Text classification 방법을 사용한 행동 인식, 손동작 인식 및 감정 인식

김기덕O
O부산대학교 전기전자컴퓨터공학과
e-mail: kimsjpk@hanmail.netO

Action recognition, hand gesture recognition, and emotion recognition using text classification method

Gi-Duk KimO
ODept. of Electricity and Electronic Computer Engineering, Pusan National University

● 요 약 ●

본 논문에서는 Text Classification에 사용된 딥러닝 모델을 적용하여 행동 인식, 손동작 인식 및 감정 인식 방법을 제안한다. 먼저 라이브러리를 사용하여 영상에서 특징 추출 후 식을 적용하여 특징의 벡터를 저장한다. 이를 Conv1D, Transformer, GRU를 결합한 모델에 학습시킨다. 이 방법을 통해 하나의 딥러닝 모델을 사용하여 다양한 분야에 적용할 수 있다. 제안한 방법을 사용해 SYSU 3D HOI 데이터셋에서 99.66%, eNTERFACE' 05 데이터셋에 대해 99.0%, DHG-14 데이터셋에 대해 95.48%의 클래스 분류 정확도를 얻을 수 있었다.

키워드: 행동 인식(action recognition), 손동작 인식(hand gesture recognition), 감정 인식(emotion recognition)

I. Introduction

딥러닝을 통한 컴퓨터 비전 기술은 인간과 컴퓨터, 로봇 간 상호작용(HCI)를 증대시키고 있다[1]. 행동 인식의 경우 운동자세 교정이나 이상행동 검출에 도움을 주며 손동작 인식을 통해 로봇에게 음성보다 쉽고 간단하게 자신의 의사를 전달할 수 있다. 그리고 감정 인식을 통하여 컴퓨터가 인간의 감정 상태에 따른 적절한 피드백을 제공하는데 도움을 줄 수 있다. 사람의 행동, 손동작 및 감정을 인식하려는 방법으로 센서를 활용한 접촉식 방법과 카메라를 이용한 비접촉식 방법이 있는데 카메라를 이용한 비접촉식의 경우 조명이나 배경의 영향을 받아 데이터 처리의 불편함과 정확성이 떨어지는 단점이 있으나 고가의 센서 장비가 필요하지 않으며 카메라로 모든 정보를 처리할 수 있어 간편한 장점이 있다. 이에 관련 기술이 딥러닝과 컴퓨터 비전을 통해 활발히 연구되고 있다. 영상 픽셀을 입력으로 하는 경우 조명의 영향과 카메라의 위치에 따라 추론 결과에 영향을 끼치게 된다. 이를 감소시키고자 카메라 영상에서 스켈레톤을 통한 특정 벡터를 추출하여 사람의 행동 인식[2], 손동작 인식[3] 그리고 감정 인식[4] 연구가 이루어지고 있다. 이와 관련하여 본 논문에서는 pose estimation을 통한 사람의 관절 및 중요 부분에서 추출된 스켈레톤 정보를 사용한 행동 인식, 센서를 사용하여 추출된 손 부위의 스켈레톤 정보를 사용한 손동작 인식, Dlib[5] 라이브러리를 사용하여 얼굴 내 눈과 입 부위에서 스켈레톤 벡터를 추출 후 감정 인식 방법을 제안한다.

II. Preliminaries

1. Datasets

1.1 SYSU 3D HOI dataset

위 데이터 세트[6]는 40명이 각각 12가지 행동에 대하여 자유롭게 행동하고 그 영상을 저장하였다. 행동은 핸드폰을 사용하여 전화를 받는 동작, 가방을 메는 동작 등을 포함한다. 총 480(12 x 40)개의 1.9초에서 21초 사이의 다른 프레임 길이 영상으로 구성되어 있다.

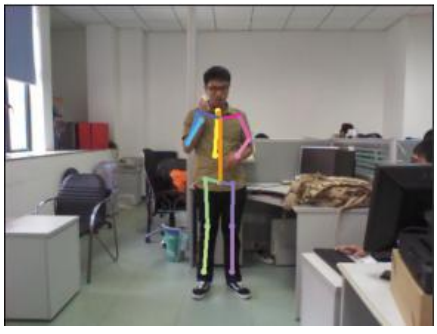

Fig. 1. SYSU dataset 그림

1.2 eNTERFACE'05 Audio-Visual Emotion Database

1,166개의 비디오 데이터로 구성되며 6개의 감정(화, 혐오, 두려움, 기쁨, 슬픔, 놀람)을 표현 클래스를 지닌다. 각 출연자는 14개국의 사람들로 81%는 남성, 19%는 여성으로 구성되어 있다.

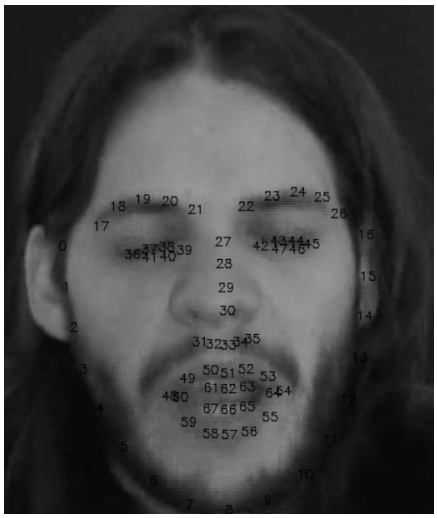

Fig. 2. eNTERFACE' 05 dataset 그림

1.3 DHG-14/28 Dataset

DHG 14/28 데이터셋[8]은 depth 이미지와 그에 대응하는 스켈레톤으로 구성된 손동작 데이터셋이다. 깊이 이미지는 Intel Realsense 카메라로 촬영되고 스켈레톤은 Intel Realsense SDK에 의해 획득하였다. 14개의 손가락 제스쳐에 대해 두 개의 손가락 데이터로 구성되었다.

Fig. 3. DHG-14/28 dataset 그림

III. The Proposed Scheme

SYSU 데이터셋의 경우 EfficientPose[9]를 사용하여 그림 1과 같이 신체의 주요부위와 관절의 포인트를 추출하였다. 그 후 식 1을 사용하여 x, y 벡터 30개를 저장하였다. 데이터 증대를 위하여 전체 프레임에서 40개의 무작위 인덱스를 뽑고 그 후 50개의 프레임의 30개 x, y 벡터를 저장하였다.

$$w^{S_{i,k}} = p_i^s - p_k^s, i \neq k$$

식 1. x, y 벡터 식

수식 1에서 i, k는 신체 주요 지점, s는 frame을 나타낸다. 전체 데이터의 모양은 (19200, 50, 30) 이며 학습 데이터와 테스트 데이터는 4:1로 나누어 학습을 진행하였다. 데이터의 모양이 임베딩 된 문장 벡터의 배치 데이터와 모양이 유사하여 text classification에 사용된 딥러닝 모델을 적용해서 행동 클래스 분류를 진행하였다. textnas[10]에 적용된 Conv1D[11], Transformer[12] 계층과 GRU[13] 층을 결합하여 딥러닝 모델을 구성하였다.

```
model = Sequential()
model.add(Conv1D(filters=64, kernel_size=3, padding='same',
                 activation='relu'))
model.add(MaxPooling1D(pool_size=2))
model.add(Bidirectional(GRU(64, return_sequences=True)))
transformer_block = TransformerBlock(128, 4, 64)
model.add(transformer_block)
model.add(layers.GlobalAveragePooling1D())
model.add(layers.Dropout(0.1))
model.add(Dense(NUM_CLASSES, activation='softmax'))

model.compile("adam", "categorical_crossentropy",
              metrics=["accuracy"])
```

Fig. 4. 학습에 사용된 딥러닝 모델 그림

eNTERFACE' 05 데이터셋[7]의 경우 Dlib 라이브러리를 사용하여 얼굴 중요 부위의 포인트를 추출하고 눈과 입술의 포인트에서 수식 1을 사용하여 x, y 벡터 데이터를 추출하였다. 데이터 증대를 위하여 전체 프레임에서 40개의 무작위 인덱스에서 그 후 30프레임 데이터를 저장하였다. 전체 데이터의 모양은 (51480, 30, 64)이다. SYSU 데이터셋과 같이 학습데이터, 테스트데이터를 4:1로 정하여 학습을 진행하였다. 딥러닝 모델은 SYSU 데이터셋에 적용한 딥러닝 모델과 같다. Dlib 라이브러리에서 얼굴 주요부위 포인트 추출 시 인식을 하지 못한 내용은 0 벡터로, 여러 개의 얼굴로 인식 시 가장 큰 영역을 지닌 부위를 얼굴로 지정하는 예외처리를 하였다. DHG-14/28의 경우 별도의 라이브러리를 통한 벡터 추출 없이 데이터셋 내 skeleton_world.txt에 저장된 66개의 특징(벡터) 데이터를 학습에 사용하였다. 20개의 무작위 인덱스를 뽑고 그 후 20개의 프레임 데이터를 저장하였다. 전체 데이터의 모양은 (56000, 20, 66)이다. SYSU 데이터셋과 같이 학습데이터, 테스트데이터를 4:1로 정하여 학습을 진행하였다. 딥러닝 모델은 SYSU 데이터셋에 적용한 모델과 같다. 학습 시 epoch는 500으로 하여 학습하였다.

Table 1. SYSU 3D HOI 데이터셋에 대한 성능 비교

Method	Accuracy
HON4D[14]	79.22%
MTDA[15]	84.21%
JOULE-SVM[16]	84.89%
CTDAN[17]	98.33%
제안한 방법	99.66%

Table 2. eNTERFACE' 05 데이터셋에 대한 성능 비교

Method	Accuracy
Rashid al[18]	80.27%
Rázuri al[19]	98.00%
제안한 방법	99.00%

Table 3. DHG-14 데이터셋에 대한 성능 비교

Method	Accuracy
Res-TCN[20]	91.1%
STA-Res-TCN[20]	93.6%
ST-GCN[21]	92.7%
DG-STA[22]	94.4%
제안한 방법	95.48%

IV. Conclusions

본 논문에서는 스켈레톤 데이터에 대해서 앙상블이 아닌 하나의 딥러닝 모델을 사용하여 행동 인식, 손동작 인식 및 감정 인식의 다양한 분야에 적용하였다. Transformer를 사용함으로써 학습 속도를 높이고 모델의 크기를 줄였으며 클래스 분류 성능 또한 높일 수 있었다. eNTERFACE' 05 dataset에서 batch size를 128로 하였을 때 1 epoch 당 22초의 학습 시간을 보였다. 학습에 사용된 컴퓨터 사양은 i5-9300, GTX 1660 ti이다. 모델 크기의 경우 1.8MB이다. 사용한 딥러닝 라이브러리는 tensorflow 2.3.0, Cuda 10.1이다.

REFERENCES

[1] Guleryuz, Onur G., and Christine Kaeser-Chen. "Fast Lifting for 3D Hand Pose Estimation in AR/VR Applications." 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018.

[2] Si, Chenyang, et al. "Skeleton-based action recognition with hierarchical spatial reasoning and temporal stack learning network." Pattern Recognition 107 (2020): 107511.

[3] Chen, Yuxiao, et al. "Construct dynamic graphs for hand gesture recognition via spatial-temporal attention." arXiv preprint arXiv:1907.08871 (2019).

[4] Ngoc, Quang Tran, Seunghyun Lee, and Byung Cheol Song. "Facial Landmark-Based Emotion Recognition via Directed Graph Neural Network." Electronics 9.5 (2020): 764.

[5] King, Davis E. "Dlib-ml: A machine learning toolkit." The Journal of Machine Learning Research 10 (2009): 1755-1758.

[6] Hu, Jian-Fang, et al. "Jointly learning heterogeneous features for RGB-D activity recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[7] Martin, Olivier, et al. "The eNTERFACE'05 audio-visual emotion database." 22nd International Conference on Data Engineering Workshops (ICDEW'06). IEEE, 2006.

[8] Dynamic Hand Gesture Recognition using Skeleton-based Features , Quentin De Smedt, Hazem Wannous and Jean-Philippe Vandeborre, 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)

[9] Groos, Daniel, Heri Ramampiaro, and Espen Ihlen. "EfficientPose: Scalable single-person pose estimation." arXiv preprint arXiv:2004.12186 (2020).

[10] Wang, Yujing, et al. "TextNAS: A Neural Architecture Search Space Tailored for Text Representation." AAAI. 2020.

[11] Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).

[12] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017): 5998-6008.

[13] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

[14] Oreifej, Omar, and Zicheng Liu. "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences." Proceedings of the IEEE conference on computer vision and pattern recognition. 2013.

[15] Zhang, Yu, and Dit Yan Yeung. "Multi-task learning in heterogeneous feature spaces." Proceedings of the National Conference on Artificial Intelligence. 2011.

[16] Hu, Jian-Fang, et al. "Jointly learning heterogeneous features for RGB-D activity recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[17] Wang, Pichao, et al. "Cooperative training of deep aggregation networks for RGB-D action recognition." arXiv preprint arXiv:1801.01080 (2017).

[18] Rashid, Munaf, S. A. R. Abu-Bakar, and Musa Mokji. "Human emotion recognition from videos using spatio-temporal and audio features." The Visual Computer 29.12 (2013): 1269-1275.

[19] Rázuri, Javier G. "Decision-making content of an agent affected by emotional feedback provided by capture of human's emotions through a Bimodal System." International Journal of Computer Science Issues 12.6

(2015).

[20] Hou, Jingxuan, et al. "Spatial-temporal attention res-TCN for skeleton-based dynamic hand gesture recognition." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[21] Yan, Sijie, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition." arXiv preprint arXiv:1801.07455 (2018).

[22] Chen, Yuxiao, et al. "Construct dynamic graphs for hand gesture recognition via spatial-temporal attention." arXiv preprint arXiv:1907.08871 (2019).