DRPose3D: Depth Ranking in 3D Human Pose Estimation

Year & Index

- IJCAI-18
- DOI:10.24963/ijcai.2018/136

Abstract excerpt

- Pain points
  - accurate 3D positions can be hard to inference
  - 2D-to-3D pose regression in two-stage methods being ill-posed
  - the ground-truth can be obtained from manual labeling, the 3D pose is hard to get without sophisticated tracking devices
- Related works
  - Learning to Rank
    - point-wise — have to extract hand designed features for each item.
    - pairwise — only learns to rank one pair of pixels in an image explicitly
    - list-wise
  - 3D Pose Estimation
    - the end-to-end methods — still need 2D image and 3D pose pairs
    - the two- stage methods
      - usually predict 2D poses first, then use optimization or machine learning methods to obtain 3D pose results.
      - only focus on inferencing possible 3D poses from human body constrains — ignore other geometric knowledge embedded in the image features.
- Contribution
  - We design a Pairwise Ranking Convolutional Neural Net- work (PRCNN) to extract the depth rankings of pairwise human joints from a single RGB image
    - PRCNN trans- forms the depth ranking problem into the pairwise classifi- cation problem by generating a ranking matrix representing the depth relations between each pair of human joints
    - the depth ranking can be identified by human intuitively and learned using the deep neural network more easily by solving classification problems
  - We propose a coarse-to-fine 3D Pose Estimator named DP- Net composed of the DepthNet and the PoseNet.
    - it regresses the 3D pose from 2D joint locations and the depth ranking matrix
    - DPNet first estimates coarse depth value that is consistent with majority entries of the depth ranking matrix then regresses the accurate 3D poses in a coarse-to-fine manner.
  - Data augmentation in 3D space for the second stage is explored — By synthesizing 3D poses and camera param- eters, 2D poses and ranking matrices can be generated ad- equately.
    - unlike previous work, synthesized cameras are put around the same circle, which is unknown in real sce- narios
    - We randomly sample camera positions on all pos- sible positions around the subject.
    - To make the augmented data obey the data distribution of training dataset, we use a statistical method to add noises.

Experiment result

The proposed DRPose3D framework achieves the-state-of- the-art results on three common protocols of Human3.6M dataset compared with both end-to-end and two-stage meth- ods [Sun et al., 2017; Fang et al., 2018; Martinez et al., 2017]. Mean per joint position errors (MPJPE) on the three proto- cols are decreased to 57.8mm(2.2% ↓), 42.9mm(6.1% ↓) and 62.8mm(13.7% ↓) respectively. And the MPJPE gap between protocol #3 and protocol #l is reduced to 5. 0mm(59.7% ↓).

It proves that our method is robust to new camera positions and our data augmentation is very effective. The experimental results show that the depth ranking is an essential geometric knowledge that can be learned, utilized and augmented in 3D pose estimation.

Method

- PRCNN predicts ranking matrix M given the image I and the 2D joint heat maps H
  - We adopt an 8-stack hourglass network [Newell et al., 2016] as our 2D pose estimator.
    - pretrained on MPII dataset
    - fine-tuned on Human3.6M
  - Residual network I" [He et al., 2016] is used as the backbone of our feature extractor.
    - $$P_{ij} = \frac{e^{F_{ij}}}{1 + e^{F_{ij}}}.$$
    - probability that ith joint has higher Z-value than jth joint
    - $$\begin{aligned} C_{ij} \equiv C(F_{ij}) &= -M_{ij} \log P_{ij} - (1 - M_{ij}) \log(1 - P_{ij}) \\ &= -M_{ij} F_{ij} + \log(1 + e^{F_{ij}}). \end{aligned}$$
    - cross entropy loss
      The final cost function is the summation of all Cij
  - Different from RankNet, where each feature depends ononе item, PRCNN requires extracting all features from one image and predicts all of the pairwise rankings together. — l9-channel tensors (3 for image and 16 for heatmaps) as inputs for our model
- DPNet regresses 3D pose given the ranking matrix M and the 2D joint locations S2D.
  - use only the predicted ranking matrix P, and 2D pose S2D as input at this stage.
  - imperfect input
    - Directly learning S3D from the ranking matrix P and S2D provides less accurate results. — ?
    - A coarse- to-fine network is proposed to resolve noisy information from the ranking matrix.
      - first part : Depth- Net — predicts coarse depth O that is consistent with the ranking ma- trix.
        - The ground-truth of O is the ranking order on Z-axis of each human joint.
        - convert the noisy ranking matrix into coarse depth values
      - second part : PoseNet — combines coarse depth values with S2D and predicts more and more accurate 3D pose S3D — use a cascaded regression network with two stages
        - The first stage predicts the S3D directly
        - the second stage predicts its residuals
        - Each stage outputs are supervised by the 3D pose ground-truth S3D
  - $\mathcal{L} = \mathcal{L}_O + \mathcal{L}_{S_{1D}} + \mathcal{L}'_{S_{1D}}.$
    Loss
- Data Augmentation — ?