OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields

abstract

- realtime system & multi-persion & 2D pose estimation
- a nonparametric representation => Part Affinity Fields (PAFs)
  (PAFs = associate body parts with individuals in the image)
- bottom-up system
  high accuracy and realtime performance, regardless of the number of people in the image
  a runtime comparison to Mask R-CNN and Alpha-Pose
- 오늘 리뷰한 논문은 2019년 버전, 이전 버전 존재(CVPR, 2017) - 개선된 점
  1) 이전 버전은 PAFs and body part location estimation를 동시에 훈련, RAF-only refinement이 더 성능이 높음 (runtime, accuracy 둘 다)
  increases both speed and accuracy by approximately 200% and 7%
  2) combined body and foot key point detector도 body only detector와 성능을 유지할 수 있음. & annotated foot dataset도 배포
  3) 일반성 입증 (vehicle keypoint estimation에도 적용 가능)
  4) open-source release, 오픈포즈 - body, foot, hand, facial keypoints

1. Introduction

bottom-up, PAFs

- challenge of human estimation: multiple people의 pose estimation

  1) unknown number of people (despite of any position or scale)
  2) difficult parts association (예) contact, occlusion, or limb articulations)
  3) runtime complexity: due to the number of people in image

- top-down방식이 아니라 bottom-up의 방식으로 해결

  top-down 방식의 치명적인 약점: person detector를 실패하면, 보완할 방법 없음, runtime은 사람 수에 비례함(각 person detector 마다 a single-person pose estimator이 run함)
  bottom-up은 detect의 실패에 robust함. 사람 수와 runtime complexity 크게 상관x, 하지만 초기 bottom-up 모델은 실질적으로 이미지마다 수 분이 걸려서, final parse(global inference)에서 cost의 이점이 없음.

2. Related work

single person pose estimation

- Single Person Pose Estimation
  traditional approach - spatial model for articulated pose: 관절의 local observations과 관절들 사이의 spatial dependencies 의 조합으로 추론
  - tree-structured graphical models
    parametrically encode the spatial relationship between adjacent parts following a kinematic chain,
  - non-tree models
    tree structure 보강 with additional edges (occlusion, symmetry, and long range relationships를 포착하기 위해)
  - CNN 적용 for 신뢰적인 local observations of body parts (정확성 크게 향상)
    global contextual cues를 고려하기
    Pfister의 large receptive fields를 가진 네트워크 설계: global spatial dependencies를 implicit하게 포착
    Wei의 convolutional pose machines architecture(multi-stage architecture based on a sequential prediction framework): 부분적인 confidence maps 개선하기 위해 global context를 반복적으로 통합 & 이전 iteration에 의해 multimodal(다양한 형태의 데이터를 입력 데이터로 사용한다는 의미)의 불확실성 보존 & 각 end of stage에서 Intermediate supervisions 시행 for 훈련 중 vanishing gradients 문제를 해결하기 위해
    Newell도 a stacked hourglass architecture에서 Intermediate supervisions이 중요하다는 것을 입증

한계: Single Person Pose Estimation을 가정하고, location and scale of the person of interest이 주어짐

2. Related work

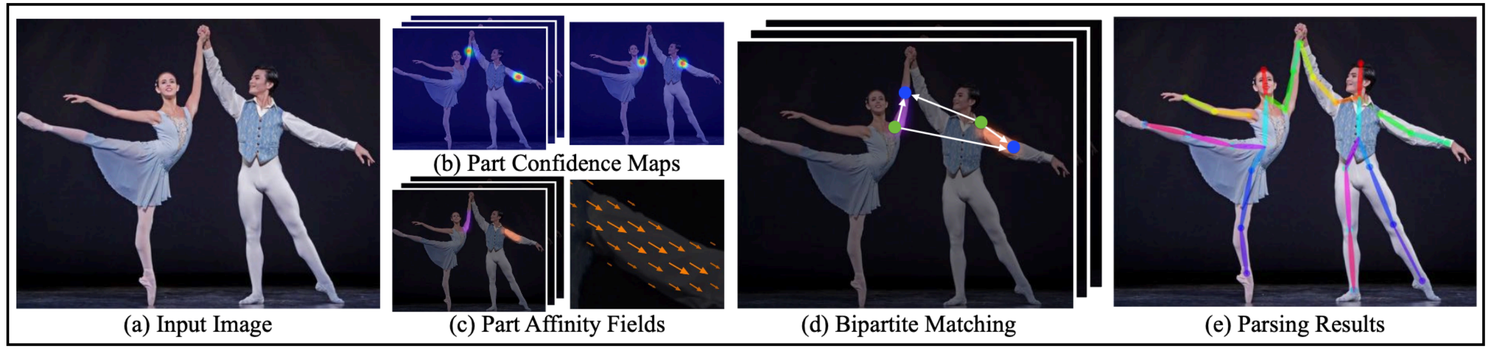
multi person pose estimation

- 과거 대부분의 multi person pose estimation의 Top-down 방식 이용
  - 1) detect people 2) estimate the pose of each person independently on each detected region
  - techniques developed for the single person case를 직접 적용 가능
  - early commitment on person detection & global inference를 요구하는 spatial dependencies (across different people) 포착 실패
  - 일부 접근법: inter-person dependencies 고려 시작
    - Eichner의 확장된 pictorial 구조: a set of interacting people and depth ordering 고려 but 여전히 person detector 요구
- bottom-up 방식
  - person detection에 의존 x & fully connected graph에서 integer linear programming를 해결하는 것은 NP-hard problem 야기 (단일 이미지 처리 시간이 몇 시간 걸림)
    - 그리디 방식
  - runtime 성능 개선 노력
- PAFs 제시: 여러사람의 관절들 사이에서 unstructured pairwise relationship를 인코딩하는 일련의 flow fields로 구성된 표현

  PAFs = a set of 2D vector fields that encode the location and orientation of limbs over the image domain
  PAF refinement is crucial for maximizing ac- curacy, while body part prediction refinement is not that important

3. 방법론

overall pipeline



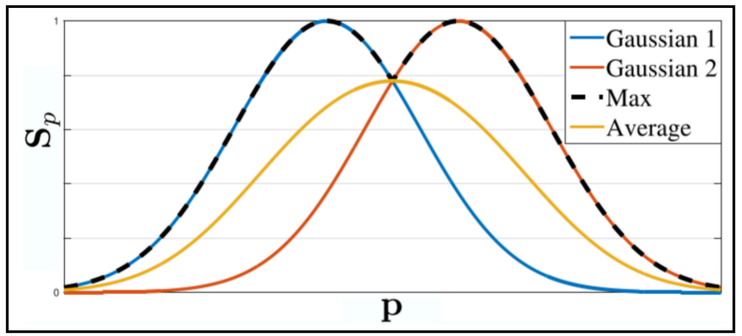
- overall pipeline
  - input: a color image of size w × h (a)
  - 1) 관절 찾기: 이미지에서 Part Affinity Fields (b) & Part Confidence Maps (c)를 거쳐 관절 예측
  - 2) 찾은 관절간의 관계 matching - Bipartite Matching (d)
  - output: 이미지의 모든 사람(each)에 대한 keypoints의 2D locations (e)
- 지도학습 - ground truth 존재
  - loss를 측정하기 위해 GT에 대한 confidence maps 생성

$$\mathbf{S}^*_{j,k}(\mathbf{p}) = \exp\left(-\frac{||\mathbf{p} - \mathbf{x}_{j,k}||_2^2}{\sigma^2}\right)$$

- 가우시안 분포 이용
- k번째 사람, j번째 관절에 대한 픽셀 p에서의 confidence map, x: k번째 사람의 j 번째 관절에 대한 GT

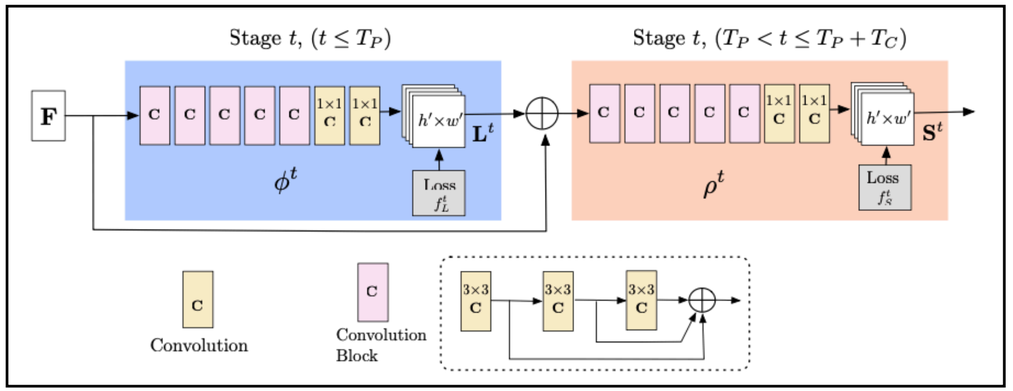
$$\mathbf{S}^*_j(\mathbf{p}) = \max_k \mathbf{S}^*_{j,k}(\mathbf{p})$$

- k번째 사람의 heat map, 각 사람의 관절 cm을 max aggregation
- peak가 중요한 정보이므로



3.1 Network Architecture

S와 L



- iterative prediction architecture
  - Blue: affinity fields that encode part-to-part association
  - Beige: detection confidence maps
  - refines the predictions over successive stages, t ∈ {1,...,T}, with intermediate supervision at each stage.
- 네트워크 2개로 구성 - Blue (네트워크 L)와 Beige (네트워크 S)
  - J: keypoint 개수, C: Limb 개수
  - 집합 L: limb 마다 PAFs(vecotor fields)를 가짐

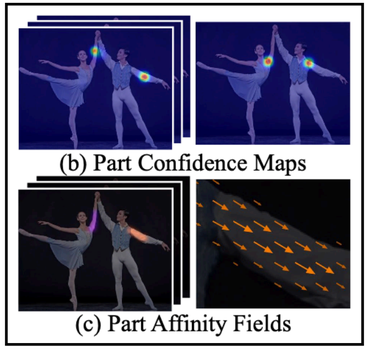
$$L = (L_1, L_2, ..., L_C)$$
$$L_c \in R^{w \times h \times 2}, c \in \{1...C\}$$

(L 결과는 c개의 관절관계만큼 있음)
(Lc = (w*h)*2크기의 실수 집합, 각 image location은 2차원 벡터를 인코딩)

  - 집합 S: keypoint마다 confidence maps를 가짐

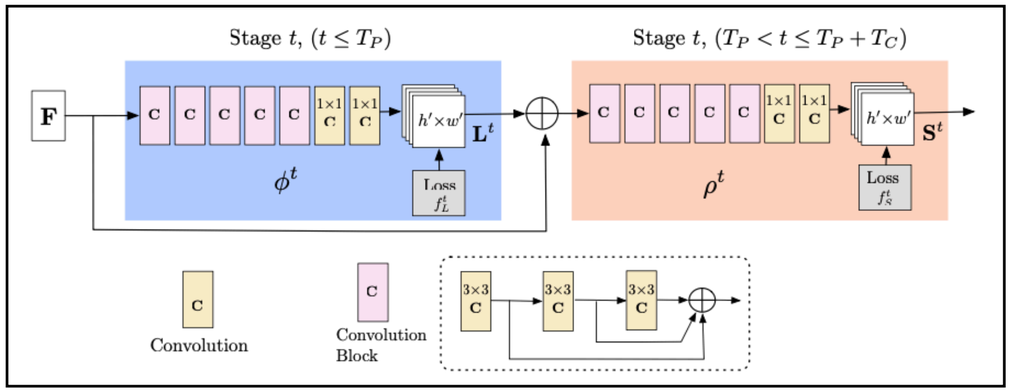
$$S = (S_1, S_2, ..., S_J)$$
$$S_j \in R^{w \times h}, j \in \{1...J\}$$

(S 결과는 j개의 관절만큼 있음)
(Sj = (w * h)크기의 실수 집합, 이미지에 대한 Heatmap)



- Confidence map & PAF는 greedy inference로 파싱함 - bipartite matchings(이분매칭)
  - 이분매칭 (=최대매칭): 가장 많이 연결되는 경우를 찾음

3.1 Network Architecture



- step
  - VGG-19로 추출된 feature F가 첫번째 stage의 input이 됨
  - 컨볼루션 연산을 통해 PAF - L 여러번 학습
  - most updated PAFs의 예측과 original image features F를 concat 후 CM - S 여러번 학습
    - PAFs는 t번의 스테이지를 반복한 결과, PAFs가 정밀할수록 전체 네트워크가 정밀해짐
  - 네트워크 직렬로 연결 (병렬에서 변경)
- prediction refine
  - **각 end of stage에서 Intermediate supervisions 시행**
    - Intermediate supervisions이란, 최종 예측 뿐만 아니라 중간 과정에서도 loss를 구함
    - 효과: 훈련 중 vanishing gradients 문제를 해결
  - **(원래 버전) several 7x7 convolutional kernels -> 3개의 연속된 3x3 convolutional kernels**
    - 계산 감소 (원래버전: 97 = 2*7^2-1, 현버전: 51 = (2*3^2-1)*3)
    - receptive field 보존
      - (각 stage의 입력 이미지에 대해 하나의 필터가 커버할 수 있는 이미지 영역의 일부 - 레이어가 깊어질수록 field는 선형적으로 증가)
    - 3개의 컨볼루션 연산에 대한 출력을 concat (DenseNet 접근법 인용)
    - non-linearity layers이 3배로 증가 => **네트워크는 lower level and higher level features 유지**
    - 이후, Sections 5.2 and 5.3에서, 정확도 속도 개선에 대한 분석 다룰 예정