수신호 인식과 표정인식을 이용한 위험 상황 인지

이은택 [1,6,O], 송윤아 [2,6], 하은겸 [3,6], 황준하 [4,6], 이연주 [5,6,*]

고려대학교 전자및정보공학과 [1], 빅데이터사이언스학부 [2], 경제통계학부 [3], 컴퓨터정보학과 [4],

응용수리과학부 [5]

고려대학교 크림슨브레인 [6]

교신저자 [*]

hoya9802, yas0531, dign7984, jack34763185, leeyeonju08@korea.ac.kr

요 약

장기화되어가는 팬데믹 상황 속에서 가정에서 보내는 시간이 증가함에 따라 증가하는 가정폭력을 피해자가 직접 해결을 요청을 하기에는 어려움이 있어 가해자에게 들키지 않고 관련 기관에 도움을 요청할 수 있는 새로운 기술의 필요하다. 본 연구에서는 카메라에 손 제스처와 얼굴표정으로 현재의 상황을 전달하는 상황을 구성하고 손 제스처 인식과 얼굴 표정 인식을 통해 상황인지와 표정감지를 하는 알고리즘을 제안한다. 이 연구에서는 Hand Tracking Module 를 이용해 손의 특징점을 감지하고 LSTM(Long Short-term Memory)([12]) 네트웍 구조를 사용하여 상황을 분류하며 표정의 감정분류를 위해 VGG16([11]) 구조를 사용하는 위험 상황 인지 알고리즘을 제안하고 그 실험결과를 소개한다.

1. 서론

전세계적으로 유례없는 코로나 19 감염병의 확산을 방지하기 위해 각국의 사회적 거리두기, 락다운(Lock Down) 정책 등으로 인해 공적 공간에서의 활동이 위축되고 사적 공간인 가정에서의 생활시간이 크게 증가함에 따라, 가정 내에서 폭력이 발생할 가능성이 높아지고 학대와 통제가 더욱 심화되어 적절한 신고 및 지원 요청에 어려움을 겪고 있다는 분석이 잇따르고 있다. 실제로 이탈리아, 미국, 스페인에서 코로나 19 로 인한 봉쇄 이후 구조전화 상담 건수와 가정폭력 신고율이 대폭 감소하였다[1]. 뿐만 아니라 피해자 지원에 필요한 경찰 및 사법 서비스 인력이 코로나 대응 인원으로 편성되면서 서비스가 축소되어 기존의 전통적인 방법으로는 피해자들을 효과적으로 보호할 수 없다는 목소리가 커지고 있다. 이 문제를 해결하기 위한 방법 중 하나로 캐나다의 여성 인권단체에서 가정폭력을 알리는 수신호 알리기 위해 유명 sns 를 통해 챌린지를 진행하였으며 이를 통해 해결된 수많은 해외 사례들이 존재한다.

모션 인식과 관련하여 다양한 연구가 진행되어 왔다. 몇 가지 논문의 사례를 살펴보려고 한다. '인공지능 기반 응급상황 예측에 대한 연구'의 경우, 환자 낙상이라는 동작에 대한 예측 시뮬레이션 모델을 다루고 있다[2]. 이 논문은 LSTM(Long Short-term Memory)과 RNN 방식을 활용하여 응급상황 예측 알고리즘 모델의 결과를 분석해 비교한다. 해당 논문의 결과는 LSTM 의 성능이 가장 우수하다고 말하고 있다. '시계열 데이터 분류와 NAS 를 통한 손동작 인식'이라는 논문을 통해서 우리는 스켈레톤 기반 손동작 인식이 시계열 데이터라는 점을 알 수 있었다[3]. 본 연구에서는 이 결과들을 반영하여 손 특징점에 대한 LSTM 을 이용해 모션 인식을 분석하고자 했다.

본 연구에서는 이 수신호와 사용자의 표정을 분석해 피해상황을 분석해 구조신호를 보내는 서비스를 제공한다. 2 장에서는 제안방법인 위험상황 인지를 위한 손동작 인식 데이터 수집에 사용된 라이브러리 설명과 표정 인식 방법에 사용된 모델에 대해 설명하고 3 장에서는 제안한 방법의 성능을 분석하고 4 장은 결론을 기술한다.

그림 1. 표정 및 손 제스처 인식

2. 영상 특징을 이용한 얼굴인식 방법

그림 2. 데이터 수집 과정
(help, handout, fist, nothing 순)

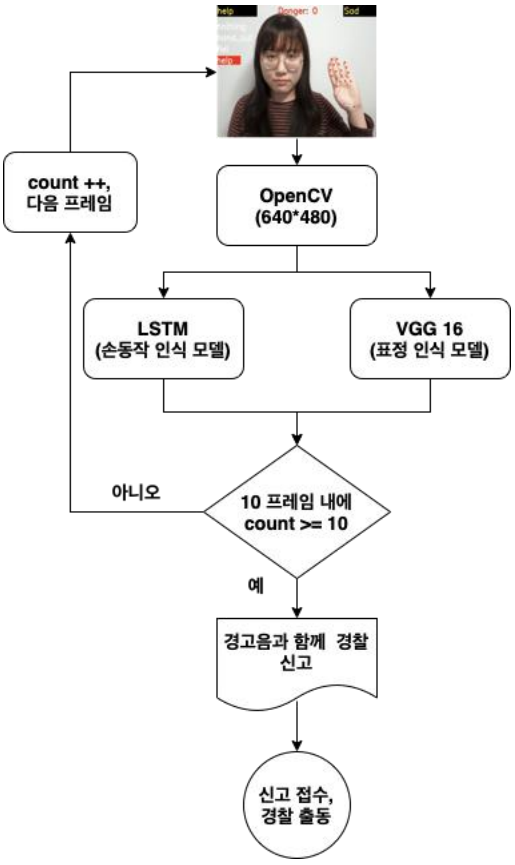

그림 3. 알고리즘 전체 flow

본 연구에서는 이 수신호와 사용자의 표정을 분석해 피해상황을 분석해 구조신호를 보내는 서비스를 제공한다. 유사한 손동작을 인식해 손동작 인식 뿐만 아니라 표정 인식 모델을 같이 제공함으로써 이와 유사한 손동작이지만 다른 의미를 가진 경우를 최대한 배제하기 위해 추가해서 좀 더 신뢰성 있는 모델을 제안한다. 모델의 학습 데이터는 구분 동작의 순서가 중요한 데이터라는 점에서 이전 상태를 통해 다음 상태를 예측할 수 있는 딥러닝 알고리즘인 LSTM 을 사용하였고 Google 사의 MediaPipe Framework 에서 제공하는 Hand Tracking Module 를 이용해 손바닥을 감지하여 손의 특징점을 감지하여 데이터를 수집하였다. 얼굴표정 인식 모델의 학습 데이터는 GitHub 사이트에서 제공하는 데이터를 이용하여 모델을 학습시켰으며, 2014 년 도 ImageNet Large Scale Visual Recognition Challenge(ILSVRC)[4][5] 대회에서 수상을 한 VGG-16 을 사용해 이미지 분류를 할 것이다. VGG-16 은 1400 만개 이상의 데이터셋 이미지로 학습하고, 키보드, 마우스, 연필 등 1000 가지 카테고리로 분류할 수 있으며, 16 개의 레이어로 구성된 컨볼루션 신경망(Convolutional Neural Network, CNN)이다[6].

데이터의 수집과정은 그림 2 과 같이 help, handout, fist, nothing 의 네가지 상황으로 수집하였고 그림 3 에서는 전체 알고리즘을 보여준다. 다음에서 알고리즘에 사용된 각 모듈을 설명한다.

2.1 OpenCV

OpenCV 는 Open Source Computer Vision Library 로 영상 처리에 대한 오픈 소스 라이브러리로, 누구나 이용할 수 있고 딥러닝을 포함한 다양한 분야에서 활용되고 있다[7]. 이 라이브러리는 실시간 이미지 처리에 많이 사용되고 있으며, 모바일 애플리케이션에서 사용되는 얼굴 인식, 문서 인식, 이미지 변환 기능 등에 해당 라이브러리가 광범위하게 사용되고 있다. OpenCV 라이브러리에는 머신러닝 기반의 이미지 추적이 가능한 Haar 특징기반 다단계 분류자가 있다. 이 방법은 2001 년 제안된 기법으로 사람의 얼굴을 픽셀화 한다. 그 후 흑백의 픽셀을 겹쳐 비교하여 해당 패턴을 비교하는 방식을 사용한다. 사람마다 얼굴은 구체적으로는 다르지만, 기본적인 얼굴 생김의 패턴은 모두 비슷하므로 흑백의 명암은 비슷할 것이라는 가정하에 해당 분류자가 제안되어 사용되고 있다[8].

2.2 MediaPipe

수신호인식 학습 데이터 수집 시 사용한 라이브러리는 MediaPipe 라는 다양한 머신러닝 모듈을 제공하는 사이트를 이용했다. MediaPipe 란 구글에서 제공하는 AI 프레임워크로서, 비디오 형식 데이터를 다양한 비전 AI 기능을 파이프 라인 형태로 손쉽게 사용할 수 있도록 제공한다. 모델 개발 및 수많은 데이터 셋을 이용한 학습도 마친 상태로 제공되므로 라이브러리를 불러 간편하게 호출해 사용하면 된다. 기본적인 얼굴인식 외에도 동작인식 등 다양

한 시각적 AI 기능들을 제공한다. 본 연구에서는 MediaPipe 에서 제공한 손가락 인식 파이썬 소스를 수정하여 손 중요 부위의 위치를 추출하고 각 위치의 차분 벡터를 얻어냈다[9].

2.3 LSTM

LSTM 은 RNN 의 한 종류로 Long Short Term Memory 의 약자이다. LSTM 은 기본 RNN 의 단점을 보완하는 모델이다. 기본적인 구조는 기본 RNN 과 동일하지만 각 반복 모듈에서 다른 구조를 가지고 있다. 반복 모듈은 4 개의 layer 가 서로 정보를 주고받는 형태의 구조로 이뤄져 있다. 4 개의 layer 로 구성되어 있기에 RNN 과 비교하면 조금 더 복잡하지만, long sequence 를 다루는 문제에서 뛰어난 성능을 가지고 있다. 해당 모델은 'cell state'라는 개념을 가진다. cell state 는 무언가를 더하거나 제거할 수 있는 능력이 있다. 이 능력은 gate 구조에 의해 제어되어진다. LSTM 은 3 개의 gate(입력게이트, 삭제게이트, 출력게이트)를 가지고 있다. 이 gate 는 cell state 를 보호하고 제어하는 역할을 수행한다[10]. 제스처 인식을 위해 LSTM 모델을 사용했다.

2.4 VGG-16

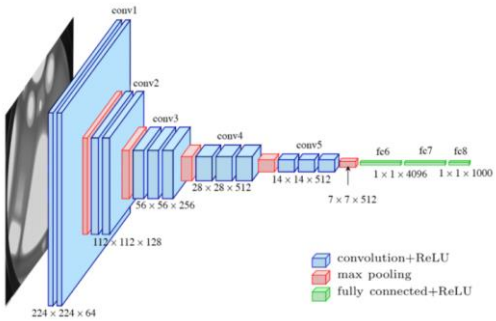

그림 4. The standard VGG-16 network architecture [11]

그림 4 에서 보이는 바와 같이 VGG-16 는 CNN 을 기반으로 둔 모델로서 신경망 모델의 깊이(Layer 수)가 16 개여서 VGG-16 이라고 불린다[11]. 하지만 실제 VGG-16 의 학습데이터의 크기는 224x224x3 였지만 본 실험에서의 데이터의 크기는 48x48x1 의 매우 작고 RGB 가 아닌 Gray 인 점을 감안하여 Overfitting 을 최소화하기 위해서 Block 별로 Dropout 를 많이 넣어서 연구를 진행하였다.

3. 실험 결과 및 분석

본 연구에서 제안하는 VGG16, LSTM 를 통한 위험신호 감지 시스템의 전체적인 프로세스는 기존의 GitHub 사이트에서 제공하는 표정 데이터셋과 실시간 손바닥 좌표 데이터를 기반으로 본 연구의 취지에 맞게 재구성하여 연구를 진행하였다.

3.1 Gesture Classification

LSTM 인풋으로 들어갈 제스처 사진을 얻기 위해서 실시간으로 연속된 30 프레임마다 손바닥의 21 개의 x, y, z 좌표를 MediaPipe 를 통해 받아서 LSTM 모델에 학습을 진행하였다.

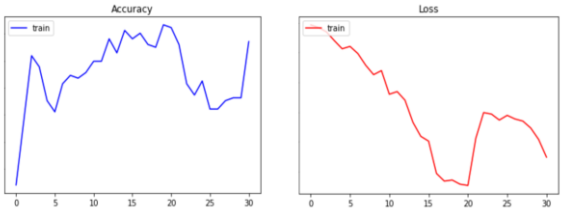

그림 5 LSTM 모델의 정확도와 손실

그림 5 에서 보이는 바와 같이 epoch 를 30 까지 실행하였으나 epoch 20 에서 10 번 연속으로 loss 가 개선되지 않아서 20 번째 epoch 의 정확도가 약 83.33%가 나왔고 검증셋의 대한 정확도는 약 91.67%가 나오게 되었다.

3.2 Emotion Classification

그림 3 과 같이 CNN 을 기반으로 둔 VGG-16 인풋으로 GitHub 에서 표정 데이터를 넣어서 진행을 하였다. 이때 데이터의 사이즈는 48x48 로 통일하였고 데이터에 맞게 Gray 스케일로 변경 후 기존 VGG-16 에서 각각의 Layer 에서 인풋의 사이즈를 고려하여 기존에 VGG-16 의 뉴런 수는 일치하게 가져가되 Dropout 의 값을 블록별로 0.4, 0.4, 0.4, 0.4, 0.7, 0.7 로 설정하여 최대한 Overfitting 이 일어나지 않도록 하였다. 그리고 10 번 연속으로 검증셋의 대한 손실함수가 감소하지 않으면 Learning rate 의 값을 줄여서 더욱 세부적으로 최적화된 값을 찾도록 하였다. 그렇게 해서 감정 정확도는 약 65.7%이고 검증셋에 대한 정확도는 약 74.13%로 나오게 되었다.

3.3 Danger Count

위에 두 모델에서 나온 결과 값을 OpenCV 에 오른쪽 상단에는 감정에 대한 예측 값과 정확도를 표기하고, 왼쪽 상단에는 제스처에 대한 예측 값과 정확도를 표기하여 만약에 Sad 표정일 때 Help 제스처를 10 frame 이상 인식하게 되면 화면에 SOS 신호를 출력하고 Danger Count 를 초기화 함으로써 모델 오류나 잘못된 신호로 인한 SOS 구조 요청을 어느정도 방지하는 효과를 볼 수 있었다.

4. 결론

본 연구에서는 실시간으로 손의 특징점을 찾아주는 MediaPipe Framework 의 Hand Tracking 모듈을 사용하여 손 제스처 인식한 데이터와 표정 영상 데이터를 기반으로 딥러닝 알고리즘 LSTM 과 VGG16 을 적용하여 실시간으로 위험신호 감지와 표정 신호 분류를 진행하였다. 손의 특징점을 찾아 21 개의 Hand Landmark 를 디텍션한 후 NumPy 배열로 변환한 데이터로 Hyperparameter 가 조정하여 LSTM 경우 검증 데이터에 대한 약 91.7%의 Testing 정확도를 산출하였다. 해당 영상 데이터의 표정을 인식하여 VGG16 알고리즘을 병렬 처리하여 표정 분류에서는 약 74.1%에 달하는 Testing 정확도를 산출하였다. 제스처 인식이 표정 인식 분류보다 다소 정확한 분류를 수행하였다. 그럼에도 불구하고 표정 인식 모델을 같이 제공함으로써 이와 유사한 손동작이지만 다른 의미를 가진 경우를 최대한 배제하기 위해 추가해 좀 더 신뢰성 있는 모델을 만들기 위해 노력하였다. 본 연구는 빅데이터라고 불리는 실시간 영상 자료를 모션인식 기반 분석으로 향후 급성장하고 있는 IOT 시장의 흐름에 올라타 웨어러블 센서, 네트워크 객체, 기존 네트워크를 활용해 삶의 다양한 부분에 응용될 것으로 기대된다[10].

감사의 글

본 연구는 2021 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2021R1A2C1008360).

참고문헌

[1] 이미정, “코로나 19 와 젠더폭력: 가정폭력 현황과 대응”, 코로나 19 관련 여성, 가족 분야별 릴레이 토론회, 제 4 차, 61 호, p.01-05, 2021

[2] 하태원, “인공지능 기반 응급상황 예측에 대한 연구”, 2021

[3] 김기덕 외 2 인, “시계열 데이터 분류와 NAS 를 통한 손동작 인식”, 한국컴퓨터정보학회 동계학술대회 논문집, 제 29 권, 제 1 호, 2021

[4] “ImageNET Large Scale Visual Recognition Challenge (ILSVRC). (n.d.).”, http://www.image-net.org/challenges/LSVRC/

[5] Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Li, F., “ImageNet: A large-scale hierarchical image database”, 2009 IEEE Conference on Computer Vision and Pattern Recognition, p.248-255, 2009

[6] 민애리, “컨볼루션 신경망의 이미지 분류를 위한 전이학습 연구”, 2020

[7] 홍두표, “OpenCV 를 활용한 고객 출입 관리시스템 설계 및 구현”

[8] Seo-Jin Hwang, “A Design and Implementation of Mask Wearing Face Detection System by OpenCV”

[9] 김기덕, “RGB 영상 데이터 기반 손동작 인식”, 한국컴퓨터정보학회 하계학술대회 논문집, 제 29 권, 제 2 호, 2021

[10] Aurelien Geron, 『Hands-on Machine Learning with Scikit-Learn, Keras & Tensorflow』, 한빛미디어, p.617-620, 2020

[11] Karen Simonyan & Andrew Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, ICLR, 2015

[12] S. Hochreiter, J. Schmidhuber, “Long Short-Term Memory”, Neural Comput., Vol. 9, 1997