2D to 3D pose reconstruction In the wild

SHOBHIT MALARYA - S20160010087
SIDDHARTH KUMAR - S20160010090

Dr. Snehasis Mukherjee

Outline of the presentation

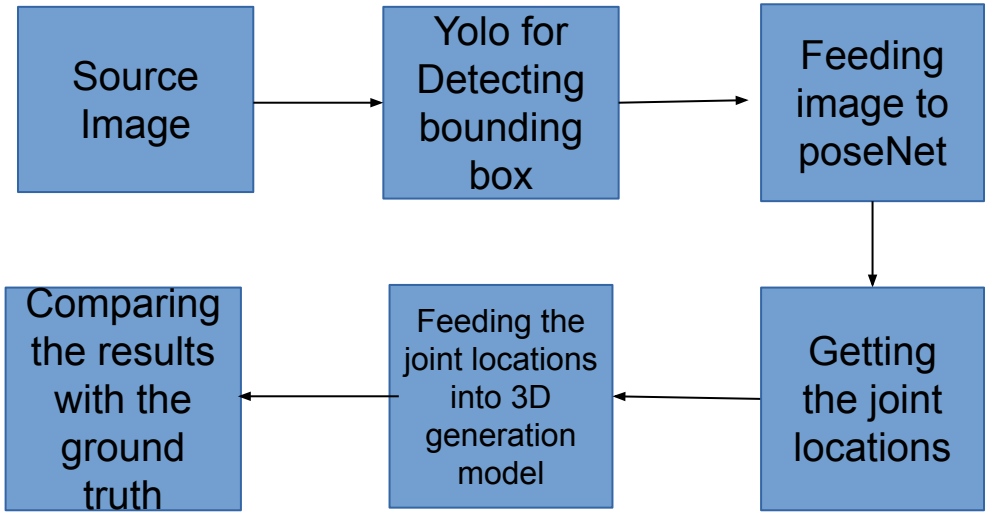
- Introduction
- Literature review
- Work done
- Discussion of results
- Summary of conclusions
- Future scope

Work done till last review

- Studying papers
  - End-to-end Recovery of Human Shape and Pose by Angjoo Kanazawa
    CVPR 2018
  - Convolutional pose machine by Shih-En Wei
    CVPR 2016
  - PersonLab by Google Inc.
    CVPR 2018

Work done till last review

- Framing Pipeline



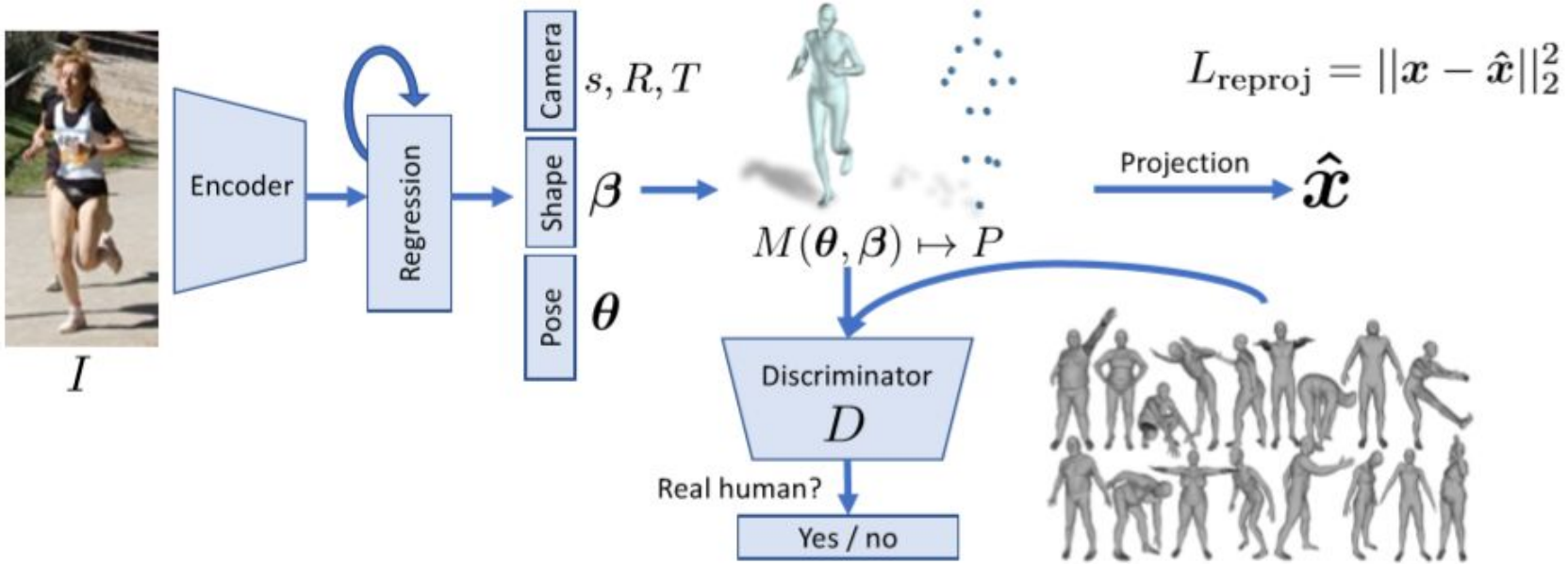
Introduction

Construction of 3D human model from 2D images.

Image source - pinterest.com

Literature review

End-to-end Recovery of Human Shape and Pose



Literature review

- Given an image, the network has to infer the 3D mesh parameters and the camera such that the 3D keypoints match the annotated 2D key points after projection.
- During training they assume that all images are annotated with ground truth 2D joints.They also consider the case in which some have 3D annotations as well. Additionally we assume that there is a pool of 3D meshes of human bodies of varying shape and pose. Since these meshes do not necessarily have a corresponding image, we refer to this data as unpaired

Literature review

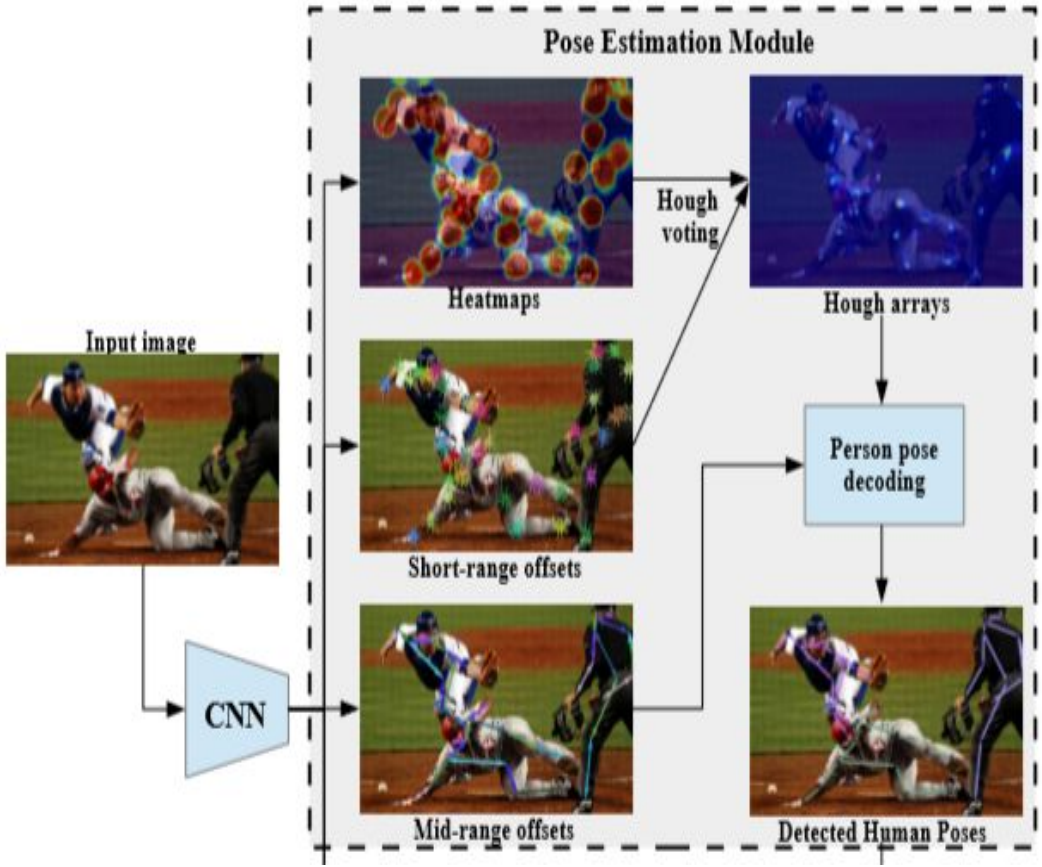
- The inferred parameters are also sent to a discriminator network whose task is to determine if the 3D parameters are real meshes from the unpaired data.
- 3D Body Representation - the set of parameters that represent the 3D reconstruction of a human body is expressed as a 85 dimensional vector $\Theta = \{\theta,\beta,R,t,s\}$. Given Θ, the projection of $X(\theta,\beta)$ is

$$\hat{x} = s\Pi(RX(\theta,\beta)) + t$$

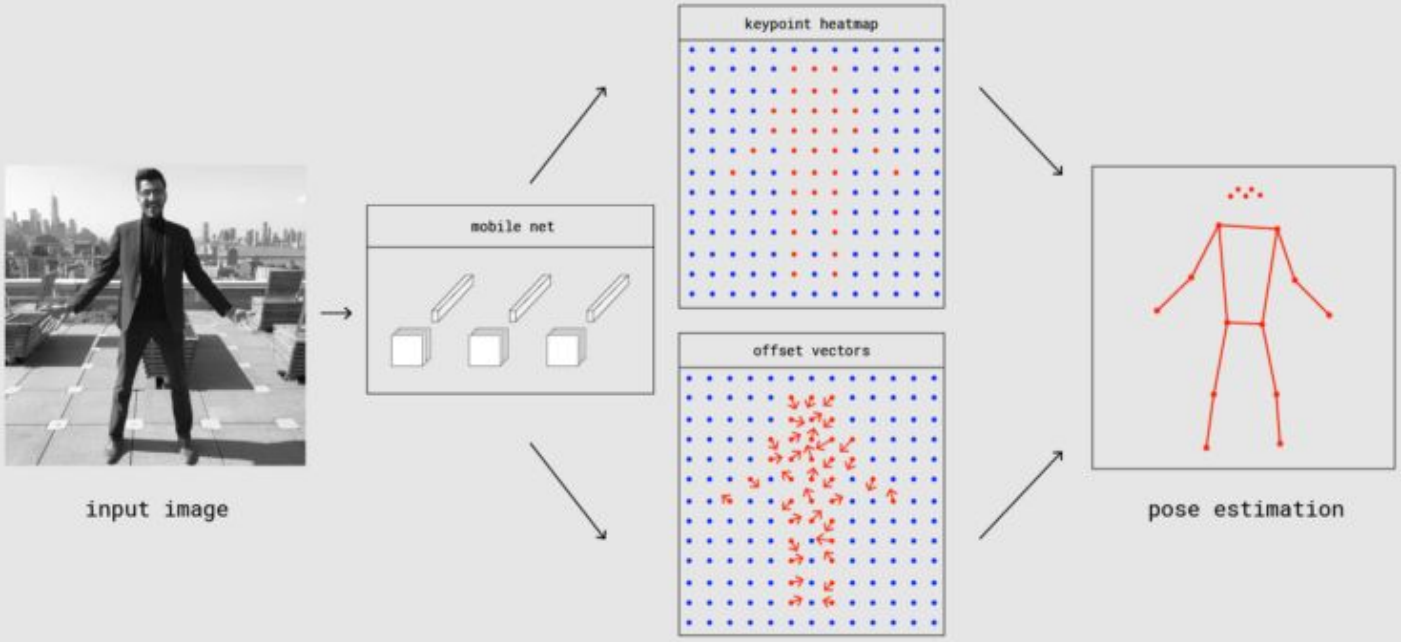
where Π is an orthographic projection.

PersonLab:

Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model



Single-Pose Detection Algorithm



PoseNet model

Output Stride and Heatmap Resolution

input image

225 x 225
(width x height)

32
8x8x17

16
15x15x17

8
29x29x17

heatmap

output stride

width resolution x height
resolution x num keypoints

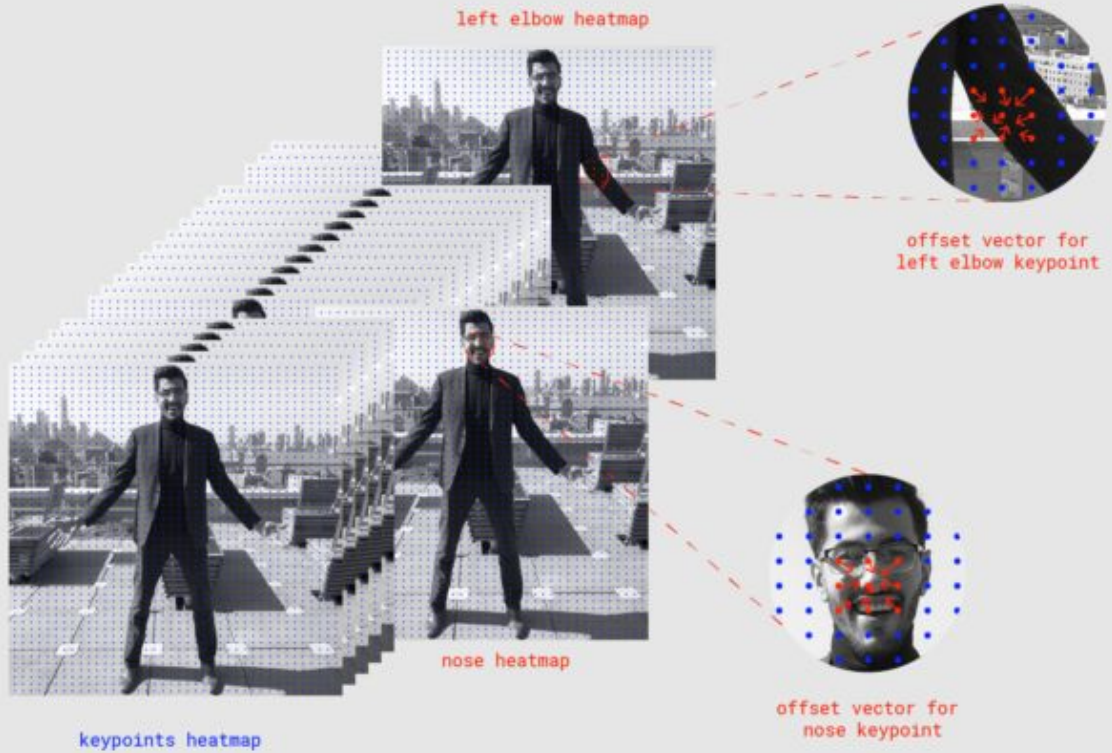
faster

slower

lower accuracy

higher accuracy

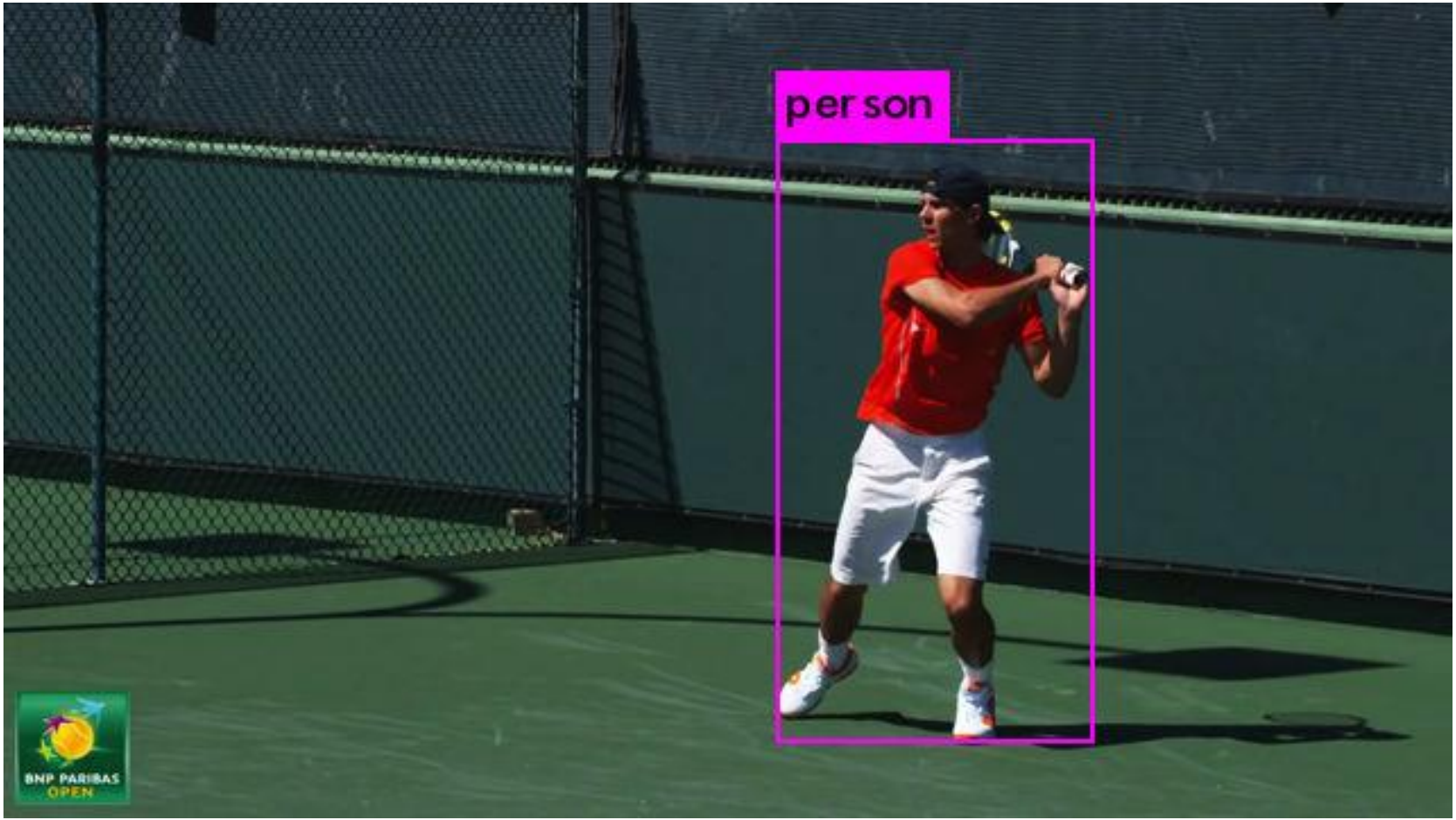
Heatmap and Offset Vector Simplification



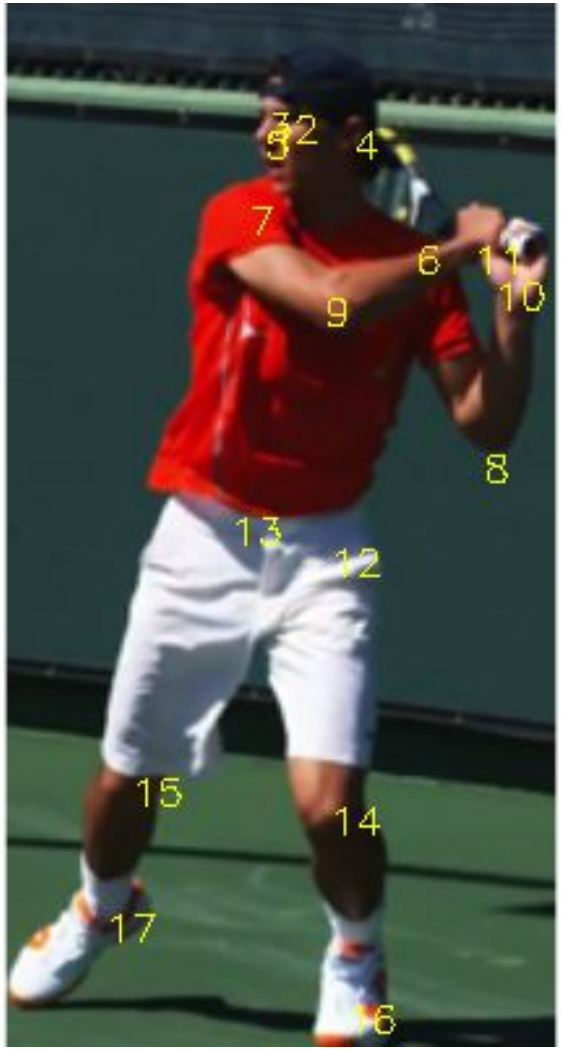
WORK DONE

SOURCE IMAGE

BOUNDING BOX



JOINT PREDICTION RESULT



Issues and Discussion of results

input

joint projection

3D Mesh overlay

3D mesh

diff vp

diff vp

input

joint projection

3D Mesh overlay

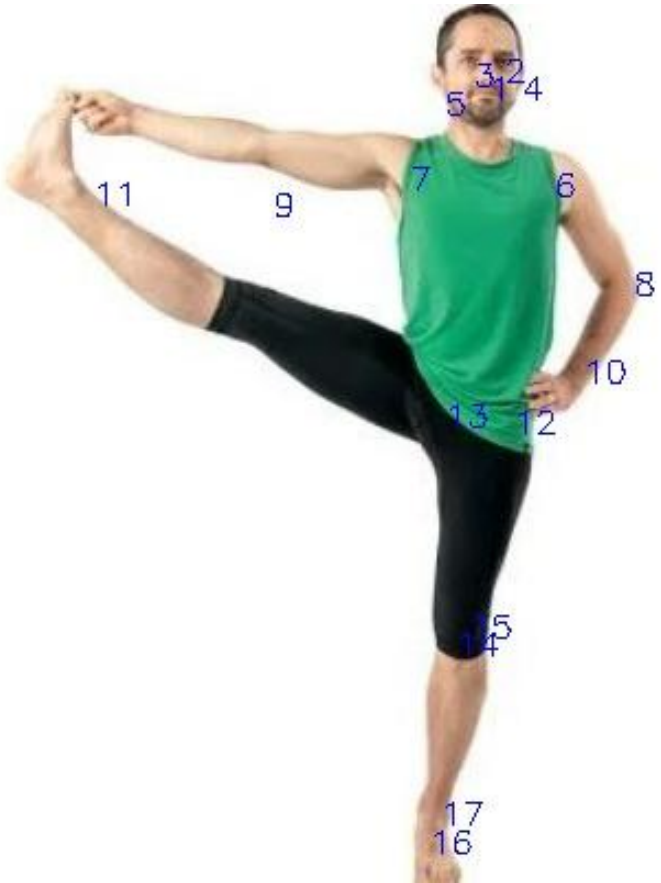
3D mesh

diff vp

diff vp

JOINT OVERLAPPING

IMAGE WITH WHITE BACKGROUND



Summary & Conclusions

- PoseNet gives all the 17 points although some are not present.
- Overlapping of few joint points especially knees and ankles.
- PoseNet is showing higher score even for true negatives.
- Cropping gives a better result than results on original image.
- Fine tuning of parameters may lead to better results.

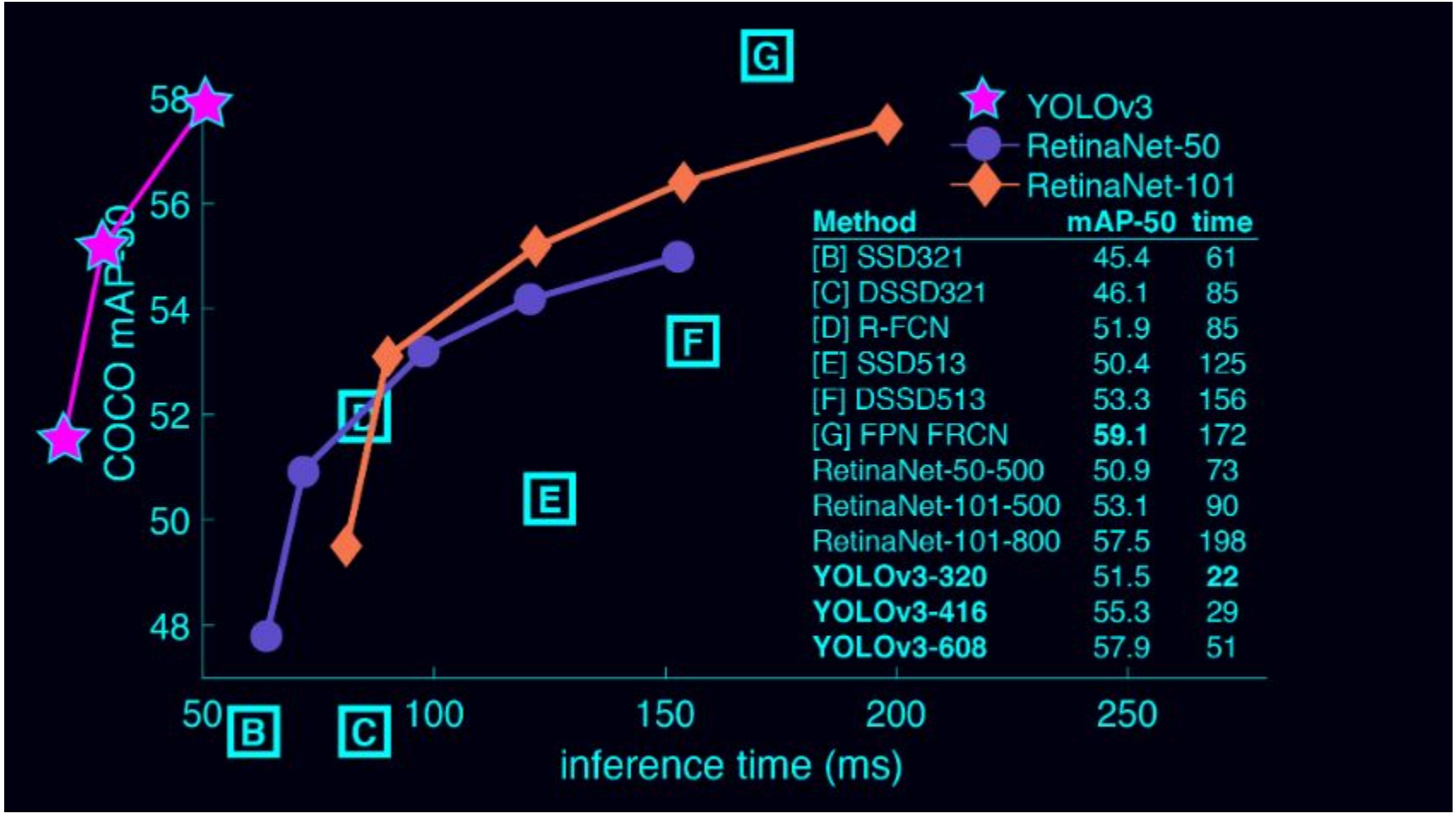
Future scope

- Increasing accuracy of 2D joint points.
- If possible then working on video instead of image.
- Observing the 3D mesh output

REFERENCES

- End-to-end Recovery of Human Shape and Pose https://github.com/akanazawa/hmr
- End-to-end Recovery of Human Shape and Pose https://arxiv.org/abs/1712.06584
- PoseNet https://medium.com/tensorflow/real-time-human-pose-estimation-in-the-browser-with-tensorflow-js-7dd0bc881cd5
- PoseNet https://github.com/tensorflow/tfjs-models/tree/master/posenet
- PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model https://arxiv.org/abs/1803.08225

THANK YOU


G
YOLOv3
RetinaNet-50
RetinaNet-101
58
56
54
52
50
48
COCO mAP-50
F
D
E
B
C
50
100
150
200
250
inference time (ms)
Method mAP-50 time
[B] SSD321 45.4 61
[C] DSSD321 46.1 85
[D] R-FCN 51.9 85
[E] SSD513 50.4 125
[F] DSSD513 53.3 156
[G] FPN FRCN 59.1 172
RetinaNet-50-500 50.9 73
RetinaNet-101-500 53.1 90
RetinaNet-101-800 57.5 198
YOLOv3-320 51.5 22
YOLOv3-416 55.3 29
YOLOv3-608 57.9 51