다중프레임 관계 신경망을 이용한 3 차원 사람 자세 추정

박성헌, 이지은, 곽노준
서울대학교 융합과학기술대학원 컴퓨터 지능 및 패턴인식 연구실
sungheonpark@snu.ac.kr, mokona85@ajou.ac.kr, nojunk@snu.ac.kr

요 약

본 논문에서는 다중프레임 관계 신경망을 이용한 RGB 영상에서의 3 차원 사람 자세 추정 방법을 제안한다. 이를 위해 기존 관계 신경망에 다중 프레임의 입력을 받아들이도록 디자인하여 기존 단일 프레임 입력의 경우에 비해 성능을 개선하였다.

1. 서론

RGB 영상에서의 사람 자세 추정은 제스처 인식, 행동 인식 등 수많은 어플리케이션으로 연결되는 중요한 문제 중 하나이다. 특히, 최근 많은 연구가 진행된 2 차원 자세 추정에 비해 3 차원 자세 추정은 아직 성능 개선의 여지를 많이 남겨 놓고 있다.

본 논문에서는 여러 장의 연속된 RGB 영상으로부터 사람의 3 차원 자세를 추정하는 알고리즘을 제시한다. 앞서 발표된 관계 신경망 [1] (Relational networks)를 이용한 단일 영상 3 차원 자세 추정 방법[2]을 확장하여 다중프레임의 영상을 이용함으로써 정확도를 향상시켰다.

2. 다중 프레임 관계 신경망을 이용한 3 차원 사람 자세 추정 프레임워크

2.1 관계 신경망

신경망에서 관계 모듈 [1]은 한 쌍의 객체의 관계를 학습하기 위한 네트워크 구조로 주로 시각적 질의 응답(Visual question answering)문제를 풀기 위해 처음 제안되었다. 관계 모듈 네트워크 구조는 다음과 같은 식으로 표현된다.

$$\mathrm{RN}(\mathbf{O}) = f\sum_{(i,j)} g(o_i, o_j)$$

f, g는 각각 네트워크를 나타내는 함수를 뜻하고 $\mathbf{O} = \{o_1, o_2, \cdots, o_n\}$는 객체로 주로 네트워크로 추출한 특징 등이 입력으로 사용된다. 따라서 위 식은 객체 간의 모든 가능한 쌍 (i,j)에 대해 특징을 각각 학습하고 그 합을 관계형 특징으로 사용하게 된다.

2.2 다중 프레임 관계 신경망을 이용한 3 차원 사람 자세 추정 방법

본 논문에서 사람의 3 차원 자세는 기존 연구를 따라서, 사람의 주요 관절을 17 개로 나누고 17 개 관절의 3 차원 위치를 추정하는 것으로 정한다. 입력으로는 RGB 영상에서 추출한 2 차원 관절 위치를 사용하고, 네트워크의 출력으로 3 차원 관절 위치를 추정하게 된다. 3 차원 관절 위치는 기준 관절(엉덩이 부근)과 상대적인 3 차원 좌표를 추정하도록 하였다. 2 차원, 3 차원 관절은 모두 벡터 형태로 변환되어 네트워크에 사용된다.

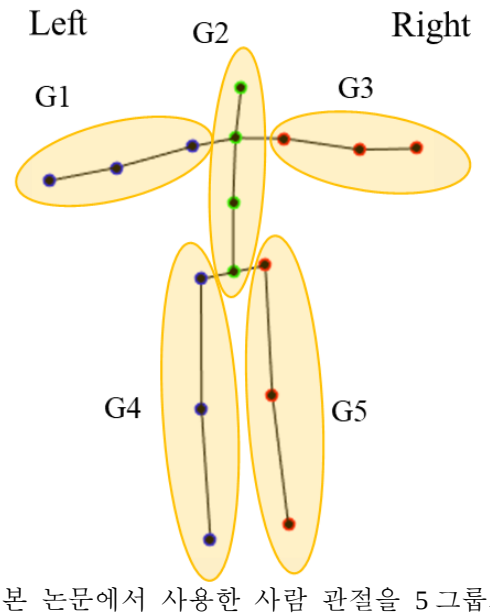


그림 1 본 논문에서 사용한 사람 관절을 5 그룹으로 분할하는 방법을 나타낸 그림.

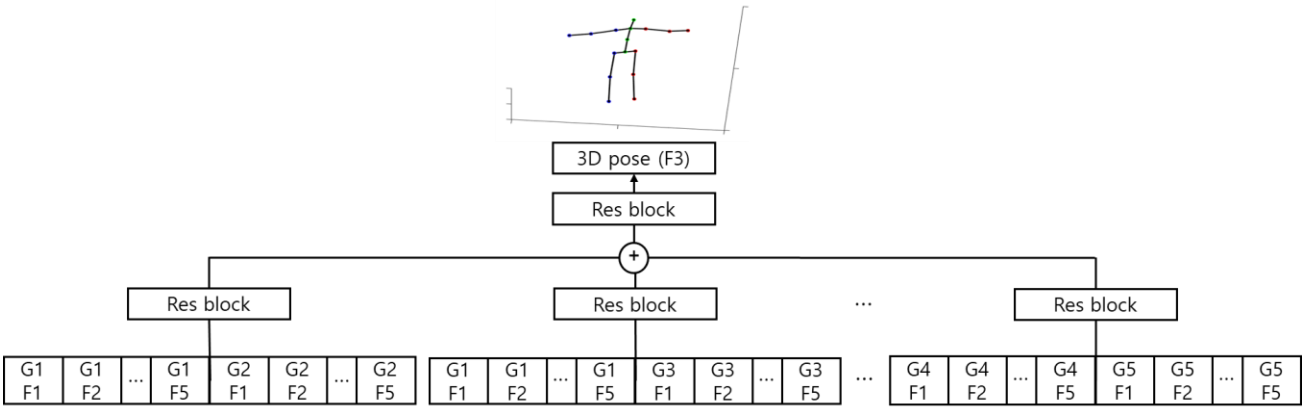


그림 2 다중프레임 관계 신경망 3 차원 사람 자세 추정 네트워크 구조.

관계 신경망을 사람 자세 추정에 활용하기 위해 우선 여러 관절을 하나의 그룹으로 묶어 몇 개의 유의미한 그룹으로 인체의 관절을 나누었다. 본 논문에서는 [2]에서 사용한 그룹화 방법을 따랐다. 그림 1 에서 볼 수 있듯이 사람의 신체를 팔, 다리, 몸통으로 나누고 그에 해당하는 관절들을 하나의 그룹으로 포함해 총 5 개의 그룹을 만들게 된다.

각 그룹에 속한 관절의 개수가 3 개 또는 4 개로 일정하지 않기 때문에, 본 논문에서 사용된 관계 신경망의 경우 관계 특징을 추출하는 함수 g가 파라미터 공유를 하지 않게 된다. 즉, 각각의 (i,j)에 대해 서로 다른 g_{ij}가 존재한다.

단일 영상의 2 차원 관절 위치를 활용하는 것보다 앞뒤 프레임의 관절 위치를 함께 고려하면 부정확한 2 차원 관절 검출에 관해서도 어느 정도 강인한 3 차원 자세 추정이 가능하다. 본 논문에서는 앞뒤 2 프레임씩 총 5 프레임의 2 차원 관절 위치를 활용해 3 차원 자세를 추정하였다.

구체적으로, 그림 2 와 같은 구조로 네트워크를 디자인하게 된다. 기본 구조는 [2]와 동일하나 입력으로 다중프레임의 2 차원 관절 위치를 넣게 된다. 이를 수식으로 나타내면 다음과 같다.

$$\mathrm{S_{3D}}(S_{2D}) = f\left(\frac{1}{n_p}\sum_{(i,j)} g_{ij}(G_{i1}, G_{i2}, \cdots, G_{if}, G_{j1}, G_{j2}, \cdots, G_{jf})\right)$$

여기서 n_p는 가능한 그룹 쌍의 총 개수 (본 논문에서는 $n_p = 10$), $\mathrm{S_{3D}}, S_{2D}$는 각각 사람의 3 차원, 2 차원 자세를 의미한다. G_{ik}는 i번째 그룹에 속하는 관절의 k번째 프레임의 값을 의미한다. 따라서 g_{ij}는 i번째 그룹과 j번째 그룹에 속한 다중 프레임의 관절들을 모두 연결한 벡터를 입력으로 받아 관계 특징을 출력으로 내게 된다. 그 이후 추출된 관계 특징들의 평균이 f에 해당하는 네트워크로 들어가게 되고 최종적으로 3 차원 관절의 좌표가 출력으로 나오게 된다. 출력으로 나오는 관절은 단일 영상에 해당하는 것으로, 입력으로 사용된 5 프레임 중 가운데에 해당하는 프레임의 3 차원 좌표를 추정하게 된다.

2.3 구현 상세

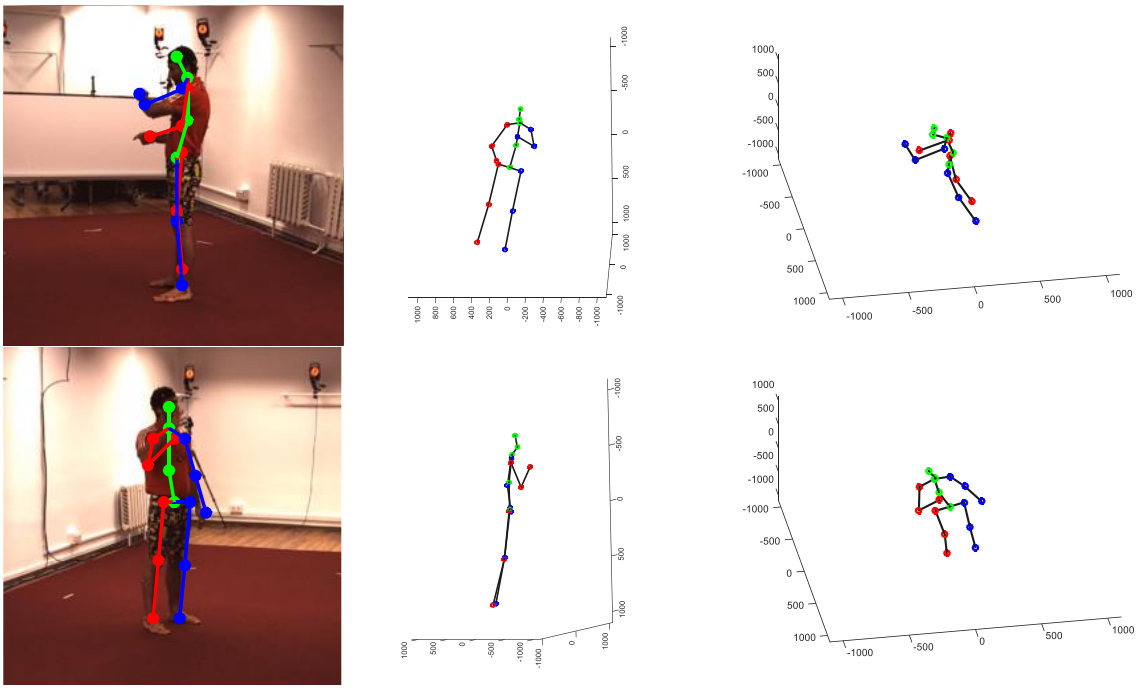
본 논문에서는 RGB 영상을 우선 데이터셋에서 제공하는 사람의 바운딩 박스로 사람 중심으로 자른 후, 256×256 크기로 리사이즈하였다. 리사이즈된 이미지에서 2 차원 좌표를 검출하기 위해 기존의 다층 모래시계 네트워크 [3]을 사용하여 컨볼루션 신경망을 통해 각 관절의 2 차원 위치를 검출하였다. 검출된 2 차원 관절 위치는 다중 프레임 관계 신경망의 입력으로 사용된다.

관계 신경망의 구조는 g_{ij}에 해당하는 관계 특징 추출 네트워크 부분의 경우 1024 차원의 완전 연결 (fully connected) 네트워크로, [4]에서 제안된 residual module 의 구조에 dropout [5]가 추가된 형태로 이용하였다. 최종 3 차원 자세를 추정하는 f 네트워크의 경우 g_{ij}와 같은 구조에 2048 차원의 완전 연결 네트워크를 사용하였다. Dropout 의 경우 g_{ij}에서는 0.25 의 확률, f에서는 0.5 의 확률을 사용하였다.

네트워크 학습에는 Adam optimizer [6]이 사용되었으며 배치 사이즈는 128 로 총 20 만회의 iteration 이 사용되었다. 학습 비율 (learning rate) 의 경우 처음 0.001 에서 4 만회 iteration 마다 0.5 씩 곱해지도록 했다.

3. 실험 결과 및 분석

본 논문에서는 가장 많이 사용되고 있는 3 차원 사람 자세 추정 데이터셋인 Human 3.6m [7]을 사용하였다. 해당 데이터셋은 총 15 개의 사람의 다양한 동작을 영상과 함께 모션캡쳐 시스템을 이용한 3 차원 ground truth 위치를 제공한다. 앞선 논문들의 연구를 따라서, 7 명의 사람 중 5 명 (S1, S5, S6, S7, S8)에 해당하는 사람을 학습 데이터로 사용하고 나머지 2 명 (S9, S11)에 해당하는 사람을 테스



입력 영상과 2 차원 입력 3 차원 추정 자세 - 뷰 1 3 차원 추정 자세 - 뷰 2

그림 3 Human 3.6m 정성적 실험 결과 예시

트 데이터로 사용했다. 각 영상을 프레임별로 늘어놓게 되면 총 약 150 만장의 학습 데이터와 50 만장의 테스트 데이터로 이루어진다.성능의 평가는 평균 관절 위치 오차 (mean per-joint position error)로 측정된다. 이는 먼저 추정된 3 차원 자세와 실제 3 차원 자세의 기준 관절을 일치시킨 뒤, 각 관절의 3 차원 거리 차이를 평균한 것이다.

표 1 에 제안된 다중프레임 관계 신경망의 성능과 최근 발표된 3 차원 자세 추정 방법들과의 성능을 비교하였다. 최근 발표된 논문들보다 좋은 성능을 보여주고 있으며, 특히 [2]보다 개선된 성능을 보임으로써 다중 프레임을 활용할 때의 장점을 증명하였다.

방법	평균 관절 위치 오차 (㎜)
Pavlakos et al.[8]	71.9
Tekin et al. [9]	69.7
Zhou et al. [10]	64.9
Martinez et al. [11]	62.9
Fang et al. [12]	60.4
Park et al. [2]	59.0
Yang et al. [13]	58.6
제안된 방법	**56.9**

표 1 Human 3.6m [7] 데이터셋의 알고리즘별 평균 관절 위치 오차

다음으로, Human 3.6m 데이터셋의 몇 가지 영상에 대해 정성적인 실험결과를 그림 3 에 보였다. 다양한 각도에서 3 차원 자세를 잘 추정하는 것을 확인할 수 있다.

4. 결론

본 논문에서는 다중 프레임을 입력으로 이용한 관계 신경망을 제시하여 단일 프레임에 비해 향상된 성능을 내는 3 차원 자세 추정 알고리즘을 제시하였다. 본 연구는 향후 가상현실, 제스처 인식, 행동 인식 등 다양한 분야에 활용될 것으로 기대한다.

감사의 글

본 연구는 한국연구재단의 차세대정보컴퓨팅기술개발사업에 의해 진행되었음 (2017M3C4A7077582).

참고문헌

[1] Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017).

A simple neural network module for relational reasoning. In Advances in neural information processing systems (pp. 4967-4976).

[2] Park, S., & Kwak, N. (2018). 3D Human Pose Estimation with Relational Networks. arXiv preprint arXiv:1805.08961.

[3] Newell, A., Yang, K., & Deng, J. (2016, October). Stacked hourglass networks for human pose estimation. In European Conference on Computer Vision (pp. 483-499). Springer, Cham

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[5] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), 1929-1958.

[6] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[7] Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2014). Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence, 36(7), 1325-1339.

[8] Pavlakos, G., Zhou, X., Derpanis, K. G., & Daniilidis, K. (2017, July). Coarse-to-fine volumetric prediction for single-image 3D human pose. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on (pp. 1263-1272). IEEE.

[9] Tekin, B., Marquez Neila, P., Salzmann, M., & Fua, P. (2017). Learning to fuse 2d and 3d image cues for monocular body pose estimation. In International Conference on Computer Vision (ICCV)

[10] Zhou, X., Huang, Q., Sun, X., Xue, X., & Wei, Y. (2017, October). Towards 3d human pose estimation in the wild: a weakly-supervised approach. In IEEE International Conference on Computer Vision.

[11] Martinez, J., Hossain, R., Romero, J., & Little, J. J. (2017, May). A simple yet effective baseline for 3d human pose estimation. In International Conference on Computer Vision (Vol. 1, No. 2, p. 5).

[12] Fang, H., Xu, Y., Wang, W., Liu, X., & Zhu, S. C. (2017). Learning knowledge-guided pose grammar machine for 3d human pose estimation. arXiv preprint arXiv:1710.06513.

[13] Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., & Wang, X. (2018, March). 3d human pose estimation in the wild by adversarial learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Vol. 1).