# 가려짐이 있는 신체의 골격 추출을 위한 Convolutional Recurrent Pose Machine

심재규[1], 김원우[1], 박광현[1]
[1] 광운대학교

# Convolutional Recurrent Pose Machine for Skeleton Extraction of the Body with Occlusion

Sim Jae-Gyu[1], Kim Won-Woo[1], Park Kwang-Hyun[1]
[1]Kwangwoon University
e-mail: worb1605@kw.ac.kr, asd130394@gmail.com, akaii@kw.ac.kr

**요 약**

실시간으로 입력되는 영상에서 신체의 골격을 추출하기 위한 기존의 Pose Machine들은 장애물에 의해 신체의 일부가 가려진 경우, 가려짐이 있는 상태로 컨피던스 맵을 출력하기 때문에 일부 관절 정보가 누락된 컨피던스 맵을 출력하게 된다. 이러한 문제를 해결하기 위해 본 논문에서는 현재 프레임의 입력 영상이 다음 프레임의 입력 영상에 영향을 끼치는 순환 구조의 네트워크를 제안하여 가려진 신체가 있는 입력 데이터에 대해서도 누락된 정보가 없는 컨피던스 맵을 추출할 수 있도록 하였다.

## 1. 서론

영상 데이터에서 신체 골격 구조를 추출하는 것은 신체를 구성하는 각 관절의 위치를 이미지 상의 2차원 좌표로 표현한 컨피던스 맵(Confidence map)을 추출하는 것을 의미한다[1].

최근 학습을 통해 이미지에서 신체의 골격 구조 추출하는 CPM(Convolutional Pose Machine) 네트워크가 발표된 바 있다.

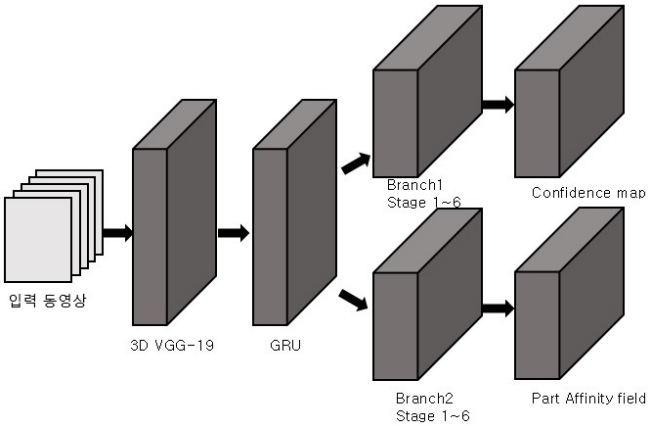
이미지 데이터에서 신체의 모든 부위가 가려진 부분 없이 잘 표현되어 있다면, 정상적으로 컨피던스 맵을 추출할 수 있다. 하지만 신체의 일부가 장애물에 의해 가려져 있다면, 한 장의 이미지에서 컨피던스 맵을 추출하는 CPM 방식으로는 기대하는 수준의 컨피던스 맵을 얻기 힘들다.

이러한 한계점을 극복하고자 본 논문에서는 순환 구조를 통하여 시간 순서로 연속된 입력 데이터를 받아들이는 CRPM(Convolutional Recurrent Pose Machine) 네트워크를 제안한다.

## 2. 본론

### 2.1 3D 합성곱 VGG-19

CPM에서 VGG-19 네트워크의 역할은 컨피던스 맵과 어피니티 필드(Affinity field)를 구하기 위한 특징 맵을 추출하는 것이다. 본 논문에서 제안하는 네트워크는 동영상 데이터를 학습 데이터로 사용하는데, 동영상 데이터를 배치 구조로 입력하여 총 5차원의 데이터를 입력 데이터로 사용한다. 이러한 이유로 CPM에서의 VGG-19는 2D 합성곱 네트워크가 적용되지만, 본 논문에서는 [그림 1]과 같이 5차원의 데이터를 합성곱 할 수 있는 3D 합성곱 VGG-19 네트워크를 통하여 특징 맵을 추출하였다[2].




[그림 1] CRPM(Convolutional Recurrent Pose Machine) 네트워크의 구조

### 2.2 GRU

입력되는 동영상 데이터의 시간적 특성을 고려하기 위해 3D 합성곱 VGG-19 네트워크를 통과한 특징 맵을 순환 구조를 가진 GRU 셀의 입력 데이터로 사용하였다[3].

### 2.3 컨피던스 맵과 어피니티 필드

컨피던스 맵은 신체의 각 관절들이 존재할 확률과 좌표를 2D 맵 형태로 표현한 것이다. 어피니티 필드는 신체의 각 뼈마디들이 존재할 확률과 위치를 2D 맵 형태로 표현한 것이다. CPM에서 컨피던스 맵과 어피니티 필드를 추출하기 위한 2D 합성곱 스테이지

[표 1] 신체의 가려짐이 있는 동영상 데이터에 대한 CPM 결과와 CRPM 결과의 mAP

| Method | Hea | Sho | Elb | Wri | Hip | Kne | Ank | mAP |
|---|---|---|---|---|---|---|---|---|
| CPM | 100.0 | 73.3 | 96.6 | 93.33 | 90.0 | 50.0 | 50.0 | 79.03 |
| Ours | 100.0 | 100.0 | 100.0 | 96.66 | 96.66 | 96.66 | 90.0 | 97.14 |

를 본 논문에서는 5차원의 동영상 데이터를 입력으로 받기 위해 3D 합성곱 스테이지로 구현하였다.

네트워크의 손실 함수는 식 (1)과 같이 컨피던스 맵과 어피니티 필드의 평균 제곱근 오차의 합을 통하여 구하였다.

$$f = \sum_{t=1}^{T} (f_S^t + f_L^t) \quad (1)$$

### 2.4 실험

입력 데이터는 유튜브 동영상 중에서 안무 동영상을 학습 데이터로 활용하였다[4].

네트워크 학습 시에는 30 프레임 길이의 동영상 935개를 사용하였고, 동영상 속 인물들의 관절 데이터는 CPM을 통과시켜 얻은 관절 정보를 검수하여 사용하였다. 학습률은 학습 스텝 단위당 학습률을 조정하는 방식으로 학습을 진행하였다.

### 2.5 평가

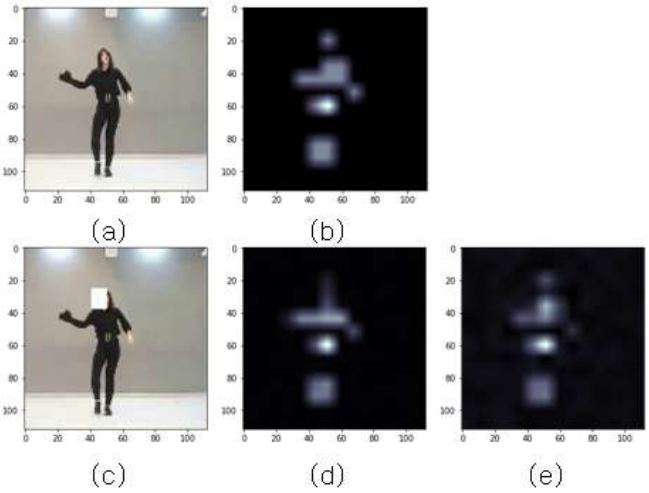
본 논문에서 제안한 네트워크의 성능을 평가하기 위해 신체의 일부가 가려진 동영상을 입력 데이터로 사용하였다. 동영상의 피사체는 13 프레임까지는 신체의 일부가 가려지지 않았고, 14 프레임부터 30 프레임까지 신체의 일부가 가려진다.

CPM의 경우 신체의 일부가 가려진 영상 데이터에 취약한 반면, 실시간으로 입력된 이미지 데이터를 순차적으로 받아들여 처리하는 CRPM에서는 [그림 2]와 같이 CPM 대비 가려짐이 없을 때의 컨피던스 맵에 더욱 가까운 모습을 나타내고 있다.

정량적인 평가를 위해 PCKh 문턱 값에 기반한 관절의 mAP(mean Average Precision) 방법을 사용하여 CPM의 결과[5]와 CRPM의 결과를 [표 1]과 같이 비교하였다.

## 3. 결론

본 논문에서는 가려짐이 있는 영상 데이터의 골격 추출에 대한 기존 연구들의 한계점을 해결하기 위해 시계열적 순서를 가진 동영상 데이터를 입력으로 받는 순환 구조의 네트워크를 제안하였다. 가려짐이 없는 신체의 영상 데이터에 대한 결과는 CPM과 CRPM 모두 큰 차이가 없었지만, 신체의 일부가 가려진 영상 데이터에서는 CRPM의 정확도가 CPM 대비 높은 것을 알 수 있다.




[그림 2] (a) 가려짐 없는 입력 이미지 (b) 가려짐 없는 입력 이미지의 컨피던스 맵 (c) 가려짐 있는 입력 이미지 (d) 가려짐 있는 이미지의 CPM 컨피던스 맵 (e) 가려짐 있는 이미지의 CRPM 컨피던스 맵

## 참고문헌

[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition(CVPR), 2017.

[2] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” Proc. of the IEEE Int. Conf. on Computer Vision(ICCV), pp. 4489-4497, 2015.

[3] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” Proc. of Conf. on Empirical Methods in Natural Language Processing(EMNLP), pp. 1724-1734, 2014.

[4] https://www.youtube.com/user/ae2chijun

[5] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, “DeepCut: Joint subset partition and labeling for multi person pose estimation,” Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition(CVPR), 2016.