되었다. 먼저, 얼굴인식에는 HOG(Histograms of Gradients)를 통해 검출하는 방법이 있다[1]. 두 번째로, ID식별에는 FaceNet을 이용한 딥 러닝 네트워크를 통해 학습된 데이터를 바탕으로 인간을 구별하는 방법이다[2]. 마지막으로 감정인식은 동영상에서 연속된 프레임의 시퀀스 정보를 받아서 다수의 딥 러닝 네트워크를 앙상블하는 방법이다[3]

하지만 ID식별 및 감정인식 어플리케이션 구현 시 표정의 변화를 고려하는 비디오 시퀀스 정보가 필요하고, 뿐만 아니라 딥 러닝 네트워크 자체의 방대한 연산을 해야 하기 때문에 실시간 구현에는 한계가 있다. 이를 해결하기 위해 Multi-Thread 기술을 이용한다. 이 기술은 CPU 프로세스 내에 같은 메모리 공간을 공유하는 다수의 스레드를 생성할 수 있기 때문에 메모리를 적게 사용하고 병렬 처리해서 실시간 구현을 가능하게 한다.

본 논문에서는 얼굴과 표정을 바탕으로 컴퓨터 비전과 딥 러닝 기술을 종합하여 ID를 식별하고, 감정을 파악하고, 이를 Multi-Thread 기술을 적용하여 서로 다른 스레드에서 병렬 처리하여 실시간으로 출력하는 알고리즘을 제안한다.

본 논문의 구성은 다음과 같다. 2절에서 제안하는 알고리즘의 프레임웍과 구현 기술들을 설명한 후 3절에서는 실험 결과를 확인한다. 마지막으로 4절에서 본 논문에 대한 결론을 맺는다.

2. 제안 알고리즘

제안하는 알고리즘의 프레임웍은 그림 1과 같다. 먼저, 입력 데이터인 단일 프레임을 ID식별, 감정인식 네트워크의 입력 데이터를 만들기 위해 얼굴인식 알고리즘을 통한 얼굴을 검출한다. 이후 검출 영역만 남도록 이미지를 자르는 전처리를 하고 나서 ID식별 네트워크의 입력으로 넣어 결과를 얻는다. 한편, 감정인식 네트워크는 비디오 시퀀스 정보가 필요하기 때문에 단일 프레임 입력이 아니라 40 프레임의 입력이 필요하다. 따라서, 전 처리된 이미지를 저장할 수 있는 버퍼에 보관한 후 40 프레임이 되면 감정인식 네트워크를 적용하여 감정 정보를 얻는다. 여기서, 판단할 수 있는 감정은 총 7가지로 angry, disgust, fear, happy, neutral, sad, surprise이다.

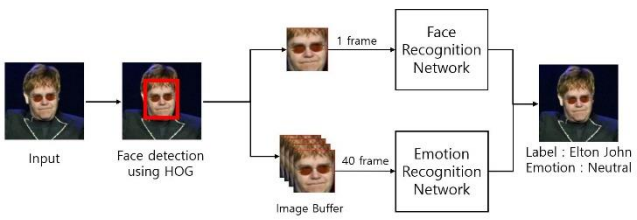

그림 1. 제안하는 알고리즘의 Framework

2.1 Multi-Thread

위 프레임웍을 실시간으로 처리하기에는 많은 시간이 소요되므로 Multi-Thread 기술을 이용하며 구현은 그림 2과 같다. 우선, 메인 스레드에서는 실시간 영상 출력과 동시에 단일 프레임을 서브 스레드로 보낸다. 이와 동시에 서브 스레드에서는 받아온 이미지를 프레임웍에 적용하여 ID식별 및

감정 결과를 얻을 수 있다. 메인 스레드와 서브 스레드는 메모리 공간을 공유하기 때문에 데이터를 주고 받을 수 있다는 점을 이용한다.

파이썬에서 Multi-Thread를 구현하기 위해 threading 모듈을 이용한다. 이 모듈을 이용하면 서브 스레드를 객체화하여 조작할 수 있다. 서브 스레드 객체 내에 프레임웍 구현 알고리즘을 포함시켜 ID와 감정을 추론한다. 또한, 모듈 내 start함수로 스레드를 적절한 시점에 동작할 수 있다. 스레드를 동작시키는 시점은 프레임이 입력으로 들어왔을 때이므로 프레임의 유무를 조건화하여 서브 스레드를 조작한다. 이 때 프레임과 같이 스레드 간 공유하는 데이터는 전역 변수로 선언해야 데이터를 주고 받을 수 있는데 프레임웍을 통한 결과인 ID 및 감정 정보도 마찬가지이다. ID와 감정 정보를 메인 스레드에서 받으면 이를 영상에 출력한다.

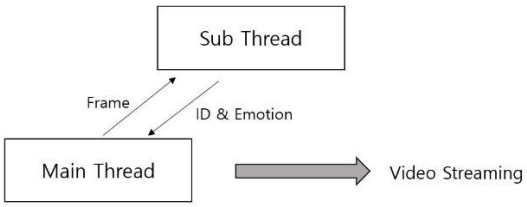

그림 2. Multi-Thread

2.2 얼굴인식

얼굴인식은 HOG(Histogram of Gradients)를 이용한다[1]. HOG는 해당 영상의 영역을 일정 크기의 셀로 분할하고, 각 셀마다 에지 픽셀의 방향에 대한 히스토그램을 구한 후 이들 히스토그램 bin 값들을 일렬로 연결한 벡터를 만든 특징점 기술자이다. HOG 기술자는 회전 혹은 형태 변화에 대한 불변성을 보이지 않기 때문에 사람, 자동차 등과 같이 내부 패턴이 복잡하지 않으면서 고유의 독특한 윤곽선을 갖는 물체를 식별하는데 적합한 특징점 기술자이다. HOG는 파이썬 오픈소스 라이브러리인 Dlib를 통해 구현한다.

2.3 ID식별

ID식별은 FaceNet 기반으로 만들어진 파이썬 오픈소스 라이브러리인 Openface를 이용한다[2]. FaceNet은 Deep Convolutional Neural Network(CNN) 구조를 이루고 있다. 이 네트워크의 출력은 128 차원의 특징점 벡터로 이루어지는데, 얼굴과 같은 원시 데이터를 N차원의 특징점 벡터로 수치화 하는 것을 embedding이라 한다. 이러한 방법으로 미리 학습된 최적의 가중치를 적용하여 특징점 벡터를 얻은 후 Support Vector Machine(SVM)을 통해 분류한다.

2.4 감정인식

감정인식은 EmotiW 2017 Challenge 논문[3]의 네트워크 구조의 일부인 Semi-supervised learning with 3D auto-encoder (S3DAE)와 Convolutional 3D with auxiliary network (C3DA)를 이용한다. 두 구조는 단일 영상이 아닌 비디오 시퀀스의 영상을 가지고 학습하였으며, 얻어진 얼굴의 특징점 벡터를 가지고 감정을 판단한다. 두 네트워크로 얻어지는 각 감정에 해당하는 확률을 가중치 합산을 통하여 사람의 내면 감정을 파악한다. 구현은 파이썬 기반 딥 러닝 라이브러리인 Keras를 이용했다.

3. 실험 결과

실험을 위해 10명의 인물의 이미지 데이터를 바탕으로 FaceNet을 적용하여 ID식별을 위한 학습을 진행했다. 이 때 연산의 효율성을 위하여 720x1280 크기를 갖는 원본 프레임을 0.25배로 축소하여 얼굴인식을 하여 전처리를 진행하고, FaceNet을 적용한다. 같은 방식으로 감정인식 네트워크에서 원본 프레임을 축소한 후 전처리를 하지만 각 네트워크의 정해진 입력 형태를 고려하여 영상의 크기를 재조정한다. 여기서, S3DAE는 112x112이고, C3DA는 224x224로 보간법으로 영상을 재조정한다. 그림 3에서의 결과를 통해 학습된 인물 중 하나로 ID 식별이 출력됨을 알 수 있다.

(a) (b)

그림 3. 실험 영상. (a) 감정 Neutral (b) 감정 Happy

	Single-Thread	Multi-Thread
평균 FPS	6.45	18.11

표 1. 평균 FPS 측정 결과

또한, (a)에서 무표정한 얼굴을 하고 있는 비디오 정보를 통해 중립 감정인 neutral을 출력하고, 마찬가지로 웃는 정보를 통해 happy 감정을 출력한다.

Multi-Thread를 유무에 따른 성능비교를 하기 위해 100 프레임을 처리하는데 걸리는 시간을 측정하여 Frame per second(FPS)를 계산했다. 측정에 이용한 PC사양은 3.2GHz i5-3470, 12GB RAM, Geforce GTX 1060 GPU이며, 측정 결과는 표1과 같이 멀티 스레드가 싱글 스레드보다 3배의 속도 차이가 난다. 싱글 스레드는 ID식별과 감정인식을 순차적으로 진행하기 때문에 프레임 주사 속도를 감소시킨다. 반면, 멀티 스레드는 병렬 처리를 통해 프레임 주사 속도를 일정하게 유지할 수 있다. 이 결과를 Multi-Thread를 통해 실시간성을 확인했다.

4. 결론

본 논문은 얼굴인식, ID식별, 감정인식이라는 3가지를 기술 실시간 영상 출력을 위해 Multi-Thread를 통해 보완하였다. 그러나 얼굴 검출 부분에서 어두운 영상의 경우에는 전 처리 단계에서 얼굴을 못 찾는 경우가 있었고, 감정인식 알고리즘

부분에서 시퀀스 정보가 처리되어야 하기 때문에 결과를 출력하기까지 많은 시간이 소요되었다. 앞으로 이를 개선하기 위한 연구를 진행할 것이다.

감사의 글

본 논문은 산업통상자원부의 산업기술혁신사업으로 지원된 연구결과입니다. [10073154, 인간 내면상태의 인식 및 이를 이용한 인간친화형 인간-로봇 상호작용 기술 개발]

참고문헌

[1] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE, 2005.

[2] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[3] Dae Ha Kim et al. 2017. Multi-modal Emotion Recognition using Semi-supervised Learning and Multiple Neural Networks in the Wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 7 pages.