German Research Center for Artificial Intelligence

REAL-TIME POSE ESTIMATION VIA STATE-OF-THE-ART COMPUTER VISION MODELS

Nghia Duong-Trung, Benjamin Paaßen, Milos Kravcik
German Research Center for Artificial Intelligence (DFKI GmbH)
consortium MILKI-PSY meeting, Aachen, 12.12.2022

SPONSORED BY THE

Federal Ministry
of Education
and Research

Content

German Research Center for Artificial Intelligence

Introduction to Pose Estimation

What is pose estimation?

German Research Center for Artificial Intelligence

- Pose estimation and tracking is a computer vision task that includes detecting, associating, and tracking semantic key points.
  - human, animal, object.
- Offline (image, replayed video) or online (real-time video).
- SOTA methods are typically based on designing the CNN architecture tailored particularly for human pose inference.

Figure: Examples of pose predictions on sports, professional and casual photos. Image source: `https://arxiv.org/abs/2103.02440`.

The importance of pose estimation

German Research Center for Artificial Intelligence

- In tradition object detection, people are perceived as a bounding box
- Pose detection and tracking -> semantically understanding of human body
- High-performing real-time pose detection and tracking will drive some of the biggest trends in computer vision
  - sport, autonomous driving, animal research

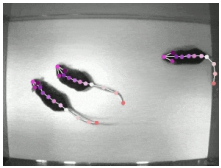


Figure: Example of animal pose predictions. Image source: `https://github.com/DeepLabCut`.

What is human pose estimation?

German Research Center for Artificial Intelligence

- Predict the locations of body parts and joints in images or videos: 2D and 3D
- Human pose modeling:
  - Kinematic: skeleton-based model: set of joint positions and limb orientations
  - Planar: contour-based model: the appearance and shape of a human body
  - Volumetric model: 3D estimation for recovering 3D human mesh

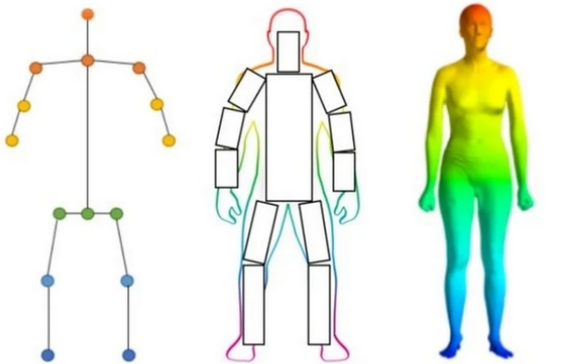


Figure: Human pose modeling. Image source: `https://arxiv.org/abs/2012.13392`.

How does pose estimation work?

German Research Center for Artificial Intelligence

- Estimate spatial positions of a body in an image or video
  - multi-pose, single pose
- Human pose estimation on the popular MS COCO dataset[a] can detect 17 different keypoints. Each keypoint is annotated with a tuple (x, y, v), where x and y mark the coordinates, and v indicates if the keypoint is visible.

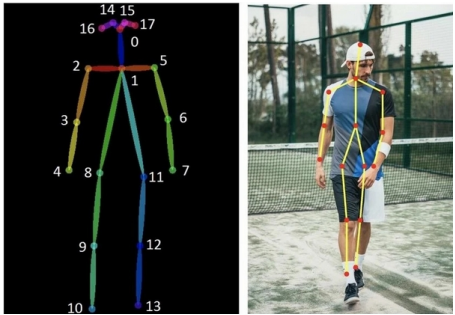


Figure: Human pose modeling. Image source: `https://viso.ai/deep-learning/openpose/`.

[a] `https://cocodataset.org/#download`

Most popular pose estimation methods

German Research Center for Artificial Intelligence

1. Deep Pose, CVPR 2014
2. PoseNet, ICCV 2015
3. DeepCut, CVPR 2016
4. OpenPose, CVPR 2017
5. Regional Multi-Person Pose Estimation (AlphaPose), CVPR 2017
6. Dense Pose, CVPR 2018
7. Deep High-Resolution Net (HRNet), CVPR 2019
8. BlazePose (MediaPipe), CVPR 2020

Use cases and applications

German Research Center for Artificial Intelligence

- Human-computer interaction, action recognition, motion analysis, augmented reality, sports and fitness, and robotics.
  - Detect gestures
  - Support the analysis of football, basketball, and sports
  - Analyze dance techniques
  - Analyze posture for body works and finesses
  - Play games via virtual tutors
  - Human tracking for consoles
- Future trends: transfer learning and edge AI

Introduction to YOLO Family Models

You only look once (YOLO)

German Research Center for Artificial Intelligence

- J Redmon et al. You only look once: Unified, real-time object detection[1]. CVPR 2016. 29950 citations
  - YOLOv1 to YOLOv3
- YOLOv4, introduced by A. Bochkovskiy et al. in 2020
- Scaled-YOLOv4, introduced by CY Wang et al. in 2021
- YOLOR, YOLOX, NanoDet-Plus
- PP-YOLOE, an industrial object detector PaddlePaddle, 2019
- YOLOv5 model v6.1, Ultralytics, 2022
- YOLOv7, CVPR 07.2022, `https://arxiv.org/abs/2207.02696`
  - There are actually two YOLOv7, the other developed on Detectron2

[1] `https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Redmon_You_Only_Look_CVPR_2016_paper.pdf`

YOLOv7 vs other object detectors

German Research Center for Artificial Intelligence

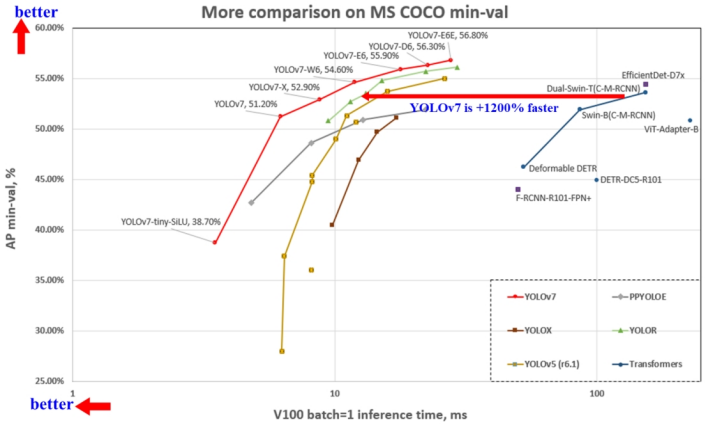


Figure: Image source: `https://arxiv.org/abs/2207.02696`, `https://github.com/WongKinYiu/yolov7`.

YOLO-Pose

YOLO-Pose paper

German Research Center for Artificial Intelligence

- Maji, D., Nagori, S., Mathew, M., & Poddar, D. (2022). YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2637-2646) [2].
  - `https://github.com/ultralytics/yolov5/`
  - `https://github.com/TexasInstruments/edgeai-yolov5`
  - `https://github.com/TexasInstruments/edgeai-yolox`
  - `https://github.com/WongKinYiu/yolov7/blob/main/tools/keypoint.ipynb`
  - PyTorch, ONNX, CoreML, TFLite
  - tested on image, video, webcam

[2] `https://openaccess.thecvf.com/content/CVPR2022W/ECV/papers/Maji_YOLO-Pose_Enhancing_YOLO_for_Multi_Person_Pose_Estimation_Using_Object_CVPRW_2022_paper.pdf`

Top-down and bottom-up approaches

German Research Center for Artificial Intelligence

- Multi-person 2D pose estimation can be categorized into top-down and bottom-up approaches
- Top-down: or two-stage approaches first perform human detection using a heavy person detector followed by estimating 2D pose for each detected person
  - Computational complexity increases linearly with the number of persons
- Bottom-up: find out identity-free keypoints of all the persons in an image in a single shot followed by grouping them into individual person instances
  - Rely on heatmaps to detect all the keypoints; require a complex post-processing to group keypoints

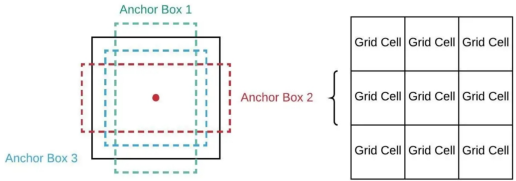
YOLO-Pose idea

German Research Center for Artificial Intelligence

- YOLO-Pose is a single shot approach like other bottom-up approaches but it does not use heatmaps.
- YOLO-Pose associates all keypoints of a person with anchors.
- Keypoints associated with an anchor are already grouped
- Independent of the number of persons in an image
- End-to-end training without independent post-processing
  - Extend the idea of IoU loss from box detection to keypoints
  - Object keypoint similarity (OKS) is used for evaluation and a loss for training

What is anchor (box) in object detection

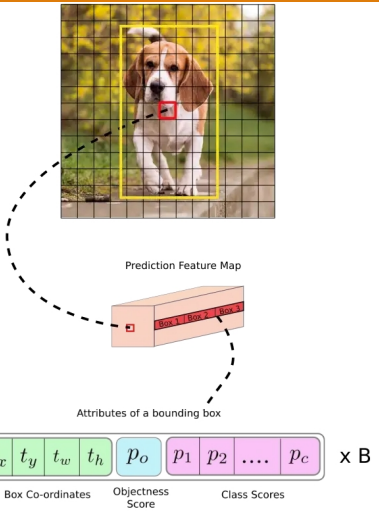
- Anchor box is a prior box that could have different pre-defined aspect ratios determined before training by running K-means on the entire dataset or a hyperparameter
- The output of convolution is a matrix as a "grid", then we assign anchor (boxes) to the grid cells, and they share the same centroid
- One we define those anchors, we can determine how much the ground truth box overlaps with the anchor boxes and pick the one with the best IoU and couple them together



Anchor example

German Research Center for Artificial Intelligence

- Each cell is assigned 3 anchors containing some set of properties (x,y,w,h,object score, classes)



Anchor in YOLO-Pose

German Research Center for Artificial Intelligence

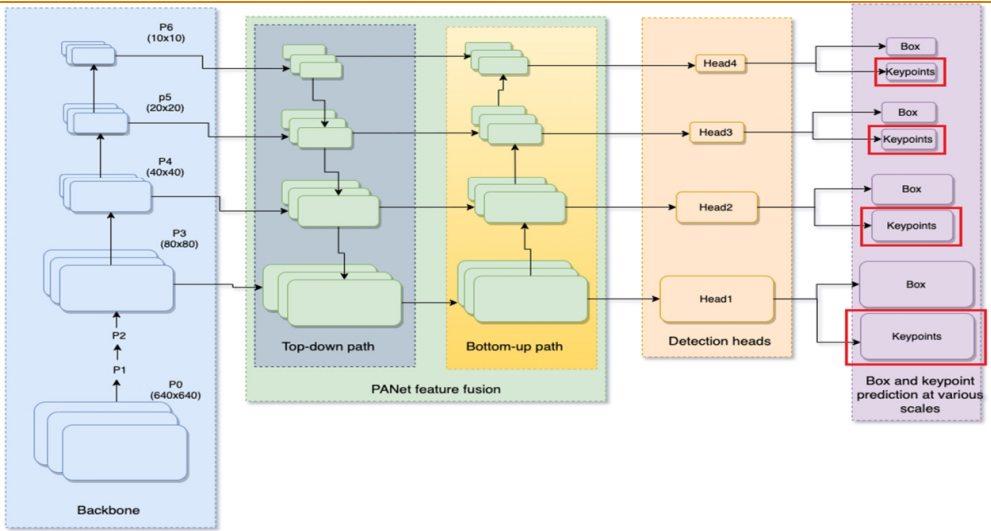
- It boils down to a single class person detection problem with each person having 17 associated keypoints, and each keypoint is again identified with a location and confidence: $\{x, y, conf\}$
- The overall prediction vector is defined as:

$$P_v = \{C_x, C_y, W, H, box_{conf}, class_{conf}, K_x^1, K_y^1, K_{conf}^1, \ldots, K_x^n, K_y^n, K_{conf}^n\} \quad (1)$$

where a box head predicts 6 elements and a keypoint head predicts 51 elements.

YOLO-Pose architecture

German Research Center for Artificial Intelligence



IoU based bounding-box loss function

German Research Center for Artificial Intelligence

For a ground truth bounding box that is matched with k^{th} anchor at location (i, j) and scale s, loss is defined as:

$$\mathcal{L}_{box}(s, i, j, k) = (1 - CIoU(Box_{gt}^{s,i,j,k}, Box_{pred}^{s,i,j,k})) \tag{2}$$

where *CIoU* is defined in another paper [3]

[3] https://ojs.aaai.org/index.php/AAAI/article/view/6999/6853

Human pose loss function formulation

German Research Center for Artificial Intelligence

- Object keypoint similarity (OKS) is the most popular metric for evaluating keypoints
- Corresponding to each bounding box, YOLO-pose model stores the entire pose information
  - If a ground truth bounding box is matched with k^{th} anchor at location (i, j) and scale s, the model predicts the keypoints with respect to the center of the anchor
  - OKS is computed for each keypoint separately and then summed to give the final OKS loss or keypoint IoU loss

Human pose loss function formulation

German Research Center for Artificial Intelligence

$$\mathcal{L}_{kpts}(s,i,j,k) = 1 - \frac{\sum_{n=1}^{N_{kpts}} \exp(\frac{2s^2k_n^2}{d_n^2})\delta(v_n > 0)}{\sum_{n=1}^{N_{kpts}} \delta(v_n > 0)} \tag{3}$$

where

$d_n =$ Euclidean distance between predicted and ground truth location for n^{th} keypoint

$k_n =$ keypoint specific weights

$s =$ scale of an object

$\delta(v_n) =$ visibility flag for each keypoint

Human pose loss function formulation

Corresponding to each keypoint, we learn a confidence parameter that shows whether a keypoint is present for that person or not

$$\mathcal{L}_{kpts\_conf}(s,i,j,k) = \sum_{n=1}^{N_{kpts}} BCE(\delta(v_n > 0), p^n_{kpts}) \tag{4}$$

where

p^n_{kpts} = predicted confidence for n^{th} keypoint

Human pose loss function formulation

Loss at location (i, j) is valid for k^{th} anchor at scale s if a ground truth bounding box is matched against that anchor. Finally, the total loss is summed over all scales, anchors and locations:

$$\mathcal{L}_{total} = \sum_{s,i,j,k} \left(\lambda_{cls}\mathcal{L}_{cls} + \lambda_{box}\mathcal{L}_{box} + \lambda_{kpts}\mathcal{L}_{kpts} + \lambda_{kpts\_conf}\mathcal{L}_{kpts\_conf}\right) \tag{5}$$

where $\lambda_{cls}, \lambda_{box}, \lambda_{kpts}, \lambda_{kpts\_conf}$ are hyperparameters chosen to balance between losses at different scales

YOLO-pose pretrained model

German Research Center for Artificial Intelligence

- The YOLO-pose model was trained on COCO dataset[4]
  - 200,000 images with 250,000 person instances with 17 keypoints
- train2017 (57k images), val2017(5k images), test-dev2017(20k images)
- `https://github.com/WongKinYiu/yolov7/tree/pose`
- `https://github.com/WongKinYiu/yolov7/releases/download/v0.1/yolov7-w6-person.pt`

[4] `https://cocodataset.org/#home`

YOLO-Pose: Demo

Demo

German Research Center for Artificial Intelligence

- `https://github.com/MILKI-PSY/yolov7-pose-bicep`
- image, video
  - Examples of teacher and students:
    `https://github.com/MILKI-PSY/yolov7-pose-bicep/tree/main/input`
  - After applying YOLO-pose, see the output files at
    `https://github.com/MILKI-PSY/yolov7-pose-bicep`
- Live demo via webcam

Remarks and Discussion

Remarks and discussion

German Research Center for Artificial Intelligence

- SOTA object detection and pose estimation model, all-in-one model
- Privacy and sensible information are preserved at the partners server
- Flexible to integrate more exercises estimation
- Flexible to input data
- Further recommendation

Thank you for listening