인페인팅 및 객체 합성을 활용한 사람 포즈 데이터 증강 기법

추연승, 김현식, 박용석
한국전자기술연구원

{piksal, hskim, yspark}@keti.re.kr

Augmentation Method for Human Pose Data Using Inpainting and Object Synthesis

Yeon-Seung Choo, Hyun-Sik Kim, and Yong-Suk Park
Korea Electronics Technology Institute (KETI)

요 약

실세계에서 사람 포즈 추정 과정에는 다양한 환경적 요인이 발생하여 사람 포즈 추정 성능에 큰 영향을 끼친다. 그 중에서도, 가장 큰 환경적 요인 중 하나는 가려짐, 잘림 현상이다. 포즈 추정 과정에서 사람 앞에 특정 객체가 자리하거나 영상의 너머 존재하지 못한 상황에서 포즈 정보를 추론해야 하기 때문이다. 본 논문에서는 이러한 문제를 해결하기 위해 인위적인 가려짐을 생성하고, 특정 부위를 제거하는 과정을 제안한다. 제안하는 방법을 활용한다면 기존의 데이터셋을 활용해서 포즈 추정 성능을 향상시킬 수 있다.

Ⅰ. 서 론

딥러닝의 발전에 따라, 사람 포즈 추정 기술은 가장 연구가 많이 이루어진 분야 중 하나이다. 이러한 사람 포즈 추정 기술은 전통적으로 top-down 및 bottom-up 으로 크게 2 가지 방법으로 나뉜다. 먼저 top-down 방법은 영상에서 사람을 먼저 검출하고, 검출된 사람들의 관절(키포인트)을 추정하는 방법으로 대표적으로 HRNet, SimpleBaseline 등이 있다 [1-2]. 다음으로 bottom-up 방법은 개별 관절들을 조합하여 사람의 전체 키포인트를 추정하는 방법을 의미한다 [3].

일반적으로 top-down 방법의 경우 사람 영역을 먼저 인지하고 개별 키포인트를 추정하기 때문에 정확도가 높은 대신 낮은 연산 속도를 나타낸다. 반대로, bottom-up 방법의 경우 개별 키포인트를 추정하여 한 명의 사람으로 나타내는 방법이기 때문에 속도는 빠르지만 상대적으로 낮은 정확도를 나타낸다. 이러한 이유로 포즈 추정 기술은 각각 과제에 맞게 알맞은 방법을 활용하는 것이 중요하다.

이러한 이유로 영상에서 사람이 특정 대상에 의해 가려지거나 사람이 잘려서 나타나는 현상은 가장 큰 장애물 중 하나이다. 왜냐하면 팔목, 손목과 같은 자식 노드의 키포인트와 부모 노드의 키포인트 간에 연결이 단절되기 때문이다. 또한 사람 영역을 추론하는 과정에서 장애물, 잘림 현상에 의해 사람 영역이 축소되어 키포인트 추정 성능이 저하되는 현상이 발생할 수 있다. 따라서, 학습 과정에서 인위적으로 잘림 및 가려짐이 발생한 학습 데이터를 생성하고 활용하여 잘림 및 가려짐에 강건한 포즈 추정 기술을 개발해야 할 필요성이 나타났다.

앞선 문제점들을 해결하기 위해, 제안하는 방법에서는 영상에서 미리 지정한 영역을 자연스럽게 제거하는 인페인팅(inpainting) 기법을 활용하여 잘림 현상과 객체를 합성하는 방법을 활용해 인위적으로 가려짐 및 잘림 현상을 구현한 학습 데이터 제작 기법을 제안한다.

Ⅱ. 본론

제안하는 방법은 앞서 언급한 바와 같이, 인페인팅과 객체 합성을 활용하여 특정 관절을 인위적으로 보이지 않도록 변형하는 과정을 나타낸다.

먼저, 제안하는 방법은 학습이 끝난 이후에 키포인트별 정확도를 추정하여 정확도가 낮은 n개의 대상 관절 $p = \{p_1, p_2, \cdots, p_n\}$를 선택한다. 이 관절들은 데이터 증강 대상이 된다.

다음으로, p가 포함된 원본 이미지 I로부터 비디오 인페인팅을 수행하고자 하는 대상 마스크 M을 지정한다. 여기서 대상 마스크 M은 사용자가 선택한 영역 부위, 혹은 미리 지정한 패치와 같이 자유롭게 선택할 수 있다.

그 다음으로 선택된 대상 마스크와 원본 이미지에 대하여 이미지 인페인팅 T를 적용한 최종 인페인팅 영상 I_T를 획득한다.

$$I_T = T(I, M) \tag{1}$$

마지막으로 원본 이미지 I에서 랜덤하게 추출된 객체 o에 대하여, 객체 합성 함수 O를 활용한 합성 영상 I_O를 다음과 같이 획득할 수 있다.

$$I_O = O(I, o) \tag{2}$$

이때, 합성 기준으로 대상 관절 p 가 미리 포함될 비율을 지정한다.

최종적으로, 인페인팅 이미지 I_T, 그리고 합성 이미지 I_O 에 대하여 각각 포즈 추정을 통해 획득한 대상 관절 p_T, p_O와 GT(Ground Truth)의 대상 관절 p_{GT}를 아래와 같이 각 MSE (Mean Squared Error)의 합을 통해 구한다. 아래 수식(3)은 최종 목적 함수를 의미하며 이때 E 는 MSE 를 나타낸다.

$$Loss = E(p_T, p_{GT}) + E(p_O, p_{GT}) + E(p_O, p_T) \quad (3)$$

위와 같은 방법을 통해, 포즈 추정에서 가려짐 및 사라짐에 강건한 포즈 추정을 수행할 수 있다. 그림 1 은 제안하는 방법을 위해 인페인팅 이미지 및 합성 이미지를 생성해내는 예시를 나타낸다.

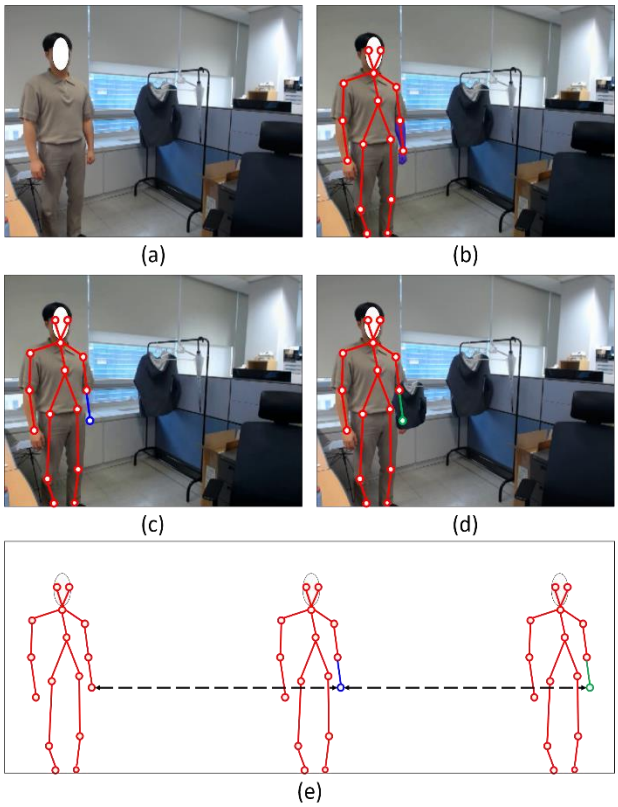

그림 1. (a): 원본 이미지. (b): 마스크 이미지 및 GT 포즈. (c) 인페인팅 이미지 및 포즈 p_T. (d): 합성 이미지 및 포즈 p_O. (e): p_{GT}, p_T, p_O 비교 예시

Ⅲ. 결론

본 논문에서는 인위적으로 가려짐 및 사라짐이 나타난 사람 포즈 추정 데이터셋 제작 방안을 제안한다.

사람 포즈 추정 과정은 크게 top-down 과 bottom-up 방법이 있고, 이는 각각 사람을 먼저 추정하는지, 관절 정보를 먼저 추정하는지에 대한 순서 차이를 나타낸다.

그러나, 두 방법 모두 발목 다음에 발, 손목 다음에 손이 나오듯 사람의 구조적인 신체 정보를 활용하기 때문에 가려짐 및 잘림 등 외부 변수에 따라 성능이 저하될 수 있다.

이러한 문제점을 해결하기 위해, 본 논문에서는 인페인팅을 활용하여 자연스러운 사라짐과 영상 내부의 객체를 활용하여 가려짐 환경을 인위적으로 만들어내어 상대적으로 기존 데이터에 비해 포즈 추정 난이도가 높은 데이터를 생성해내는 방안을 제안한다. 그 결과, 포즈 추정 모델은 정확도가 낮은 관절 정보에 대하여 추가적인 학습을 통해 가려짐과 잘림 등 외부 요소에 강건한 포즈 추정을 할 수 있다.

ACKNOWLEDGMENT

이 논문은 2024 년도 정부(산업통상자원부)의 재원으로 한국산업기술평가관리원의 지원을 받아 수행된 연구임 (No. 20009940, 개인 상황 인지를 기반으로 스마트 홈 케어가 가능한 Home Ambient Intelligence Display (HAID) 및 디자인 개발).

참 고 문 헌

[1] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep High-Resolution Representation Learning for Human Pose Estimation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[2] B. Xiao, H. Wu, and Y. Wei, “Simple Baselines for Human Pose Estimation and Tracking,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[3] Z. Cao, G. Hidalgo, T. Simon, S. -E. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 43, no. 1, pp. 172-186, 1 Jan. 2021.