단일 이미지에 기반을 둔 사람의 포즈 추정에 대한 연구 동향

Recent Trends in Human Pose Estimation Based on a Single Image

조정찬
Jungchan Cho

(13120) 경기도 성남시 수정구 성남대로 1342, 가천대학교 소프트웨어학과
thinkai@gachon.ac.kr

요 약

최근 딥러닝 기술이 발전함에 따라 많은 컴퓨터 비전 연구 분야에서 주목할 만한 성과들이 지속적으로 나오고 있다. 단일 이미지를 기반으로 사람의 2차원 및 3차원 포즈를 추정하는 연구에서도 비약적인 성능향상을 보여주고 있으며, 많은 연구자들이 문제의 범위를 확장하며 활발한 연구 활동을 진행하고 있다. 사람의 포즈 추정은 다양한 응용 분야가 존재하고, 특히 이미지나 비디오 분석에서 사람의 포즈는 행동 및 상태, 의도 파악을 위한 핵심 요소가 되기 때문에 상당히 중요한 연구 분야이다. 이러한 배경에 따라 본 논문은 단일 이미지를 기반으로 한 사람의 포즈 추정 기술에 대한 연구 동향을 살펴보고자 한다. 강인하고 정확한 문제 해결을 위해 다양한 연구 활동 결과가 존재한다는 점에서 본 논문에서는 사람의 포즈 추정 연구를 2차원 및 3차원 포즈 추정에 대해서 나누어 살펴보고자 한다. 끝으로 연구에 필요한 데이터 세트 및 사람의 포즈 추정 기술을 적용하는 다양한 연구 사례를 살펴볼 것이다.

Abstract

With the recent development of deep learning technology, remarkable achievements have been made in many research areas of computer vision. Deep learning has also made dramatic improvement in two-dimensional or three-dimensional human pose estimation based on a single image, and many researchers have been expanding the scope of this problem. The human pose estimation is one of the most important research fields because there are various applications, especially it is a key factor in understanding the behavior, state, and intention of people in image or video analysis. Based on this background, this paper surveys research trends in estimating human poses based on a single image. Because there are various research results for robust and accurate human pose estimation, this paper introduces them in two separated subsections: 2D human pose estimation and 3D human pose estimation. Moreover, this paper summarizes famous data sets used in this field and introduces various studies which utilize human poses to solve their own problem.

논문투고일자 : 2019. 08. 06 심사일자 : 2019. 08. 07 게재확정일자 : 2019. 10. 21

키워드: 딥러닝, 사람 포즈, 사람 포즈 추정, 행동인식, 연구 동향
Keyword: Deep learning, human pose, human pose estimation, action recognition, research trends

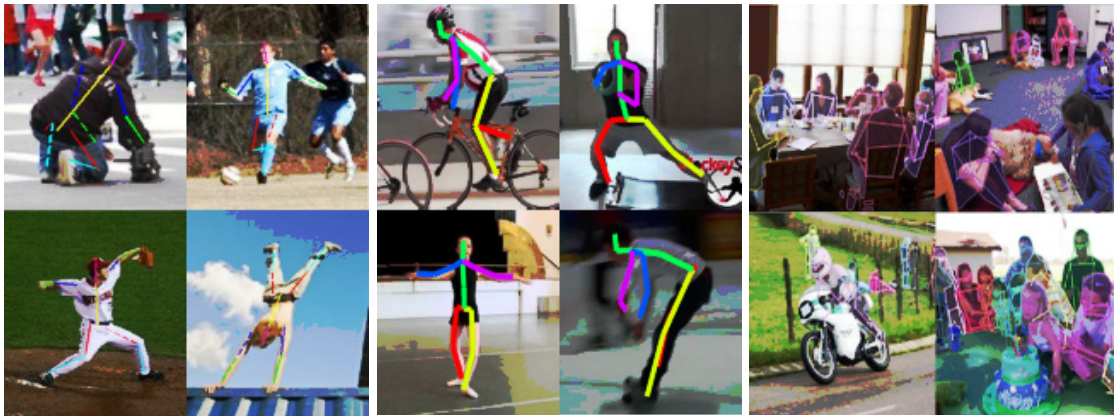
1. 서론

사물인터넷(IoT) 클라우드 등 정보통신기술의 발달로 데이터 규모가 폭발적으로 커지고 있다. 특히 CCTV가 광범위하게 보급되고 소셜네트워크서비스(SNS)가 활성화되면서 이미지와 동영상 데이터가 기하급수적으로 증가하고 있다. 이러한 동영상 데이터의 급격한 증가는 인터넷뿐만 아니라 가상현실(VR)과 증강현실(AR) 같은 대용량 트래픽 소스(source)가 증가된 것에 기인한다. 시스코(CISCO)에서는 IP 비디오트래픽이 2022년까지 전체 IP 트래픽의 82%를 차지할 것으로 예측하였다[1]. 이와 같이 이미지·동영상 데이터의 증가 추세가 사람이 관리할 수 있는 수준을 넘어섬에 따라 이미지·동영상의 자동분석기술에 대한 관심이 커지고 있으며 국내에서도 관련 연구가 활발하게 진행되고 있다 [2-5].

이미지·동영상의 자동분석을 위해서는 사람의 상태와 행동에 대한 이해가 수반되어야 하며, 이는 비정상 행동 탐지, 자율주행 자동차에서의 활용, 로봇 컨트롤 등에서 광범위하게 활용될 수 있다는 점에서 중요하다. 이에 따라 본 논문에서는 단일 이미지 데이터를 이용한 사람의 포즈 인식에 관한 연구 동향을 살펴보고자 한다.

사람의 포즈 추정 문제는 사람의 주요 관절의 위치를 추정하는 문제로 정의할 수 있다[6]. 예를 들어 (그림 1)과 같이 사람의 머리, 목, 어깨, 팔꿈치, 손목, 엉덩이, 무릎, 발목 등의 좌표를 추정하는 것이다. 이러한 사람의 포즈를 추정하는 문제는 초기에 HoG (Histogram of oriented gradients)[10] 묘사자(descriptor)와 그림구조모델(pictorial structure model)을 이용하여 연구되어 왔으나[6] 2012년 이미지넷챌린지(ImageNet Large Scale Visual Recognition Challenge)[11,12] 이후 딥러닝(deep learning) 기술이 사람의 포즈를 추정하는 문제에도 적용되면서 비약적으로 성능이 향상되어 왔다.

본 논문에서는 먼저 2장에서 포즈 추정 연구에 관한 국외의 기술 동향을 2차원 포즈 추정과 3차원 포즈 추정으로 나누어 분석한다. 다음으로 3장에서는 사람의 포즈 추정에 관련한 국내의 연구 동향을



(이미지 출처, 좌측: Leed Sports Pose Dataset[7], 중간: MPII Human Pose Dataset[8], 우측: COCO 2019 Keypoint Detection Task[9])
(그림 1) 사람의 포즈 추정에 대한 데이터 세트의 예

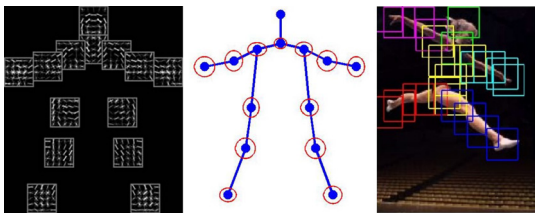
파악하였고, 4장에서 새로운 연구를 위한 데이터 세트 및 포즈 추정 결과를 활용한 여러 연구들을 분석하여 시사점을 도출하고자 하였다.

2. 국외의 연구 동향

단일 이미지를 바탕으로 사람의 포즈를 추정하는 연구는 크게 사람의 2차원 포즈 추정과 3차원 포즈 추정으로 구분할 수 있다. 본 장에서는 관련 연구 사례를 통해 기술 동향을 분석하고자 하였다. 최근 국내에서도 사람의 포즈 추정 기술에 대해 활발히 연구되고 있으므로 1) 외국 연구 사례를 먼저 살펴보고 2) 국내의 연구결과는 3장에서 따로 살펴보도록 한다.

2.1. 2차원 포즈 추정 기술

사람의 포즈를 추정하는데 있어서 가장 직관적인 접근 방법은 객체 인식(object detection) 기술을 직접적으로 적용하는 것이다. 즉, 사람의 각 관절을 독립된 객체로 보고 이미지에서 서로 다른 다수의 객체를 찾아내는 방식이다. 그러나 이 방식의 경우 분별력이(discriminative power) 약한 사람의 관절에는 독립적으로 적용이 어려운 문제가 있다. 예를 들어 팔꿈치와 같은 관절이 독립적인 패치 이미지로 사용되는 경우를 생각해 보자. 이 경우 패치 이미지가 어떤 관절의 부분에 속하는지 인식하는 것은 사람에게도 어려운 문제라는 것을 알 수 있다. 이에 따라 딥러닝 기술을 활용하기 이전의 포즈인식 연구에서는 HoG 묘사자를 이용하여 각 관절의 후보를 찾고 이들 간의 연결 관계를 수학적으로 모델링함으로써 해결하고자 하였다. 이는 관절 간의 관계 모델링을 통해 얼굴과 같은 분별력이 큰 관절의 주변에서 어깨와 같은 분별력이 약한 관절을 찾는 접근 방식이다. 이에 대한 대표적인 방식으로 Yang 외 연구진[6]의 연구를 들 수 있는데 이는 관절과 관절의 관계를 스프링으로 모델링한 그림구조 모델에 기반을 둔 방식이다(그림 2). 그러나 HoG 묘사자는 같은 관절은 유사하게 서로 다른 관

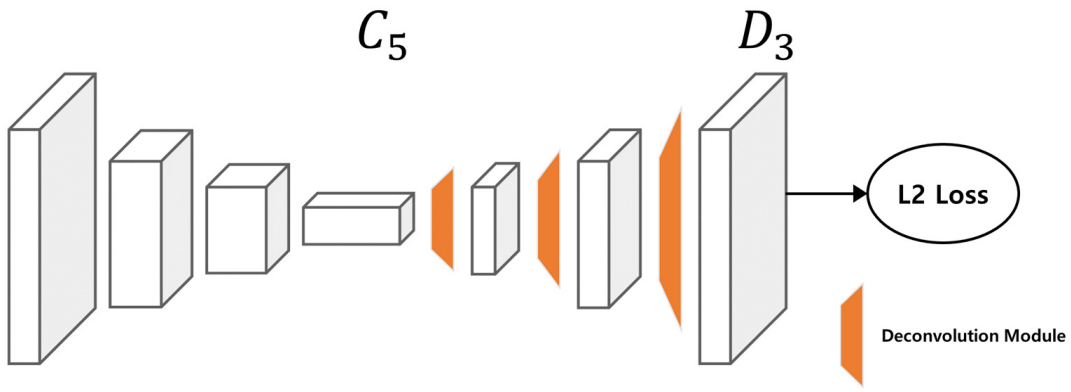


(이미지 출처: Articulated pose estimation with flexible mixtures-of-parts[6])
(그림 2) 초기의 단일 이미지를 통한 사람의 2차원 포즈 추정 연구의 예

절은 다르게 표현할 수 있는 표현력(representation power)이 부족하여 이를 활용한 사람의 포즈 인식 성능은 만족할만한 수준에 도달하지 못하였다.

한편, 2012년 AlextNet[12]이 딥러닝을 사용하여 이미지넷챌린지에서 압도적인 차이로 우승을 차지하고, 이후 컴퓨터 비전의 여러 연구들에서 딥러닝 기술을 활용하기 시작하였다. 포즈를 추정하는 연구에 있어서도 딥러닝 기술이 활용되기 시작하였다. Tosheve 외 연구진[13]은 사람의 포즈를 추정하는 문제를 심층 신경망(Deep Neural Network; DNN)에 기반을 둔 회귀 모델(regressor) 학습으로 정의하고, 이러한 회귀 모델에 케스케이드(cascade) 방식을 적용하여 높은 정밀도에 도달할 수 있는 방법을 제안하였고, 실험결과에서 이전의 방법과 비교하여 상당히 향상된 성능을 보여주었다. 이에 따라 여러 연구들에서 사람의 포즈를 인식하는 데 있어서 딥러닝을 활용하고 있는 실정이다[14,16-18,21].

Wei 외 연구진[14]은 관절과 관절간의 이미지 의존적 공간 모델(image-dependent spatial model)을 학습하는 프레임워크를 제안하였다. 이후 추정 결과를 정교화(refinement)하기 위하여 이전 단계로부터 도출된 신념 지도(belief map)에서 작동하는 순차적인 구조를 제안하였고 LSP[7], FLIC[15], MPII[8] 데이터 세트에서 향상된 결과를 보여주었다. 또 다른 대표적인 연구로 Newell 외 연구진[16]이 제안한 적층 모래시계(stacked hourglass) 신경망을 들 수 있다. Newell 외 연구진의 연구에서는

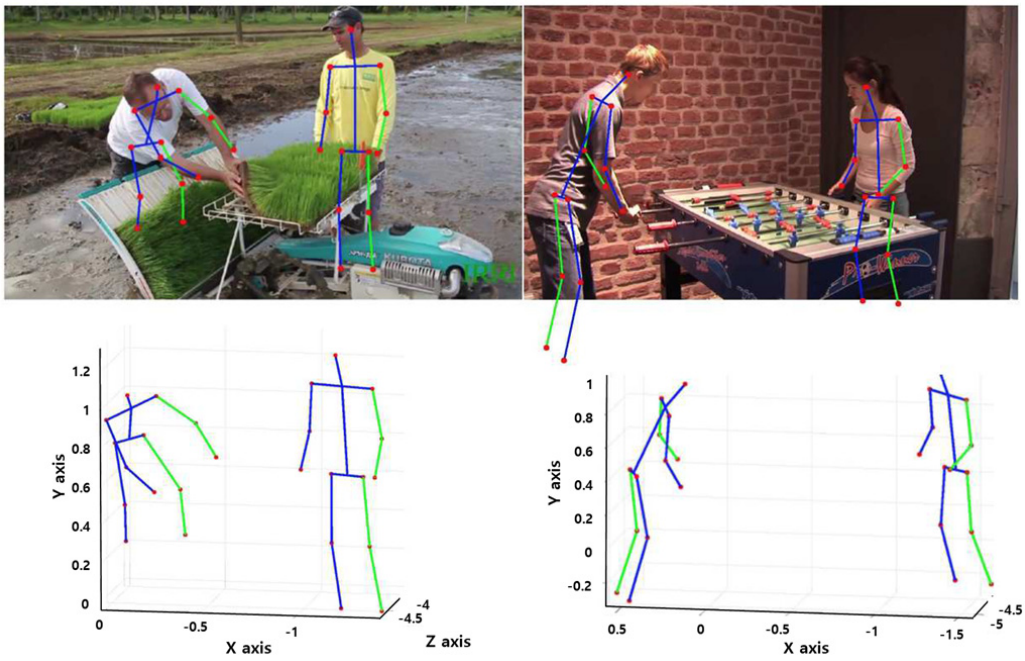


(그림 3) 이미지 출처: 잔류 네트워크에 기반 Xiao 외 연구진이 제안한 간단한 구조의 네트워크 [18].

풀링(pooling)과 업샘플링(upsampling)에 기반을 둔 모래시계 형태의 모듈을 제안하였으며 이러한 모듈을 연속적으로 적층함으로써 FLIC 데이터 세트[15]와 MPII 데이터 세트[8]에서 상당히 높은 성능을 보여 주었다. 반면에 이러한 방식은 사람 또는 사물에 가려진 관절이나 사진 촬영 후 이미지에서는 보이지 않는 관절, 복잡한 배경 등이 있는 경우에는 문제가 있었다. 이를 해결하기 위해 Chen 외 연구진[17]은 캐스케이드 피라미드 네트워크(CPN: Cascaded Pyramid Network) 구조를 제안하였는데 이는 "어려운(hard)" 관절 검출 문제를 보다 쉽게 만드는 것을 목표로 하는 것으로 눈, 손과 같이 "단순한" 관절의 위치를 추정하는 GlobalNet과 GlobalNet의 결과를 이용하여 보다 어려운 관절의 위치를 정교화하는 RefineNet으로 구성되어 있다.

이후 계속적인 연구들에서 포즈 인식 성능이 향상되어 왔으나 네트워크 구조의 복잡도가 증가하는 문제가 있었다. Xiao 외 연구진[18]은 이러한 문제점을 지적하고 잔류 네트워크(Residual Network)[19]에 기반을 둔 간단한 구조만으로도 기존의 방법에 비하여 향상된 성능을 달성하는 것이 가능함을 실험을 통해 증명하였고(그림 3) 포즈연구의 새로운 아이디어 평가 시 기준선으로 활용할 수 있다는 시사점을 도출하였다. 최근 Sun 외의 연구진[20]은 기존의 방법이 공간적으로 줄어들어 생성된 저해상도 특징(feature) 표현에서 포즈 위치에 관한 고해상도 표현을 다시 복구하는 한계가 있음을 지적하였다. 이에 따라 해당 연구진은 네트워크는 전체 과정에서 고해상도 표현을 유지하고 깊이를 더해갈수록 저해상도 네트워크를 병렬로 연결한 네트워크를 제안하였으며, COCO 데이터 세트[9]에서 가장 좋은 성능을 보여주고 있다.

앞서 살펴 본 사람의 포즈를 추정하는 방법은 단일 이미지에서 사람을 찾고 한 사람의 포즈를 정확히 추정하는 하향식(top-down) 방식의 연구이다. 이러한 방식은 이미지에서 사람의 수가 증가하는 경우에 계산량이 사람 수에 비례하여 증가하는 문제점이 있다. 이러한 한계를 극복하기 위한 방법으로 상향식(bottom-up) 방식의 연구가 있다. 이 상향식 방식은 이미지에서 먼저 사람의 관절을 찾고 관절들 간의 연관성을 찾아서 사람의 포즈를 추정하는 방식이다. 이를 통하여 다수의 사람이 있는 이미지에서 계산의 효율성을 추구하면서 가림에 의한 정확도 하락을 방지하고자 한다. 대표적으로 Carnegie Mellon University(CMU)의 Cao 외 연구진[21]이 제안한 OpenPose연구가 있다. 이는 파트 근접 필드(Part Affinity Fields)와 같은 비모수적 표현을 통해 신체 부위 영상에서 관절의 연결관계를 학습하는 방식이며, 이를 적용하여 사람의 인원수와 무관하게 실시간으로 성능 달성이 가능함을 확인하였다.



(그림 4) 이미지 출처: 단일 이미지를 활용한 3차원 포즈 추정 연구의 예[30]

2.2. 3차원 포즈 추정 기술

마커를 이용하여 3차원 정보를 얻는 모션캡쳐(motion capture) 기술은 영화 제작 등에서 빈번하게 활용되고 있다. 그러나 이 기술은 3차원 정보를 수집하는 장비가 고가이며, 실외 사용이 어려운 한계가 있다. 이러한 한계를 극복하기 위하여 단일 이미지를 입력으로 사용하여 사람의 3차원 포즈를 추정하는 연구가 활발하게 진행되고 있다(그림 4).

한 장의 이미지에서 사람의 3차원 포즈를 추정하는 것은 2차원 포즈 추정의 정확도에 의존한다. 뿐만 아니라 3차원 포즈 추정은 내재적 모호함을 가지고 있어 사람의 2차원 포즈 추정보다 어려운 문제가 된다[22]. 이러한 내재적 어려움으로 인해 초기의 3차원 포즈 추정 연구는 1) 사람의 2차원 포즈 추정에서 발생하는 노이즈의 정확한 모델링(noise modeling), 2) 강인한 3차원 복원기술의 적용의 두 가지 방향을 혼합한 형태로 진행되었다. Simo-Serra 외의 연구진[22]은 2차원 포즈 추정에서 발생한 노이즈를 모양(shape) 공간상에 전파하고 사람의 3차원 형상이 나타낼 수 있는 구조학적 정보를 구속조건(constraint)으로 하여 가장 그럴듯한 결과를 선택함으로써 3차원 포즈를 추정하는 방식을 제안하였다. 또한 Simo-Serra 외 연구진[23]은 2차원 포즈 추정과 3차원 포즈 추정이 독립적으로 수행됨에 따라 2차원 정보의 오류 수정이 불가함을 지적하였다. 이를 해결하기 위한 방법으로 해당 연구진은 베이시안(Bayesian) 프레임워크를 적용하여 2차원 추정과 3차원 추론 문제를 결합함으로써 보다 좋은 성능을 도출하는 방법을 제안하였다.

3차원 포즈 추정에서는 2차원 포즈 추정 결과를 이용하는 방법 외에 2차원 특징을 입력으로 하고 3차원 복원을 결과로 하는 회귀 문제로 설계하려는 방식이 존재한다. Agrawal 외 연구진[24]은 실루엣을 입력으로 하여 관련 벡터 머신(Relevant Vector Machine; RVM)[25]을 이용하여 3차원 포즈를 추정하는 방식을 제안하였으며, Kostrikov 외 연구진[26]은 회귀 포레스트(regression forest)를

〈표 1〉 단일 이미지 기반을 둔 사람의 2차원 및 3차원 포즈 추정에 사용되는 데이터 세트

데이터 세트	관절의 수	사람 수	2차원 포즈	3차원 포즈	비고
LSP [8]	14	2,000	✓	-	-
LSP Extended[45]	14	10,000	✓	-	-
FLIC [15]	10	5,003	✓	-	상반신 포즈
MPII [8]	16	40,000	✓	-	410가지의 사람 행동
COCO [9]	17	250,000	✓	-	객체 인식/분할, 컨텍스트(context) 인식, 캡션 등
Human EVA [46]	15	40,000	✓	✓	6가지 행동
Human 3.6M [47]	24	3,600,000	✓	✓	17가지 행동
POSE TRACK [48]	15	〉276,000	✓	-	-
MPI-INF-3DHP [29]	28	〉1,300,000	✓	✓	다양한 포즈를 포함한 8가지 활동(activity)
MuCo-3DHP [32]	28	-	✓	✓	MPI-INF-3DHP 이용한 학습용 합성 데이터 (복수의 사람)
MuPoTs-3D [32]	28	〉8,000	✓	✓	복수의 사람

이용하는 방식을 제안하였다.

지금까지 살펴본 방식은 딥러닝에 기반을 두지 않고 2차원 이미지를 이용하여 3차원 포즈를 추정하였다. 그러나 최근에 3차원 포즈 추정에 있어서도 딥러닝을 활용한 다수의 연구들이 있다[27-30, 32-33]. 특히 회귀 문제에서 딥러닝에 기반을 둔 방법들이 우수한 성능을 보임으로써 여러 3차원 포즈 추정 연구들에서는 2차원 이미지 또는 2차원 포즈 추정 결과로부터 3차원 포즈를 추정하는 회귀 문제로 설계하고 있다. Pavakos 외 연구진[27]은 2차원 포즈 추정에서 제안된 모래시계 구조를 응용하여 기존의 방식보다 좋은 성능을 보여주었다. Yang 외 연구진[28]은 실험 환경이 아닌 일반적인 환경에서의 3차원 포즈 추정 성능 향상을 위하여 하드 코딩(hard-coded)된 포즈 추정의 구속조건을 정의하는 대신 정답과 추정된 3차원 포즈를 구별하기 위한 다중 소스 구별자(multi-source discriminator)를 설계하고 적대적학습(adversarial learning) 프레임워크를 활용하는 방법을 제안하였다. Mehta 외 연구진[29]은 일반적인 환경을 포함하는 2차원 포즈 데이터와 실험실 환경에서 촬영된 3차원 포즈 데이터를 지식 전이(knowledge transfer) 방법으로 학습하여 일반적인 환경의 단일 이미지에서 3차원 포즈를 추정하는 방법을 제안하였다. 뿐만 아니라 제안하는 방법의 테스트를 위하여 마커 없는 모션캡쳐 방식을 통하여 얻어진 일반적인 환경의 3차원 포즈 추정에 대한 데이터 세트 MPI-INF-3DHP도 소개하였다<표 1>.

이미지에서 한 사람의 3차원 포즈를 추정하는 방법 외에 다수의 사람의 포즈를 추정하는 방법도 제안되었는데 Rogez 외 연구진[30]은 Ran 외 연구진[31]에 의해 제안된 객체 검출 방법인 Faster R-CNN에서 영감을 얻어 이미지의 서로 다른 위치에서 잠재적 기준 포즈를 제안(proposal)하고, 제안된 포즈에 분류기를 이용하여 점수를 부여하였다. 이후 회귀 방법으로 학습된 네트워크를 이용한 포즈 정교화 및 인접 포즈의 가설 통합을 통하여 여러 사람의 2차원 및 3차원 포즈를 동시에 추정할 수 있도록 하였다. Metha 외 연구진[32]도 일반적인 환경의 단일 이미지에서 여러 사람의 3차원 포즈 추정을 위한 새로운 싱글샷(single-shot) 방법을 제안하였다. 이 방법은 한 사람의 포즈 추정을 넘어 신체부위의 연관을 사용함으로써 단일 이미지에서 다수의 사람에 대한 3차원 포즈를 추정한다. 또한

제안된 방법을 학습하기 위하여 정교한 다수 사람의 상호작용과 실제 이미지를 보여주는 첫 번째 대규모 훈련 데이터인 MuCo-3DHP를 소개하였으며, 테스트를 위해 새로운 3D의 포즈가 주어지는 다중 사용자 테스트 세트 MuPoTs-3D도 소개하였다<표 1>.

최근에는 이미지 내 사람의 주요 관절의 3차원 포즈를 추정하는 것을 넘어 고밀도 깊이 정보를 추론하는 연구도 제안되었다. Alp Güler 외 연구진[33]은 RGB 이미지와 인체의 표면에 기반을 둔 표현 사이의 조밀한 대응관계를 확립하여 COCO 데이터 세트[9]에 기반을 둔 고밀도 깊이 정보를 포함하는 데이터 세트를 소개하였고, 컨볼루션 신경망을 학습함으로써 한 장의 이미지에서 사람의 고밀도 깊이 정보를 추정하는 방법을 제안하였다.

3. 국내의 연구 동향

3.1. 2차원 포즈 추정 기술

2차원 포즈 추정 연구와 관련하여 차은미와 이경미[34]의 연구에서는 사람의 머리, 몸통 등의 구성요소를 검출하고 추적한 후 이에 대한 확률 전파를 이용하여 사람의 2차원 자세를 추정하였다. 오수영과 한준희의 연구[35]에서는 관절단위의 파트가 아닌 의미론적으로 파트를 검출하고 영역별로 매칭된 파트로 자세를 추정하였다. 이와 같이 국내에서도 사람의 2차원 포즈 추정 관련 연구들이 존재하지만 2차원 포즈 추정에 관한 방법론적 연구보다는 이를 활용한 사람 추적, 사람 수 분석, 행동분석을 주제로 하는 논문들이 주로 이루어지고 있었다[36-37].

최근 사람의 2차원 포즈 추정과 관련하여 국내에서도 우수한 연구 결과가 발표되었다. 문경식 외 연구진[38]은 이전까지 제안된 사람의 2차원 포즈 추정 방법들이 유사한 오류 분포를 갖고 있음에 주목하여 이러한 오류를 수정하는 정교화 네트워크를 제안하였고 다양 포즈 인식 방법에 제안된 방법을 결합하는 경우에 포즈 인식 성능이 일관되게 개선되는 것을 확인하였다.

3.2. 3차원 포즈 추정 기술

국내에서는 사람의 2차원 포즈 추정 연구보다 3차원 포즈 추정 관련 연구들이 보다 활발히 진행되고 있다. 조정찬 외 연구진[39]은 3차원 포즈 추정 시 2차원 포즈 추정에서 발생되는 위험을 줄이고 학습과 테스트를 위한 데이터 세트가 서로 다른 경우에서의 성능 하락을 최소화하기 위하여 포즈 선택과 포즈 변환방법을 제안하였다. 또한 차건호 외 연구진[40]은 컨볼루션 신경망을 이용하여 2차원 단서로부터 3차원 추정에 대한 다수의 후보를 생성하고 이러한 정보를 종합하여 보다 강인한 하나의 3차원 포즈 추정결과를 도출하는 방법을 제안하였다.

박성헌 외 연구진[41]은 3차원 포즈 추정의 회귀 문제를 풀기 위하여 이미지 특징과 추정된 2차원 포즈 정보를 연결하여 3차원 포즈 추정 성능을 향상시키는 방법을 제안하였다. 장주용의 연구[42]에서는 2차원 포즈 추정에서 노이즈를 제거하고 동시에 3차원 복원을 수행하여 효과적으로 3차원 포즈를 추정하였다. 뿐만 아니라 장주용과 이경무의 연구[43]에서는 추정된 2차원 포즈 정보에 기반을 둔 무작위 필드(Conditional Radom Field; CRF) 모델과 2차원과 3차원 정보 사이의 회귀 분석을 이용하여 3차원 포즈를 추정하였다.

사람의 3차원 포즈 추정에 대해서 상당한 개선이 이루어졌으나 대부분의 경우 단일 사람에 대한 3차원 포즈를 추정하는 것에 집중되어 있었다. 최근 문경식 외 연구진[44]은 한 장의 RGB 이미지에 기반을 두어 다수 사람의 3차원 포즈를 추정하고, 완전한 학습에 기반을 둔 카메라 거리 인식의 하향식 접근 방법을 제안하였다. 이 방법은 사람의 포즈와 카메라와의 상대적 거리까지 추정이 가능하여 다수의 3차원 포즈 추정에 있어 더욱 입체적인 결과를 준다.

4. 사람의 포즈 추정관련 문제에 대한 고찰

본 장에서는 1) 현재 사용 가능한 데이터 세트와, 2) 포즈 추정을 이용한 대표적인 응용 사례를 살펴봄으로써 새로운 포즈 추정 연구에 대한 시사점을 도출하고자 한다.

4.1. 포즈 인식 연구를 위한 데이터 세트

앞서 살펴본 것과 같이 한 장의 이미지를 사용하여 정확하게 사람의 2차원 및 3차원 포즈를 추정하는 연구는 매우 중요한 분야이다. 이를 위한 실내외 환경에서의 포즈 분석에 대한 알고리즘 성능 평가를 위하여 <표 1>과 같은 데이터 세트가 주로 사용된다. 그러나 사람의 포즈 정보 수집은 객체 인식 데이터 수집보다 어려움이 있다. 이에 따라 사람의 포즈 인식 연구를 위한 데이터 세트와 데이터 세트에 포함된 이미지의 수가 이미지 분류 또는 객체 인식과 같은 연구에 비하여 부족한 실정이다.

최근 Fridman 외 연구진[49]은 운전자의 행동과 자동차와의 상호작용에 대한 연구에서 사람의 포즈 정보를 중요한 요소로 활용하였다. 이러한 상황에서는 특수한 자세뿐만 아니라 구조물에 의해서도 가림이 빈번하게 발생하기 때문에 이에 대한 강인한 포즈 인식은 기존 연구에서 해결하고자 하는 것과는 다른 문제에 해당된다. 그러나 이러한 특정 상황에서의 포즈 추정 기술에 대한 연구와 이를 위한 데이터 세트가 많지 않은 실정이다. 따라서 이러한 문제 해결을 위한 데이터 세트 개발이 필요하다.

4.2. 포즈 추정 기술을 확장한 연구사례

사람의 포즈는 행동 분석에 있어서는 핵심 정보라고 볼 수 있다. Wang 외 연구진[50]은 사람의 포즈에 대한 구조학적 정보를 공간적으로 모델링함으로써 행동인식에 있어 향상된 결과를 도출하였다. 행동인식과 관련하여 포즈정보의 중요성은 Jhuang 외 연구진[51]의 연구에도 잘 나타나 있다. 그들은 비디오에서 사람의 행동 분석을 목적으로 사용 가능한 특징을 상위 수준(high-level)의 사람 포즈 정보, 중간 수준(mid-level)의 광학적 흐름(optical flow), 하위 수준(low-level)의 이미지 기울기(gradient)로 구분하여 행동인식 성능을 측정하였다. 그 결과 상위 수준의 특징인 사람의 포즈 정보를 활용하는 것은 하위/중간 수준의 기본 특징을 보완하여 행동인식 성능을 개선 한다는 실험 결과를 발표하였다. 비디오에서의 사람의 행동 인식뿐만 아니라 이미지에 기반을 둔 사람의 행동인식에서도 사람의 포즈 정보는 중요한 특징으로 활용되고 있다. Zhao 외 연구진[52]은 이미지에서 추출된 포즈 정보는 관절의 의미론적 상호작용이라는 가정하에 신체 주요 관절 부위의 형태에 따라 의미를 부여하고 컨볼루션 신경망을 학습함으로써 기존의 방법과 비교하여 향상된 결과를 보여주었다.

최근에는 포즈 정보와 적대적 생성망을 활용하여 이미지를 생성하는 연구도 활발하게 진행되고 있다. Chan 외 연구진[53]은 포즈 정보를 활용한 동작 전달 방식을 제안하였는데, 이 연구에서는 전문가의 춤 동작이 있는 비디오 소스 데이터로부터 프레임별로 연속적인 포즈정보를 추출하였다. 그리고 적대적 생성망을 이용하여 포즈 움직임 정보를 타겟(target) 사람의 이미지 정보와 매핑함으로써 마치 타겟 사람이 실제로 춤을 추고 있는 것과 같은 비디오 생성 결과를 보여주었다. Neverova 외 연구진[54]은 서로 다른 뷰(view)에서 촬영된 한 사람의 영상정보와 표준화된 깊이정보모델을 공통 표면기반 좌표계(surface-based coordinate)에 매핑하여 적대적 생성망을 학습함으로써 입력되지 않은 포즈의 이미지를 자연스럽게 생성할 수 있는 방법을 제안하였다.

위에서 살펴본 것과 같이 포즈 인식 기술의 정확도 향상으로 인하여 포즈 인식기술을 활용한 연구 분야는 행동인식을 넘어서 행동 이미지 생성 분야에 까지 확장되고 있다. 이러한 이미지 생성 기술은 보다 정교한 가상의류 착용 기술, 학습을 위한 가상 데이터 세트 생성 및 비디오 영상 합성기술 등에도 활용할 수 있으므로 이에 대한 심도 있는 연구가 필요하다. 뿐만 아니라 이미지 생성기술의 발전으로

인하여 원본과 위조의 구분이 점차 어려워지고 있으며 이를 악용하는 것이 가능하므로 위조 이미지 및 위조 동영상 탐지 기술 역시 지속적으로 연구가 필요한 분야 중 하나이다.

5. 결론

본 연구에서는 단일 이미지를 기반으로 하여 사람의 2차원 추정과 3차원 추정에 대한 국내외 연구 동향을 분석하였고, 딥러닝이 등장하면서 사람의 포즈를 추정하는 기술이 상당한 기술적 진보를 이루었음을 확인하였다. 뿐만 아니라 사람의 포즈 추정 기술 연구에 활용되는 데이터 세트와 사람의 포즈 추정 정보를 활용하는 주요 연구들을 소개함으로써 본 분야에 대한 대략적인 흐름을 파악하는데 있어서 기초정보를 제공하였다.

해외에서 다양한 연구들이 이루어지고 있는 것에 비교하여 국내에서는 사람의 포즈와 관련된 연구들이 부족한 실정이다. 본 연구에서 소개한 사람의 포즈 인식 관련 연구 동향, 연구에 활용 가능한 데이터 세트 및 최신 응용 연구 사례들을 통하여 국내에서도 사람의 포즈 인식 기술에 대한 연구뿐만 아니라 이를 활용한 다양한 연구가 지속적으로 확대되길 기대한다.

Acknowledgement

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.B0101-15-0266, 실시간 대규모 영상 데이터 이해·예측을 위한 고성능 비주얼 디스커버리 플랫폼 개발)

참고문헌

[1] https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html

[2] 배창석, 김보경, "동영상 시맨틱 이해를 위한 시각 동사 도출 및 액션넷 데이터베이스 구축," 한국차세대컴퓨팅학회 논문지, 제14권, 제5호, Oct. 2018.

[3] 리준, 현종환, 최호진, "TSSN : 감시 영상의 강우량 인식을 위한 심층 신경망 구조," 한국차세대컴퓨팅학회 논문지, 제14권, 제6호, Dec. 2018.

[4] 최현종, 노대철, 김태영, "이중흐름 3차원 합성곱 신경망 구조를 이용한 효율적인 손 제스처 인식 방법," 한국차세대컴퓨팅학회 논문지, 제14권, 제6호, Dec. 2018.

[5] 후세인 탄베르, 칸 살만, 무함마드 칸, 이미영, 백성욱, "구조적인 유사성에 기반한 다중 뷰 비디오의 효율적인 키프레임 추출," 한국차세대컴퓨팅학회 논문지, 제14권, 제6호, Dec. 2018.

[6] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Jun. 2011.

[7] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in Proc. of the British Machine Vision Conference, Aug. 2010.

[8] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2014.

[9] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, and C. L. Zitnick, "Microsoft coco: Common objects in context," in Proc. of the European conference on computer vision, Sep. 2014.

[10] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2005.

[11] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li,

and L. Fei-Fei,"ImageNet: A Large-scale hierarchical image database,"in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2009.

〔12〕 A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks,"in Proc. of Neural Information Processing Systems, Dec. 2012.

〔13〕 A. Toshev, and C. Szegedy,"Deeppose: Human pose estimation via deep neural networks,"in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2014.

〔14〕 S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh,"Convolutional pose machines,"in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2016.

〔15〕 B. Sapp, and B. Taskar,"MODEC: Multimodal decomposable models for human pose estimation,"in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2013.

〔16〕 A. Newell, K. Yang, and J. Deng,"Stacked hourglass networks for human pose estimation,"in Proc. of the European conference on computer vision, Oct. 2016.

〔17〕 Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun,"Cascaded pyramid network for multi-person pose estimation,"in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2018.

〔18〕 B. Xiao, H. Wu, and Y. Wei,"Simple baselines for human pose estimation and tracking,"in Proc. of the European conference on computer vision, Sep. 2018.

〔19〕 K. He, X. Zhang, S. Ren, and J. Sun,"Deep residual learning for image recognition,"in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2016.

〔20〕 K. Sun, B. Xiao, D. Liu, and J. Wang,"Deep high-resolution representation learning for human pose estimation,"in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2019.

〔21〕 Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh,"OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields," arXiv preprint, arXiv:1812.08008, 2018.

〔22〕 E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer,"Single image 3D human pose estimation from noisy observations," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2012.

〔23〕 E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer,"A joint model for 2D and 3D pose estimation from a single image,"in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2013.

〔24〕 A. Agarwal and B. Triggs, "3D human pose from silhouettes by relevance vector regression," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2004.

〔25〕 M. E. Tipping, "The relevance vector machine," in Proc. of Neural Information Processing Systems, Nov. 2000.

〔26〕 I. Kostrikov, and J. Gall, "Depth sweep regression forests for estimating 3D human pose from images,"in Proc. of the British Machine Vision Conference, Sep. 2014.

〔27〕 G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis,"Coarse-to-fine volumetric prediction for single-image 3D human pose,"in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2017.

〔28〕 W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, "3D human pose estimation in the wild by adversarial learning,"in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2018.

〔29〕 D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3D human pose estimation in the wild using improved CNN supervision,"in Proc. of International Conference on 3D Vision, Oct. 2017.

〔30〕 G. Rogez, P. Weinzaepfel, and C. Schmid, "LCR-Net: Localization-classification-regression for human pose,"in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2017.

〔31〕 S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks,"in Proc. of Neural Information Processing Systems, Dec. 2015.

〔32〕 D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt, "Single-shot multi-person 3D pose estimation from monocular RGB,"in Proc. of International Conference on 3D Vision, Sep. 2018.

〔33〕 R. Alp Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild,"in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2018.

〔34〕 차은미,이경미,"구성요소 기반 확률 전파를 이용한 2D 사람 자세 추정,"한국HCI학회 학술대회, Feb. 2007.

〔35〕 오수영,한준희,"파트 기반 영역 매칭을 이용한 사람 자세 추정,"한국정보과학회 학술발표논문집, Dec. 2015.

〔36〕 이경미,김혜정,이윤미,"사람 자세 추정을 위한 모델 기반 추적,"한국HCI학회 학술대회, Feb. 2006.

〔37〕 박준혁, 이종석,"자세 추정을 이용한 비디오에서의 사람 수 측정 및 분류,"한국정보과학회 학술발표논문집, Dec. 2017.

〔38〕 G. Moon, J. Y. Chang, and K. M. Lee, "Posefix: Model-agnostic general human pose refinement network,"in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2019.

〔39〕 J. Cho, M. Lee, and S.i Oh, "Single image 3D human pose estimation using a procrustean normal distribution mixture model and model transformation," Computer Vision and Image Understanding, vol. 155, pp. 150-161, Feb. 2017.

〔40〕 G. Cha, M. Lee, J. Cho, and S. Oh, "Deep pose consensus networks," Computer Vision and Image Understanding, vol. 182, pp. 64-70, May 2019.

〔41〕 S. Park and N. Kwak, "3D human pose estimation with relational networks,"in Proc. of the British Machine Vision Conference, Sep. 2018

〔42〕 J. Chang, "DR-Net: Denoising and reconstruction network for 3D human pose estimation from monocular rgb videos," IET Electronics Letters, vol. 54, no. 2, pp. 70-72, 2018.

〔43〕 J. Chang and K. Lee, "2D-3D pose consistency-based conditional random fields for 3D human pose estimation," Computer Vision and Image Understanding, vol. 169, pp. 52-61, 2018.

〔44〕 G. Moon, J. Chang, and K. Lee, "Camera distance-aware top-down approach for 3D multi-person pose estimation from a single

RGB image,"in Proc. of the International Conference on Computer Vision, Oct. 2019.
〔45〕 S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation,"in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2011.
〔46〕 L. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," International journal of computer vision, vol. 87, no. 1-2, pp. 4, 2010.
〔47〕 C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, No. 7, Jul. 2014.
〔48〕 D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3D human pose estimation in the wild using improved CNN supervision,"in Proc. of International Conference on 3D Vision, Oct. 2017.
〔49〕 L. Fridman, D. E. Brown, M. Glazer, W. Angell, S. Dodd, B. Jenik, B., J. Terwilliger, A. Patsekin, J. Kindelsberger, L. Ding, S. Seaman, A. Mehler, A. Sipperley, A. Pettinato, B. Seppelt, L. Angell, B. Mehler, B. Reimer,"Mit autonomous vehicle technology study: Large-scale deep learning based analysis of driver behavior and interaction with automation,"arXiv preprint, arXiv:1711.06976, 2017.
〔50〕 C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition,"in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2013.
〔51〕 H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and. M. J. Black,"Towards understanding action recognition,"in Proc. of the International Conference on Computer Vision, Dec. 2013.
〔52〕 Z. Zhao, H. Ma, and S. You, "Single image action recognition using semantic body part actions,"in Proc. of the International Conference on Computer Vision, Oct. 2017.
〔53〕 C. Chan, S. Ginosar, T. Zhou, and A. A. Efros,"Everybody dance now,"arXiv preprint, arXiv:1808.07371, 2018.
〔54〕 N. Neverova, R. Alp Guler, and I. Kokkinos, "Dense pose transfer,"in Proc. of the European conference on computer vision, Sep. 2018.

저자소개

◆ 조정찬

•2010년 중앙대학교 학사
•2016년 서울대학교 박사
•2016년~2019년 삼성전자 책임연구원
•2019년~현재 가천대학교 소프트웨어학과 조교수
•관심분야: 딥러닝, 컴퓨터 비전