Grid search와 Transformer를 통한 그룹 행동 인식

김기덕O, 이근후*
*(주)쓰리아이퓨처,
O(주)쓰리아이퓨처
e-mail: lghoo@naver.com*, kimsjpk@naver.comO

Group Action Recognition through Grid search and Transformer

Gi-Duk KimO, Geun-Hoo Lee*
*3Ifuture,
O3Ifuture

● 요 약 ●

본 논문에서는 그리드 탐색과 트랜스포머를 사용한 그룹 행동 인식 모델을 제안한다. 추출된 여러 사람의 스켈레톤 정보를 차분 벡터, 변위 벡터, 관계 벡터로 변환하고 사람별로 묶어 이를 TimeDistributed 함수에 넣고 풀링을 한다. 이를 트랜스포머 모델의 입력으로 넣고 그룹 행동 인식 분류를 출력하였다. 논문에서 3가지 벡터를 입력으로 하여 합치고 트랜스포머 계층을 거친 모델과 3가지 벡터를 입력으로 하고 계층적으로 트랜스포머 모델을 거쳐 행동 인식 분류를 출력하는 두 가지 모델을 제안한다. 3가지 벡터를 합친 모델에서 클래스 분류 정확도는 CAD 데이터 세트 96.6%, Volleyball 데이터 세트 91.4%, 계층적 트랜스포머 모델은 CAD 데이터 세트 96.8%, Volleyball 데이터 세트 91.1%를 얻었다

키워드: 그룹 행동 인식(group action recognition), 트랜스포머(transformer), 계층적 구조(hierarchical structure)

I. Introduction

CCTV를 통한 이상 상황 탐지의 경우 사람의 집중력 감소로 이를 극복할 수 있는 행동 인식을 통한 이상 상황 인식 시스템이 필요하다. 철도 역사나 키즈 까페의 CCTV 영상에서는 다수의 사람이 등장하며 이에 대해 그룹 행동 인식을 통하여 사고에 대해서 즉각적으로 대응함으로써 사회 안전망 구축에 기여할 수 있다.[1] CAD 데이터 세트와 Volleyball 데이터 세트에서 동영상 클립 데이터에서 각 사람 개인의 행동과 그룹의 행동 클래스를 라벨링 하였는데 본 논문에서는 개인별 행동 라벨링은 사용하지 않고 (데이터 배치 수, 사람 수, 프레임 수, 특징)의 4차원 데이터를 입력으로 TimeDistributed 함수와 풀링 과정을 통해 3차원 벡터로 변환한 후 트랜스포머 계층을 거친 그룹 행동 인식 알고리즘을 제안한다. RGB 데이터를 입력으로 하는 경우 모델에서 데이터 크기의 증가로 학습 시간 증가 및 모델의 크기가 증가한다. pose estimation을 통하여 사람의 스켈레톤 데이터를 추출함으로써 입력의 크기를 줄이고 적은 학습 자원에서 효율적으로 딥러닝 네트워크 모델 학습이 가능하다.

본 논문의 구성은 2장에서 관련 연구로 그룹 행동 인식 관련 연구 발달 과정을 살펴보고 3장에서 논문에서 사용한 모델에 대해서 설명하고 모델의 인식 정확도 결과를 기술한다. 그리고 4장에서는 결론에 관해서 기술한다.

II. Preliminaries

1. Related works

1.1 그룹 행동 인식

RNN(Recurrent Neural Network) 기반 그룹 행동 인식은 2016년 Ibrahim 등이 Volleyball 데이터 세트를 공개하여 이를 대상으로 그룹 행동 인식 알고리즘이 연구되었다. Ibrahim 등은 개인별로 RNN 모델을 거치고 이를 합쳐 다시 RNN을 적용해 그룹의 행동 인식 결과를 얻는 계층적 RNN 모델[2]을 사용하였다. 그 후 Zappardino 등은 라벨링 된 사람의 영역에서 pose estimation 결과를 추출하고 추출된 스켈레톤 이미지를 입력으로 하는 네트워크[3]를 제안하였다. Zhou 등은 스켈레톤 벡터와 함께 사람 간 상관관계 등을 계층적으로 학습하는 모델[4]을 제안하였다.

1.2 트랜스포머

RNN의 경우 시계열 데이터를 처리하기 위한 모델로 자연어 처리 등 여러 분야에 사용되어 왔으나 긴 대화는 장기 의존성 문제가 발생한다. 트랜스포머[5]는 기계번역을 위해 고안된 모델로 MLP(Multi-Layer-Perceptron)와 Attention 방법을 사용해 RNN이 가지는 장기 의존성 문제를 해결하였다. 트랜스포머는 인코더와 디코더로 구성되며 인코더, 디코더 내 각 층의 계층은 단어 사이의 관계를 계산하는 멀티 헤드 어텐션으로 구성된다. 계층에서 query, key, value 벡터의 연산을 통해 벡터 간 다양한 관계에 대해 학습이 가능하다.

III. The Proposed Scheme

본 논문에서는 그룹 행동 인식 데이터 세트로 CAD 데이터 세트[11]와 Volleyball 데이터 세트[2]를 사용하였다. 데이터 세트의 라벨링 데이터로 사람의 bounding box 영역, 개인행동의 클래스 정보, 그룹 행동의 클래스 정보를 담고 있다. 본 논문에서는 Zhou의 깃허브[6]에 올려진 CAD 데이터 세트와 Volleyball 데이터 세트에서 추출한 스켈레톤 특징을 가공하여 Luvison 등이 제안[7]한 차분 벡터, 변위 벡터, 관계 벡터로 변형하고 딥러닝 모델의 입력으로 사용하였다. 변형에 사용한 식은 수식 1, 2와 같다. 관계 벡터는 표 1과 같이 상대 포인트 좌표에서 대응하는 포인트 지점 좌표를 뺀 벡터이다. TimeDistributed 함수를 사용하여 (데이터 배치 수, 사람의 수, 프레임 수, 특징)의 4차원 데이터를 입력으로 받아 Conv1D, MaxPooling, GRU, GlobalAveragingPooling을 거쳐 3차원 데이터로 변형하였다. Zappardino 등이 제안한 딥러닝 모델에 영감을 받아 가공한 3가지 벡터를 입력으로 받고 위에서 언급한 방법으로 3차원 벡터로 변형하고 Concatenate를 통해 합친 후 GRU, Attention을 거친 모델을 사용하였다. 모델은 그림 1과 같다.

$$v_i^s = \frac{P_i^{s+1} - P_i^s}{\triangle T} | 1 < s < \Upsilon$$

수식. 1. 변위 벡터 식

$$w_{i,k} = p_i^s - p_k^s | \ i \neq k$$

수식. 2. 차분 벡터 식

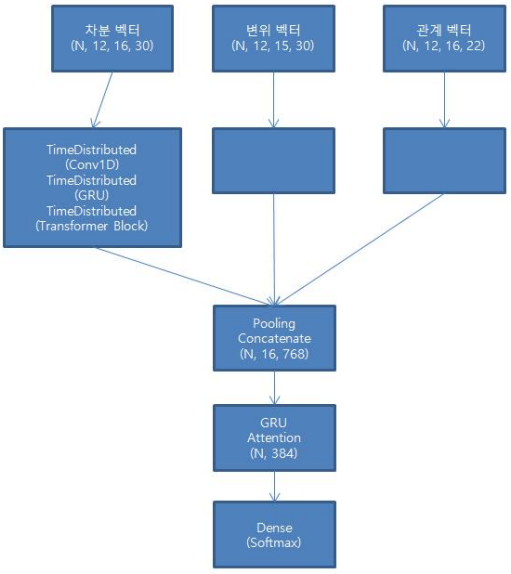

Fig. 1. Concatenate 모델 구조

Table 1. 관계 벡터 포인트 대응 쌍

신체 부위	상대 포인트
머리, 왼쪽 손, 오른쪽 손	허리
머리, 왼쪽 손, 왼쪽 발	오른쪽 엉덩이
목, 오른쪽 손, 오른쪽 발	왼쪽 엉덩이
왼쪽 손, 오른쪽 손	머리

그리고 Zhou 등의 모델에 영감을 받아 층 별로 3가지 벡터를 입력으로 넣고 위의 방법에 따라 3차원 벡터로 변형 후 트랜스포머 모델을 거치는 모델을 사용하였다. 성능 강화를 위해 출력된 3차원 벡터를 skip connection으로 연결하였고 이에 의한 특징을 줄이기 위해 1x1 Conv2D 계층을 거쳐 특징의 수를 줄였다. 1, 2, 3층에 각 벡터를 배치하고 트랜스포머 1개 또는 2개 계층을 거치는 모델을 구성하고 그리드 탐색을 통해 48($3!* 2^3$)개의 파라미터를 가지고 모델을 학습하여 테스트 데이터 최고 정확도의 모델을 탐색하였다. 모델은 그림 2와 같다.

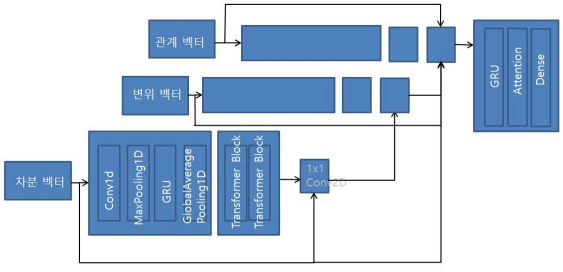

Fig. 2. 계층적 트랜스포머 구조

CAD 데이터 세트의 경우 train, test 데이터 세트 비율을 4:1로 정하고 데이터를 섞어서 학습하였다. volleyball 데이터 세트의 경우 다른 논문에서 적용한대로 train, validation, test 영상을 학습 이전에 분리하여 학습을 진행하였다. 학습에 사용한 코드는 블로그에 공개하였다.(https://blog.naver.com/kimsjpk/223000102967)

Table 2. CAD 데이터 세트 성능 비교

알고리즘	정확도
AT[8]	92.8%
SACRF[9]	95.2%
GroupFormer[10]	96.3%
COMPOSER[4]	96.2%
제안한 모델 1	96.6%
제안한 모델 2	96.8%

Table 3. Volleyball 데이터 세트 성능 비교

알고리즘	정확도
Zappardino et al[3]	91.0%
SACRF	95.0%
GroupFormer	95.7%
COMPOSER	94.6%
제안한 모델 1	91.4%
제안한 모델 2	91.1%

Fig. 3. 흐릿한 이미지 예시

IV. Conclusions

본 논문에서는 그룹 행동 인식을 위해 스켈레톤 특징을 차분, 변위, 관계 벡터로 변형하고 각 사람을 시간 순으로 배치하여 4차원 데이터를 학습할 방법을 제안하였다. 이를 통해 CAD 데이터 세트에서 SOTA(State Of The Art)를 달성하였다. Volleyball 데이터 세트의 경우 그림 3과 같이 흐릿한 이미지에서 사람의 올바른 스켈레톤 특징을 추출하지 못해서 성능이 낮게 나온 것으로 추측한다. 앞으로 흐릿한 이미지에서 올바른 스켈레톤 특징을 추출하는 알고리즘을 적용하면 더욱 높은 그룹 행동 인식 정확도를 얻을 수 있을 것이다.

REFERENCES

[1] TANG, Yansong, et al. Mining semantics-preserving attention for group activity recognition. In: Proceedings of the 26th ACM international conference on Multimedia. 2018. p. 1283-1291.

[2] IBRAHIM, Mostafa S., et al. A hierarchical deep temporal model for group activity recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 1971-1980.

[3] ZAPPARDINO, Fabio, et al. Learning group activities from skeletons without individual action labels. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021. p. 10412-10417.

[4] ZHOU, Honglu, et al. COMPOSER: Compositional Reasoning of Group Activity in Videos with Keypoint-Only Modality. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV. Cham: Springer Nature Switzerland, 2022. p. 249-266.

[5] VASWANI, Ashish, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.

[6] https://github.com/hongluzhou/composer

[7] LUVIZON, Diogo Carbonera; TABIA, Hedi; PICARD, David. Learning features combination for human action recognition from skeleton sequences. Pattern Recognition Letters, 2017, 99: 13-20.

[8] GAVRILYUK, Kirill, et al. Actor-transformers for group activity recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. p. 839-848.

[9] PRAMONO, Rizard Renanda Adhi; CHEN, Yie Tarng; FANG, Wen Hsien. Empowering relational network by self-attention augmented conditional random fields for group activity recognition. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. Springer International Publishing, 2020. p. 71-90.

[10] LI, Shuaicheng, et al. Groupformer: Group activity recognition with clustered spatial-temporal transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. p. 13668-13677.

[11] CHOI, Wongun; SHAHID, Khuram; SAVARESE, Silvio. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In: 2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops. IEEE, 2009. p. 1282-1289.