딥러닝 기반의 실시간 입모양 인식 시스템 구현

*조동훈, **김원준
건국대학교
*johun204@gmail.com, **wonjkim@konkuk.ac.kr

Real-Time Lip Reading System Implementation Based on Deep Learning

*Dong-Hun Cho **Won-Jun Kim
Konkuk University

요약

입모양 인식(Lip Reading) 기술은 입술 움직임을 통해 발화를 분석하는 기술이다. 본 논문에서는 일상적으로 사용하는 10개의 상용구에 대해서 발화자의 안면 움직임 분석을 통해 실시간으로 분류하는 연구를 진행하였다. 시간상의 연속된 순서를 가진 영상 데이터의 특징을 고려하여 3차원 합성곱 신경망 (Convolutional Neural Network)을 사용하여 진행하였지만, 실시간 시스템 구현을 위해 연산량 감소가 필요했다. 이를 해결하기 위해 차 영상을 이용한 2차원 합성곱 신경망과 LSTM 순환 신경망 (Long Short-Term Memory) 결합 모델을 설계하였고, 해당 모델을 이용하여 실시간 시스템 구현에 성공하였다.

1. 서 론

입모양 인식 기술은 입의 움직임으로 발화를 분석하는 기술이다. 여러 명이 동시에 말하거나 잡음 환경에서 떨어지는 음성인식의 정확도에 대해 보조 역할로써 사용할 수도 있다. 기존에도 딥러닝을 이용해 입모양 인식 시스템을 설계하기 위한 많은 연구가 수행되었다[1]. 본 연구에서는 웹캠을 이용한 실시간 분석을 위해 적은 연산량과 높은 정확도의 새로운 모델을 설계하여 구현해보았다.

2. 본 론

2.1. Dataset

학습에 사용한 데이터로는 학습에 사용하기에 적절한 수의 데이터 개수와 다양한 성별과 인종의 발화 영상을 가진 Ouluvs2 Database를 사용하였다. Ouluvs2 Database는 일상에서 사용되는 10개의 상용구를 정면에서 3회 발화한 영상을 포함한다. 10개의 상용구는 'Excuse me', 'Goodbye', 'Hello', 'How are you', 'Nice to meet you', 'See you', 'I am sorry', 'Thank you', 'Have a good time', 'You are welcome' 으로 구성되어있다. 영상 데이터는 전체 52명의 실험자의 발화 영상이며 39명은 남성, 13명은 여성이다. 본 연구에서는 무작위 40명의 영상을 학습용으로, 다른 12명의 영상을 검증용으로 사용하였다.

2.2. 데이터 전처리 과정

각각의 영상 데이터에서 입술 영역을 검출하기 위해 dlib[2]를 사용하였다. dlib는 얼굴에서 특징점을 검출하여 **Fig. 1** 과같이 반환한다. 입술 영역만을 사용하여 진행하기 위해 49~68번째 특징점 영역을 1:2 비율로 잘라내어 **Fig. 2** 와 같이 사용하였다. 잘라낸 입술 영역의 이미지는 1채널의 흑백 이미지로 변환하였으며 16*32 픽셀 크기로 변환하였다. Data Augmentation을 위해 입술 영역을 중심을 기준으로 8방향 (좌측 상단, 상단, 우측 상단, 우측, 우측 하단, 하단, 좌측 하단, 좌측)으로 10 픽셀씩 이동하여 잘라내었으며 좌우 반전을 통해 16개의 데이터를 생성하였다. 마지막으로 전체 픽셀에 대해 Z-Score 정규화를 진행하였다. Z-Score 정규화는 식 (1)과 같이 표현될 수 있다.

$$z = \frac{(x-\mu)}{\sigma} \tag{1}$$

x: 값, μ: 평균, σ: 표준편차

Fig. 1. dlib Facial landmarks

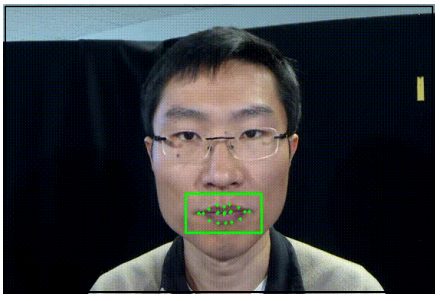

Fig. 2. 입술 영역 검출

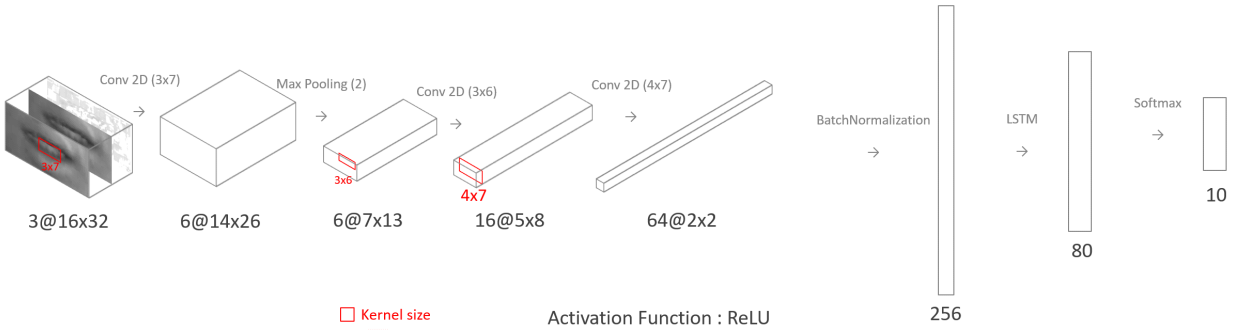

Fig. 3. 차 영상을 이용한 2차원 합성곱 신경망과 LSTM 순환 신경망 결합 모델

3. 모델 설계

3.1. 3차원 합성곱 신경망 모델

영상 데이터는 시간의 순서에 따른 시계열 데이터이다. 연속된 프레임에 대해 상관성을 부여하기 위해 연속된 8개의 프레임을 결합하여 8채널 이미지를 입력으로 하는 3차원 합성곱 모델을 설계하였다. 8개의 3차원 합성곱 레이어와 Softmax 함수를 이용하여 10개의 결과로 Classification 하는 것으로 65.1%의 정확도로 인식에 성공했다.

3.2. 3차원 합성곱 신경망과 LSTM 순환 신경망 결합 모델

앞선 3차원 합성곱 신경망의 모델에 LSTM 순환 신경망을 결합한 모델이다. 이전 프레임과 현재 프레임의 연산 결과를 순환 신경망에 함께 입력하여 시계열 데이터 처리의 정확성을 높였다. 동일하게 Softmax 함수를 이용하여 89.4% 정확도로 인식에 성공했다. 하지만 모든 프레임에 대해 3차원 합성곱 연산을 수행하기 때문에 실시간 시스템 구현에 성능 저하가 발생하였고, 이를 해결하기 위해 3차원 합성곱을 사용하지 않는 새로운 모델을 설계하였다.

3.3. 차 영상을 이용한 2차원 합성곱 신경망과 LSTM 순환 신경망 결합 모델

2차원 합성곱을 사용하며 연속된 프레임 간의 상관성을 부여하기 위해 두 프레임의 차 영상 (Difference image)을 이용하였다. 모델의 입력 데이터는 3채널 이미지로 구성하였는데, 순서대로 이전 프레임의 영상, 현재 프레임의 영상, 두 프레임의 차 영상으로 구성하였다. 설계한 모델은 Fig. 3 과 같다. 인식 정확도는 76.7%으로 앞서 수행한 3차원 합성곱 신경망과 LSTM 순환 신경망 결합 모델보다 다소 낮았지만, 적은 연산량으로 실시간 시스템 구현에 더 적합하다고 판단하였다.

4. 실험결과

실험은 GTX 1050을 이용하여 진행하였으며 구현을 위해 Python과 Pytorch를 사용하였다. 설계 모델은 차 영상을 이용한 2차원 합성곱 신경망과 LSTM 순환 신경망 결합 모델을 사용하였다. Ouluvs2 Database와 웹캠 환경에는 차이가 있기 때문에, Ouluvs2 Database를 이용하여 학습한 Pre-training 모델에 웹캠으로 촬영한 영상으로 Fine-tuning을 진행하여 인식 정확도를 높였다. 각 문장마다 발화 길이가 다르기 때문에 발화의 시작과 끝을 인식하기 위해, 입술의 움직임을 감지하여 1초간 움직임이 없을 경우 정지 상태로 전환되도록 구현하였다.

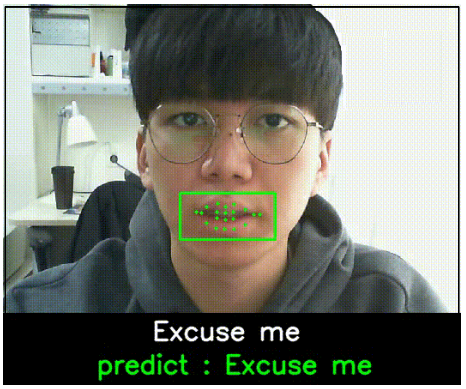

Fig. 4. 웹캠을 이용한 실시간 입모양 인식 결과

5. 결 론

본 논문에서는 실시간 입모양 인식 시스템 구현에 적합한 몇 가지 모델 구조를 제안하였다. 첫 번째 방법으로 3차원 합성곱 신경망 모델을 제안하였는데, 비교적 낮은 정확도와 많은 연산량으로 인해 적합하지 않았다. 두 번째 방법으로는 3차원 합성곱 신경망과 LSTM 순환 신경망 결합 모델을 제안하였는데, 높은 정확도로 인식에 성공했지만 많은 연산량으로 인해 실시간 시스템 구현에 어려움이 있었다. 마지막으로 차 영상을 이용한 2차원 합성곱 신경망과 LSTM 순환 신경망 결합 모델을 제안하였다. 차 영상을 이용해 움직임이 없는 배경 영역을 제거할 수 있어 적은 연산량으로도 효과적인 인식이 가능했으며, 해당 모델을 이용해 실시간 시스템 구현에 성공하였다.

감사의 글

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구 결과로 수행되었음" (No.2018-0-00213, SW중심대학(건국대학교))

참고문헌

[1] Ivan Fung, Brian Mak, "END-TO-END LOW-RESOURCE LIP-READING WITH MAXOUT CNN AND LSTM", 2018 IEEE

International Conference on Acoustics, Speech and Signal Processing (ICASSP '18)

[2] D. E. King, "Dlib-ml: A machine learning toolkit," Journal of Machine Learning Research, vol. 10, pp. 1755-1758, 2009.