KIPS Trans. Softw. and Data Eng.
Vol.11, No.12 pp.517~524
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2022.11.12.517

Semantic Occlusion Augmentation for Effective Human Pose Estimation

Hyun-Jae Bae† · Jin-Pyung Kim†† · Jee-Hyong Lee†††

ABSTRACT

Human pose estimation is a method of estimating a posture by extracting a human joint key point. When occlusion occurs, the joint key point extraction performance is lowered because the human joint is covered. The occlusion phenomenon is largely divided into three types of actions: self-contained, covered by other objects, and covered by background. In this paper, we propose an effective posture estimation method using a masking phenomenon enhancement technique. Although the posture estimation method has been continuously studied, research on the occlusion phenomenon of the posture estimation method is relatively insufficient. To solve this problem, the author proposes a data augmentation technique that intentionally masks human joints. The experimental results in this paper show that the intentional use of the blocking phenomenon enhancement technique is strong against the blocking phenomenon and the performance is increased.

Keywords : Data Augmentation, Occlusion, Human Pose Estimation, Deep Learning

가려진 사람의 자세추정을 위한 의미론적 폐색현상 증강기법

배 현 재† · 김 진 평†† · 이 지 형†††

요 약

사람의 자세추정(Human pose estimation)은 사람의 관절 키포인트를 추출하여 자세를 추정하는 방법이다. 폐색현상(Occlusion)이 발생하면, 사람의 관절이 가려지므로 관절 키포인트 추출 성능이 낮아진다. 폐색현상은 총 3가지로 행동할 때 스스로 가려짐, 다른 사물에 의해 가려짐과 배경에 의해 가려짐으로 크게 나뉜다. 본 논문에서는 폐색현상 증강기법을 활용하여 효과적인 자세추정방법을 제안한다. 자세추정방법이 지속적으로 연구되어왔지만, 자세추정방법의 가려짐 현상에 관한 연구는 상대적으로 부족한 상태이다. 이를 해결하기 위해 저자는 사람의 관절을 타겟팅하여 의도적으로 가리는 데이터 증강기법을 제안한다. 본 논문에서의 실험 결과는 의도적으로 폐색현상 증강기법을 활용하면 폐색현상에 강인하며 성능이 올라간 것을 보여준다.

키워드 : 데이터증강, 폐색현상, 자세추정, 딥러닝

1. 서 론

사람의 자세추정(Human Pose Estimation)은 눈, 골반, 무릎과 같은 사람의 해부학적인 관절이나 중요 신체부위를 키포인트(Keypoint)로 지정하여 사람의 자세를 추정하는 기술이다. 사람의 자세추정 방법은 행동인식[1], 자세추적[2], 영상분할[3], 자율주행[4], 시니어 모니터링[5] 그리고 사회적 행동 분석[6] 등 여러 분야에 활용된다. 다양한 분야에서 활용되기 위해서는 사람의 관절 키포인트가 정확하고 폐색현상(Occlusion)에 강인해야 한다. 폐색현상으로 인한 문제로 관절 키포인트 추출 성능이 낮아지며, Fig. 1과 같이 스스로 가려짐[7], 다른 사물에 의해 가려짐과[8] 배경에 의해 가려짐으로 크게 나뉜다. 기존 연구방법에서는 일반적으로 리소스나 비용이 큰 영상분할(Semgmentation)[9]과 깊이지도(Depth Map)[10]를 활용하는 방법으로 진행돼왔다. 본 논문에서는 강인한 관절 키포인트 추출을 위해 사람의 관절 키포인트 뿐만 아니라 신체 부위를 타겟팅하여 고의로 폐색현상을 유발하는 방식의 데이터 증강기법을 제안한다.

폐색현상을 해결하기 위한 방법 중 하나로 비용이 들지 않는 데이터 증강기법을 통한 방법이 있다. 기존의 데이터 증강기법[11-13]으로는 접기(Flipping), 돌리기(Rotation)와 크기변환(Scaling) 등이 있다. 이와 같은 방법들은 이미지 분류(Image Classification), 객체탐지(Object Detection), 영상분할(Segmentation) 등에 활용된다. 자세추정방법에서의 데이터 증강방법[14, 15]은 상체와 하체의 키포인트를 자르기(Cropping)하는 경우가 있다. 본 논문에서는 폐색현상을

† 준 회 원 : 성균관대학교 소프트웨어학과 석사과정
†† 정 회 원 : (재)차세대융합기술연구원 선임연구원
††† 종신회원 : 성균관대학교 소프트웨어학과 교수
Manuscript Received : August 12, 2022
First Revision : September 23, 2022
Accepted : October 4, 2022
* Corresponding Author : Jee-Hyong Lee(john@skku.edu)
* Corresponding Author : Jin-Pyung Kim(jpkim@snu.ac.kr)

Fig. 1. The Different Types of Occlusion States: a) Self Occlusion b) Inter Object Occlusion c) Background Occlusion

해결하기 위해 관절과 신체의 부위를 타겟팅하여 강인한 데이터 증강방법을 제안한다. 본 논문의 기여(Contribution)는 총 3가지로 볼 수 있다. 첫 번째, 사람의 자세추정방법에서 데이터 증강기법은 폐색현상문제에 효과가 있음을 입증하였으며 두 번째로, 폐색현상 문제를 해결하기 위한 데이터 기반 증강방법을 제안하였다. 마지막으로, 데이터증강방법은 폐색현상에 대한 강인함을 입증하였다.

2. 관련 연구

2.1 사람의 자세추정(Human pose estimation)

딥러닝에서의 사람자세추정방법은 크게 2가지로 분류된다. 하향식(Top-Down) 방식[15]과 상향식(Bottom-Up) 방식[16]으로 나누어진다. Top-Down 방식[15]은 이미지에서 사람을 먼저 탐지하고, 탐지된 바운딩박스 내에서 사람의 관절 키포인트를 추출하는 방법이다. Bottom-Up 방식[16]은 이미지에서 사람의 관절 키포인트를 먼저 추정하고, 키포인트간의 상관관계를 분석하여 자세를 추정하는 방식이다. 자세추정방법의 벤치마크로 사용되는 COCO[17]와 MPII[18] 데이터셋이 있다. 최근 Top-Down의 자세추정방법이 대부분의 SOTA(The state of the art)를 달성했다. Top-Down 방식은 크게 2가지로 분류된다. 첫 번째는 각각의 관절 키포인트를 예측하는 것이고 두 번째는 관절 키포인트가 존재할 만한 위치를 확률적 히트맵(Heatmap)으로 계산하고 키포인트의 위치를 추정하는 방법이다.

SimpleBaseline[19]은 출력 피쳐(Feature)의 해상도를 확대하기 위해 디컨볼루션(Deconvolution) 층들을 거의 추가하지 않으므로써 간단하지만, 모델의 무게가 가볍고 비용적으로 효율적인 방법론을 제안하였다. HRNet[15]은 고해상도 피쳐맵(Feature Map)을 유지하면서 Receptive field가 확대되어 공간(Spatial) 정보까지 풍부하게 학습할 수 있다.

Fig. 2. Framework of Top-down Pipeline

2.2 폐색현상 발생 시 자세추정(Occlusion in human pose estimation)

자세추정방법 중 폐색현상 문제는 항상 발생하였지만, 현재까지도 문제를 해결하지는 못하였다. 최근, 자세추정방법에서의 폐색현상 문제를 해결하기 위해 많은 연구들이 시도되고 있다. 영상분할[9]기법과 이미지의 깊이(depth) 정보[10]를 활용하는 연구 뿐만 아니라, 이미지의 히트맵을 활용하여 3D 재건축(Reconstruction)[20] 기법이 있다. 이러한 방법들과는 달리 본 논문에서 제안하는 방법은 사람의 관절 혹은 팔, 다리 부분을 의도적으로 타겟팅한 폐색현상의 증강기법이다.

2.3 데이터 증강(Data augmentation)

데이터 증강방법은 분류, 객체탐지 등 모델의 강인함을 증가시키기 위해 가장 간단하고 강력한 방법론 중 하나이다. 이미지 분류와 객체탐지 분야에서는 이미지 일부분을 제거해주는 흐리기(Blurring), 자르기(Cutout) 그리고 돌리기(Rotation) 등이 있다. 최근 연구로 믹스업(Mixup)[21]은 이미지 분류 분야에서 주로 사용되며 두 개의 이미지를 임의로 선택한 람다(lambda)를 활용하여 합치는 방식이다. 자세추정방법을 위한 증강방법은 보통 스케일링(Scaling), 돌리기(Rotation) 그리고 접기(Flipping) 등을 사용한다.

최근 하향식(Top-Down) 방식[22]에서는 상체 혹은 하체의 관절 키포인트를 나누어 증강하는 방식으로 연구됐다. 본 논문에서는 폐색현상에 강인하고, Pascal VOC 2012 데이터셋을 활용한 사람의 관절, 팔과 다리 부분을 타겟팅한 증강방법을 제안한다.

3. 의미론적 폐색현상 증강방법

제안하는 방법은 Pascal VOC 2012 데이터셋을 활용하였고, 영상분할(Segmentation)된 객체의 마스크를 폐색현상으로 사용하였다. Fig. 3에서의 모델은 HRNet[15]을 활용하여 관절 키포인트를 추출하였고, 관절 키포인트를 타겟팅하여 마스킹된 객체로 폐색현상을 의도적으로 발생시켰다. 기존의 방식은 영상분할(Segmentation)[9], 깊이지도(Depth map)[10] 카메라 등을 활용하여 연구를 진행하였다. 기존의 방식들은 모델의 성능을 높여 보다 정확한 관절 키포인트를 통해 폐색현상이 발생하더라도 추출과 자세를 추정하는 방법이지만, 이러한 방법들은 딥러닝 모델의 리소스(Resource)가 크고 비용(Cost)이 많이 드는 단점이 있다.

폐색현상을 고려한 데이터 증강방법은 예전부터 최근까지 연구가 계속 진행되어 있지만, 자세추정방법 분야에서는 연구가 다양하게 진행되고 있지 않다. Fig. 4에서 자르기(Cutout)와 같은 데이터 증강방법은 사람의 관절키포인트를 모두 가려버리기 때문에 폐색현상이 발생하면, 사람 관절의 컨피던

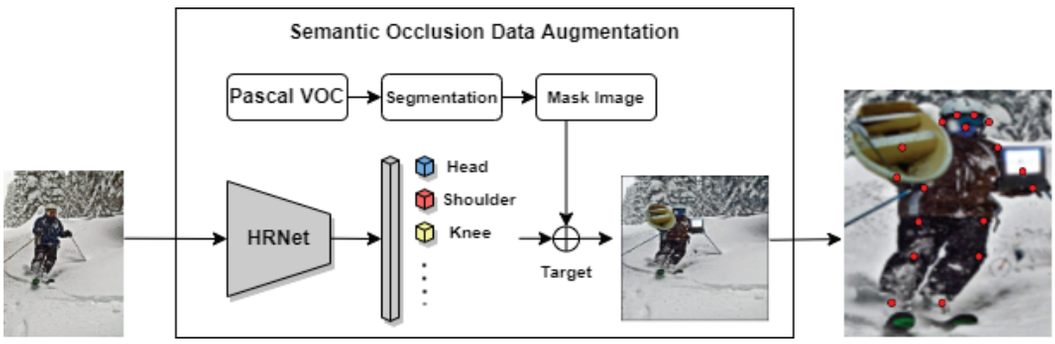


Fig. 3. Overview of Our Approach. The Input Image is Fed to the HRNet to Obtain Keypoint Groups of Augmentation Parameters Which are used to Target the Randomly Selected Parts

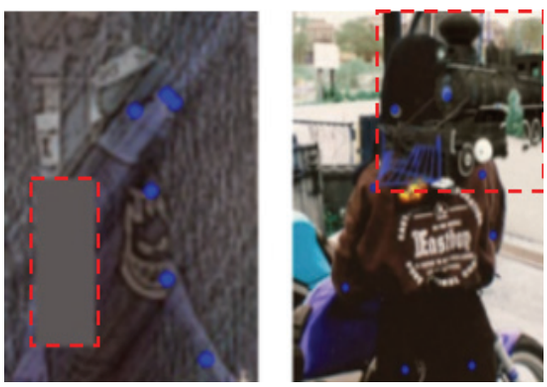


Fig. 4. The Different Types of Target Occlusion: Left) Cutout Occlusion, Right) Semantic Occlusion

스맵(Confidence Map) 성능이 낮아지게 되는 문제점이 있다[23]. 기존의 문제점은 사람의 관절 키포인트를 모두 가리는 경우지만, 본 논문에서 제안하는 방법은 기존 방법과 차별화되었으며, 의도적으로 사람의 관절키포인트를 가려 의미론적인 폐색현상을 만들어낸다. 관절 키포인트를 모두 가려버리면 관절과 관절간의 관계성이 불명확해지기 때문에 관절 추출 성능인 컨피던스맵 성능이 낮아지게된다. 뿐만 아니라, 제안하는 방법에 추가로 흐리기와 자르기 등 여러 폐색현상을 적용하였을 때, 가장 강인하고 의미론적인 폐색현상이 무엇인지 실험으로 확인하였다.

사람의 크기(Scale)은 매우 다양하지만, 기존의 증강방법은 이미지 크기에 강인하지(Robust) 않다. 그뿐만 아니라, 데이터 증강 시 기존 정보를 포함하지 못할 때는 의미론적인 특징을 잃어 원본 데이터의 클래스와는 달라진다[24]. 마지막으로 데이터 크기와 종류에 따른 최적의 폐색현상 마스크를 찾기 어렵다. 그러므로, 기존 방식의 단점들을 보다 효율적으로 해결하기 위해 의미론적 폐색현상 데이터 증강방법을 제안한다. 본 논문은 2020년 ICPR에서 발표한 논문[24]에서 '데이터 증강방법은 폐색현상에 도움을 주지 않는다'라는 주장을 반박하기 위해 제안하는 방법을 [24]에 적용하였다. 본 논문은 자세추정방법에서 의미론적 폐색현상 증강방법을 적용하여 폐색현상이 발생한 상황에서도 키포인트 추출의 성능을 올리는 것이 목표다.

Fig. 5. The Different Types of Data Augmentation : a) Blurring, b) Cutout, c) Cutmix

실험에서는 다른 증강방법들을 비교하여 폐색현상이 일어났을 때, 어떤 증강기법이 폐색현상에 도움이 되는지 확인하고자 한다. 또한, 타겟팅을 관절 키포인트로 했을 때와 여러 개의 관절 키포인트를 부분적으로 가렸을 때의 성능 비교를 확인하였다.

데이터 증강방법의 비교는 Fig. 5와 같이 흐리기(Blurring), 자르기(Cutout) 그리고 컷믹스(Cutmix)[25] 순으로 진행하였다. 컷믹스는 하나의 이미지 안에 두개의 레이블을 가진 이미지를 적당한 비율로 배치하여 두 개의 레이블이 하나의 이미지에 비율에 맞게 배치되는 방법이다. 데이터 증강방법에 타겟팅을 적용하여, 관절 키포인트를 의도적으로 가렸을 때의 성능 비교 후 폐색현상에 가장 효율적인 데이터 증강방법을 제안한다.

4. 실험 결과 및 평가

4.1 데이터셋

본 논문에서는 MS COCO 데이터셋[17]과 MPII 데이터셋[18]에 대해 폐색현상 증강기법을 적용하였다. COCO 데이터셋은 17개의 관절 키포인트로 라벨링되어 있으며 250,000명의 사람객체와 200,000장의 이미지로 구성되어 있다. 본 논문에서의 모델은 COCO train2017 데이터셋으로 학습되었으며, 학습 데이터셋은 150,000명의 사람객체와 57,000장의 이미지로 구성되어 있다. 검증은 COCO val2017 데이터셋으로 5,000장으로 구성되어 있다. COCO의 평가 방법은 Equation (1)과 같이 OKS(Obejct Keypoint Simi-

larity)로 관절 키포인트가 얼마나 유사하게 추출되는지 확인하는 방법이다. d_i는 GT(Ground-Truth)와 발견된 키포인트와의 유클리디안(Euclidean) 거리 값이며, v_i는 GT의 플래그(Flag) 값이다.

$$OKS = \frac{\sum_i e^{-\frac{d_i^2}{2s^2k_i^2}} \delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad (1)$$

MPII 데이터셋은 16개의 관절 키포인트로 라벨링 되어 있으며 40,000명의 사람객체와 25,000장의 이미지로 구성되어 있다. MPII 데이터셋의 이미지는 유튜브(Youtube) 비디오에서 추출되었고 이미지에서는 관절에 대한 좌표뿐만 아니라 3차원 토르소(Torso) 그리고 머리 방향 등 410개의 다양한 활동 레이블링이 존재한다. MPII의 평가 방법은 PCK (Percentage of Correct Key-points)로 몸통(Torso) 다이어미터(diameter)를 기준으로 진행하며, 예측값과 정답(Ground-Truth)의 거리가 특정 임계값(Threshold) 안에 들어오는지 확인하는 방법이다.

4.2 실험 환경

하드웨어 실험 환경은 3.60GHz의 Intel(R) core I7-9700K CPU와 NVIDIA Tesla V100 32GB * 3 GPU (40TFLOPS), 32GB RAM, Linux Ubuntu 16.04 운영체제가 설치된 데스크톱 PC에서 실험하였다. 딥러닝 프레임워크 중 많이 사용되는 PyTorch 1.8.x 버전을 사용하여 실험을 진행하였다.

4.3 모델의 실험 결과 및 평가

본 논문에서는 COCO와 MPII 데이터셋으로 파라미터는 Table 1과 같이 학습을 진행하였으며, 저자가 제안하는 의미론적 폐색현상 데이터 증강방법을 COCO와 MPII 데이터셋에 적용하였다. 저자는 Hourglass, ResNet, HrNet 네트워크에 데이터 증강방법을 적용하여 실험을 비교하였다. 기존의 데이터 증강방법은 대부분 랜덤플립(Random Flip), 회전(Rotation) 그리고 크기변환(Scale)만 적용하였다. 네트워크 학습은 오픈 플랫폼인 '파이토치(PyTorch)'를 사용하였다. Table 2에서의 SO(Semantic Occlusion)은 사람의 관절을 타겟팅하여 의도적으로 가리는 데이터 증강방법이다. AP (Average Precision)는 추정자세와 정답자세의 유사성을 나타내는 척도로 OKS에 따라 계산되는 지표이다.

저자는 에폭(Epoch) 170부터 200까지는 학습률(Learning Rate) 감소 비율을 10으로 설정하였고, 마지막 에폭은 210으로 설정하였다. COCO 학습데이터셋은 각각의 human 박스 Ground-Truth가 있으며, 기본 해상도는 256 × 192 크기이다. 저자는 랜덤플립(Random Flip), 랜덤회전(−40°, 40°) 그리고 랜덤스케일(0.7, 1.3)을 적용하였다. COCO 검증 데이터셋은 Li 등[26]이 제안한 바운딩박스(Bounding Box) 예측 방법을 활용하였다. 저자는 반전된 이미지에 대해서도 자세를 예측하였고, Fig. 6과 같이 마지막 예측값을 얻기 위해 히트맵(Heatmap)을 평균하여 계산하였다. MPII 데이터셋은 학습과 검증 모두 신체의 크기와 중앙값이 제공된다. 주어진 값을 사용하여 타겟된 사람객체 주변의 이미지를 잘라낸다. 그리고 256×256 또는 384×384 크기로 변환해준다. 증강방법은 랜덤플립, 랜덤회전(−30°, 30°) 그리고 랜덤스케일(0.75, 1.25)을 적용하였다.

Table 1. Parameter Values of the Proposed Model

Parameters	Values(functions)
Number of Keypoints	COCO : 17
	MPII : 16
Loss Function	Cross Entropy
Optimizer	Adam
Epochs	HRNet : 210
	ResNet : 210
	SimpleBase : 140
Shuffle	True
Pin Memory	True
Augmentations	Blurring
	Cutout
	Cumix
Augmentations Probability	0.5
Targeting	Keypoint
	Part

Table 2. Human Pose Estimation Results on the MS COCO 2017 Validation Dataset

Augmentation	Arch	Input Size	AP
Baseline	hourglass	256×192	66.9
+SO	hourglass	256×192	70.3
Baseline	resnet_50	256×192	70.4
+SO	resnet_50	256×192	69.2
Baseline	pose_hrnet_w32	256×192	74.1
+SO	pose_hrnet_w32	256×192	74.3

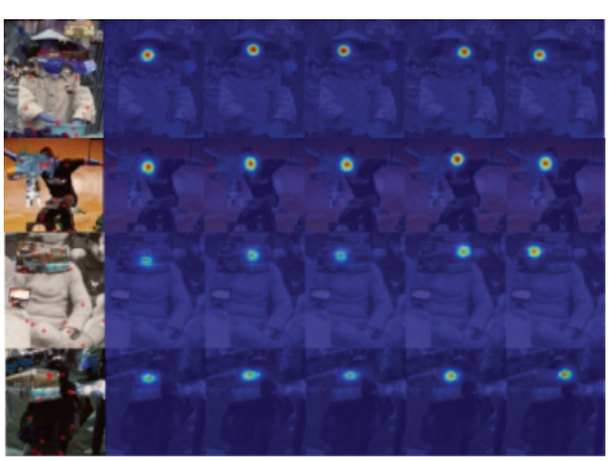


Fig. 6. The Result of Keypoint Heatmap with Occlusion

본 논문에서는 검증 프로토콜을 따라 2개의 벤치마크(Benchmark) 데이터셋에 제안하는 방법들로 성능을 확인하였다. 백본(Backone) 네트워크로 HRNet을 사용하였다. “W32”와 “W48”은 각각 HRNet의 마지막 세 단계에서 고해상도 서브 네트워크의 채널 크기를 표현한 것이다. 채널 크기가 클수록 제안하는 SO(Semantic Occlusion) 방법이 효과적이다. Table 2는 기존의 방법론들을 사용한 베이스라인과 본 논문에서 제안하는 SO(Semantic Occlusion) 방법의 성능을 백본마다 비교하였다. 백본 네트워크 중에서 HRNet의 성능이 가장 좋음을 확인하였다. Table 3은 여러 제안하는 증강방법들을 비교하고, 관절 키포인트를 타겟팅 하거나 신체 부분을 가리는 여러 관절키포인트를 타겟팅하여 비교 실험을 진행하였다. AP는 정답자세와의 유사성을 나타내는 척도로 AP(M)은 객체의 크기에 따른 평가 방법으로 32의 제곱보다 크며 96의 제곱 픽셀보다 작을 때의 성능이다. AP(L)은 위와 같은 방법이며, 96의 제곱 픽셀보다 큰 경우일때의 성능이다. AR은 객체의 여러 IoU(Interjection of Union) 값들의 평균값을 의미한다. 실험은 COCO val2017 데이터셋을 활용하였고, 증강방법 중 Cutout과 Blur기법을 관절 키포인트에 타겟팅한 성능이 가장 좋았다.

베이스라인 대비 w32에서는 성능이 0.4%과 w48에서는 성능이 1.1%가 증가하였다. 이미지 입력 사이즈는 각각 256×192와 384×288로 실험을 진행하였다.

Table 4는 본 논문에서 제안하는 여러 증강방법들을 Table 2와 동일하게 비교하고 데이터셋만 MPII 데이터셋으로 변경하여 실험을 진행하였다. MPII 데이터셋의 실험 결과 중 컷믹스기법을 활용한 증강방법의 성능이 가장 좋다. 베이스라인 대비 w32에서는 성능이 0.7%와 w48에서는 성능이 0.9%가 증가하였다. 이미지의 입력 사이즈는 모두 256×256으로 동일하게 실험을 진행하였다.

Table 3. Human Pose Estimation Results on the MS COCO 2017 Validation Dataset - HRNet

Augmentation	Arch	Input Size	AP	AP 0.5	AP 0.75	AP(M)	AP(L)	AR
Baseline	pose_hrnet_w32	256×192	74.3	90.6	81.7	70.7	80.7	78.8
Blur(parts)	pose_hrnet_w32	256×192	74.1	90.3	81.1	70.6	80.2	78.5
Blur(parts) +SO	pose_hrnet_w32	256×192	**74.6**	**90.4**	**81.9**	**71.1**	**81.2**	**80**
Cutout(keypoint)	pose_hrnet_w32	256×192	74.5	90.5	81.7	70.9	80.7	78.8
Cutout(keypoint) +SO	pose_hrnet_w32	256×192	**74.6**	90.4	81.7	**71.1**	**81.2**	**80**
Cutout(parts)	pose_hrnet_w32	256×192	74.5	90.5	81.6	70.9	80.7	78.8
Cutout(parts) +SO	pose_hrnet_w32	256×192	**74.8**	90.5	**82.2**	**71.1**	**81.8**	**80.1**
Cutout(keypoint) + Blur(keypoint)	pose_hrnet_w32	256×192	74.3	90.5	81.1	70.8	80.6	78.6
Cutout(keypoint) + Blur(keypoint) +SO	pose_hrnet_w32	256×192	**74.7**	90.2	**82**	**71**	**81.5**	**80**
Cutout(parts) + Blur(parts)	pose_hrnet_w32	256×192	74.3	90.4	81.2	70.6	80.5	78.6
Cutout(parts) + Blur(parts) +SO	pose_hrnet_w32	256×192	**74.5**	90.4	**81.9**	**71.1**	**81.1**	**79.8**
Cutmix	pose_hrnet_w32	256×192	74.4	90.7	81.5	71.1	80.5	78.8
Cutmix +SO	pose_hrnet_w32	256×192	**74.7**	90.3	**82.1**	**71.1**	**81.4**	**80**
Augmentation	Arch	Input Size	AP	AP 0.5	AP 0.75	AP(M)	AP(L)	AR
Baseline	pose_hrnet_w48	384×288	76.3	90.8	82.9	72.3	83.4	81.2
Blur(parts) +SO	pose_hrnet_w48	384×288	76.4	90.8	83.2	72.5	83.4	81.4
Cutout(keypoint) +SO	pose_hrnet_w48	384×288	76.8	90.9	83.3	72.9	83.9	81.7
Cutout(parts) +SO	pose_hrnet_w48	384×288	76.7	90.7	83.4	72.8	83.6	81.6
Cutout(keypoint) + Blur(keypoint) +SO	pose_hrnet_w48	384×288	**77.2**	**91**	**84**	**73.4**	**84**	**82**
Cutout(parts) + Blur(parts) +SO	pose_hrnet_w48	384×288	76.9	91	83.5	73	84	81.7
Cutmix+SO	pose_hrnet_w48	384×288	76.7	90.9	83.4	72.8	83.9	81.7

Table 4. Human Pose Estimation Results on the MPII Validation Dataset - HRNet

Augmentation	Arch	Input Size	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean	Mean @0.1
Baseline	pose_hrnet_w32	256×256	97.1	95.9	90.3	86.4	89.1	87.1	83.3	90.3	37.7
Blur(keypoint)	pose_hrnet_w32	256×256	97.1	95.9	90.3	85.8	89.2	86.4	82.2	90	37.5
Blur(keypoint) +SO	pose_hrnet_w32	256×256	97.2	95.8	89.9	85.9	88.7	86.4	82.4	89.9	37.3
Blur(keypoint) +SO	pose_hrnet_w48	256×256	97.2	96	90.7	86.6	89.4	86.6	82.8	90.3	38.2
Cutout(part)	pose_hrnet_w32	256×256	97.5	96.2	90.7	86.6	89.2	86.5	83	90.4	38
Cutout(part) +SO	pose_hrnet_w32	256×256	97.2	96.5	91.5	86.9	90.2	87.3	83.6	90.9	38.3
Cutout(part) +SO	pose_hrnet_w48	256×256	**97.3**	**96.6**	**91.4**	**87.1**	**90.7**	**88**	**84.4**	**91.2**	38.9
Cutmix	pose_hrnet_w32	256×256	97.4	96.3	90.9	86.6	89.5	86.4	82.9	90.5	38.2
Cutmix +SO	pose_hrnet_w32	256×256	97.3	96.7	91.6	87	89.7	87.4	84.4	91	38
Cutmix +SO	pose_hrnet_w48	256×256	**97.4**	**96.5**	**91.7**	**87.2**	**90.2**	**88**	**84.6**	**91.2**	39

Table 5. Human Pose Estimation Results on the MS COCO 2017 Test-dev Dataset - HRNet

Augmentation	Arch	Input Size	AP	AP 0.5	AP 0.75	AP(M)	AP(L)	AR
Blur(parts) +SO	pose_hrnet_w32	256×192	73.6	92.1	81.8	70.2	79.4	78.9
Cutout(keypoint) +SO	pose_hrnet_w32	256×192	74	92.4	82.2	70.8	79.6	79.3
Cutout(parts) +SO	pose_hrnet_w32	256×192	73.9	92.3	82.2	70.7	79.6	79.3
Cutout(keypoint) + Blur(keypoint) +SO	pose_hrnet_w32	256×192	73.9	92.4	82.2	70.6	79.6	79.3
Cutout(parts) + Blur(parts) +SO	pose_hrnet_w32	256×192	73.8	92.2	82	70.5	79.5	79.1
Cutmix +SO	pose_hrnet_w32	256×192	73.8	92.3	82.1	70.7	79.4	79.2
Augmentation	Arch	Input Size	AP	AP 0.5	AP 0.75	AP(M)	AP(L)	AR
Blur(parts) +SO	pose_hrnet_w48	384×288	75.5	92.5	83.4	72	71.4	80.6
Cutout(keypoint) +SO	pose_hrnet_w48	384×288	75.9	92.6	83.7	72.3	82	81
Cutout(parts) +SO	pose_hrnet_w48	384×288	75.8	92.6	83.5	72.2	81.7	80.9
Cutout(keypoint) + Blur(keypoint) +SO	pose_hrnet_w48	384×288	**76.3 (COCO SOTA #13)**	**92.7**	**84.2**	**72.9**	**82.1**	**81.3**
Cutout(parts) + Blur(parts) +SO	pose_hrnet_w48	384×288	76	92.7	83.6	72.5	82	81.1
Cutmix +SO	pose_hrnet_w48	384×288	75.9	92.6	83.6	72.3	82	81

Table 5는 폐색현상 증강방법들을 w32와 w48에 적용하여 COCO test-dev 2017 데이터셋의 테스트 성능을 비교한 결과이다. 이 중 Cutout과 Blurring 방법을 관절 키포인트를 타겟으로 진행하였을 때, 성능이 가장 높았다. Fig. 7, 8, 9)는 COCO 테스트 데이터 세트의 제안 방법으로 얻은 일부 자세 추정 결과를 보여준다. 본 논문에서는 저자가 제안하는 의도적으로 관절 키포인트를 가리는 의미론적 폐색현상 방법을 적용하였을 때, COCO와 MPII 데이터셋에서 모두 성능이

Fig. 7. Example of Self Occlusion on the COCO Test Set

Fig. 8. Example of Inter Object Occlusion on the COCO Test Set

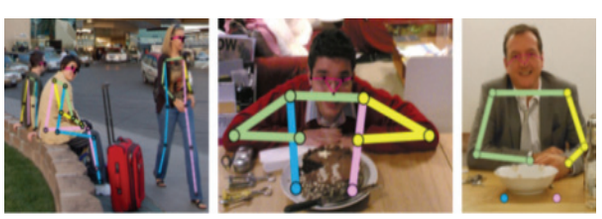


Fig. 9. Example of Background Occlusion on the COCO Test Set

올라간 것을 확인하였다. Cutout, Blurring 그리고 Cutmix을 같이 적용하여도 성능이 올라가고 이 중에서 Cutout과 Blurring을 같이 사용하였을 때, 성능이 가장 많이 오른 것을 확인하였다. 본 논문에서 제안하는 방법은 Microsoft사의 주관대회인 'COCO Keypoint Challenge'에서 SOTA(State-of-the-art) 13등을 달성하는 성능을 나타내었다

5. 결론 및 향후 방향

본 논문에서는 SO(Semantic Occlusion)라는 폐색현상 증강방법을 활용한 자세추정방법 연구를 목적으로 자세추정 모델들의 관절 키포인트 추출에 대한 성능 평가 후, 성능이 좋은 방법론을 확인하였다.

자세추정방법 분야에서의 증강방법을 소개하였고, 해당 분야에서 폐색현상이 발생했을 때 성능을 높이는 방법론을 제안하였다. HRNet를 활용하여 폐색현상을 관절부분에 의도적으로 타겟팅하여 학습하였다. 벤치마크와 다양한 실험들을 통해 저자가 제안하는 방법론의 성능을 입증하였다. 저자는 본 논문에서 제안하는 방법론을 통해 앞으로 자세추정방법 뿐만 아니라 폐색현상 관련연구에 영감을 제공할 수 있기를 바란다. 추가적으로 SCI에 제안할 방법론으로는 폐색현상을 사람의 이상행동인식에도 적용하고 효과를 확인할 예정이다.

References

[1] H. J. Bae, G. J. Jang, Y. H. Kim, and J. P. Kim, "LSTM (long short-term memory)-based abnormal behavior recognition using AlphaPose," *KIPS Transactions on Software and Data Engineering*, Vol.10, No.5, pp.187-194, 2021.

[2] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran, "Detect-and-Track: Efficient pose estimation in videos," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[3] Z. Li, X. Chen, W. Zhou, Y. Zhang, and J. Yu, "Pose2body: Pose-guided human parts segmentation," *In 2019 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, pp.640-645, 2019.

[4] Z. Fang and A. M. Lopez, "Intention recognition of pedestrians and cyclists by 2d pose estimation," *arXiv preprint arXiv:1910.03858*, 2019.

[5] P. A. Dias, D. Malafronte, H. Medeiros, and F. Odone, "Gaze estimation for assisted living environments," In *the IEEE Winter Conference on Applications of Computer Vision*, pp.290-299, 2020.

[6] L. Ladicky, P. H. S. Torr, and A. Zisserman, "Human pose estimation using a joint pixel-wise and part-wise formulation," In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp.3578-3585, 2013.

[7] Y. Huang, B. Sun, H. Kan, J. Zhuang, and Z. Qin, "Followmeup sports: New benchmark for 2d human keypoint recognition," *arXiv preprint arXiv:1911.08344*, 2019.

[8] T. Golda, T. Kalb, A. Schumann, and J. Beyerer, "Human pose estimation for real-world crowded scenarios," *arXiv preprint arXiv:1907.06922*, 2019.

[9] P. S. R. Kishore, S. Das, P. S. Mukherjee, and U. Bhattacharya, "Cluenet : A deep framework for occluded pedestrian pose estimation," 12 2019.

[10] U. Rafi, J. Gall, and B. Leibe, "A semantic occlusion model for human pose estimation from a single depth image," In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp.67-74, 2015.

[11] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020.

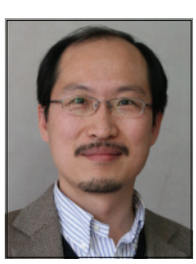
[12] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, Vol.6, No.1, pp.60, 2019.
[13] L. Taylor and G. Nitschke, "Improving deep learning using generic data augmentation," *arXiv preprint arXiv:1708.06020*, 2017.
[14] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multiscale structure-aware network for human pose estimation," CoRR, *arXiv preprint arXiv:1803.09894*, 2018.
[15] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep HighResolution representation learning for human pose estimation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
[16] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," CoRR, *arXiv preprint arXiv:1812.08008*, 2018.
[17] T.-Y. Lin et al., "Microsoft COCO: common objects in context," CoRR, *arXiv preprint arXiv:1405.0312*, 2014.
[18] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2014.
[19] B., Xiao, H., Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
[20] N. D. Reddy, M. Vo, and S. G. Narasimhan. "Occlusion-net: 2d/3d occluded keypoint localization using graph networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[21] H. Guo, Y. Mao, and R. Zhang, "Mixup as locally linear out-of-manifold regularization," CoRR, *arXiv preprint arXiv:1809.02499*, 2018.
[22] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. "Cascaded pyramid network for multi-person pose estimation," In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.
[23] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, "How robust is 3D human pose estimation to occlusion?," *arXiv preprint arXiv: 1808.09316*, 2018.
[24] R. Pytel, O. S. Kayhan, and J. C. van Gemert, "Tilting at windmills: Data augmentation for deep pose estimation does not help with occlusions," *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021.
[25] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.
[26] W. Li et al., "Rethinking on multi-stage networks for human pose estimation," *arXiv preprint arXiv:1901.00148*, 2019.

배 현 재

https://orcid.org/0000-0002-2164-0125
e-mail : jason0425@skku.edu
2019년 ~ 2020년 (재)차세대융합기술연구원 인턴연구원
2020년 ~ 2021년 (재)차세대융합기술연구원 연구원
2021년 ~ 현 재 성균관대학교 소프트웨어학과 석사과정
2022년 ~ 현 재 ㈜클레버러스 대표이사
관심분야 : Pose Estimation, Object Detection, Action Recognition

김 진 평

https://orcid.org/0000-0003-4840-7216
e-mail : jpkim@snu.ac.kr
2006년 성균관대학교 전자전기컴퓨터공학과 (석사)
2014년 성균관대학교 전자전기컴퓨터공학과 (박사)
2016년 ~ 2018년 한국철도기술연구원 선임연구원
2018년 ~ 2019년 한국도로공사 도로교통연구원 책임연구원
2019년 ~ 현 재 (재)차세대융합기술연구원 선임연구원
관심분야 : Artificial Intelligence & Computer Vision



이 지 형

https://orcid.org/0000-0001-7242-7677
e-mail : john@skku.edu
1993년 한국과학기술원 전산학과(학사)
1995년 한국과학기술원 전산학과(석사)
1999년 한국과학기술원 전산학과(박사)
2002년 ~ 현 재 성균관대학교 소프트웨어학과 교수
관심분야 : Machine Learning, Deep Learning, Intelligence System