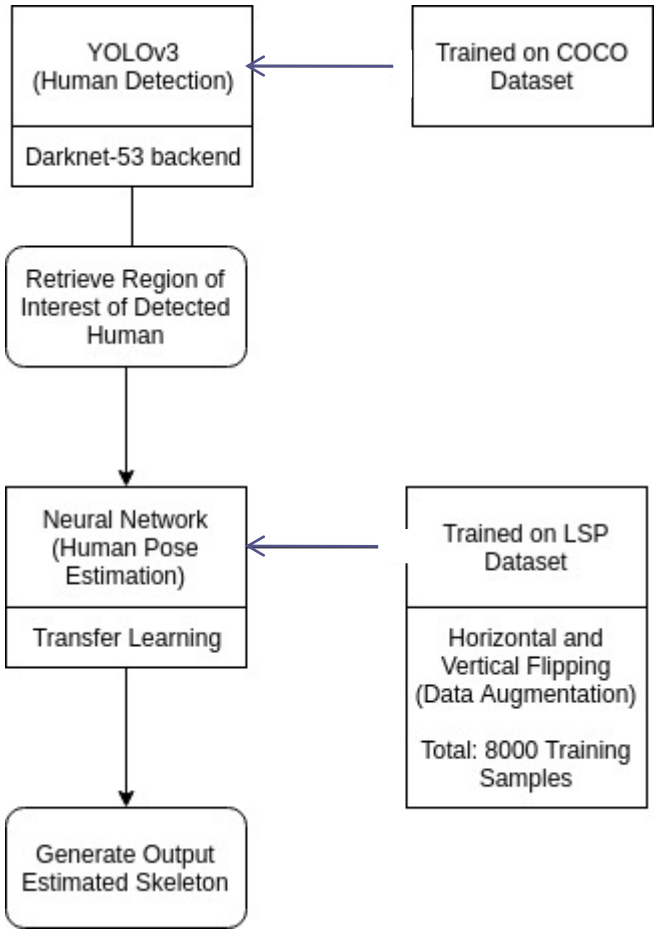
HUMAN POSE ESTIMATION
USING DEEP LEARNING APPROACHES

Exploratory Project

Lakshya Gupta(17075029)
Janhavi Gupta(17075061)
2nd-Year(B.Tech)
Department of Computer Science and Engineering

Framework

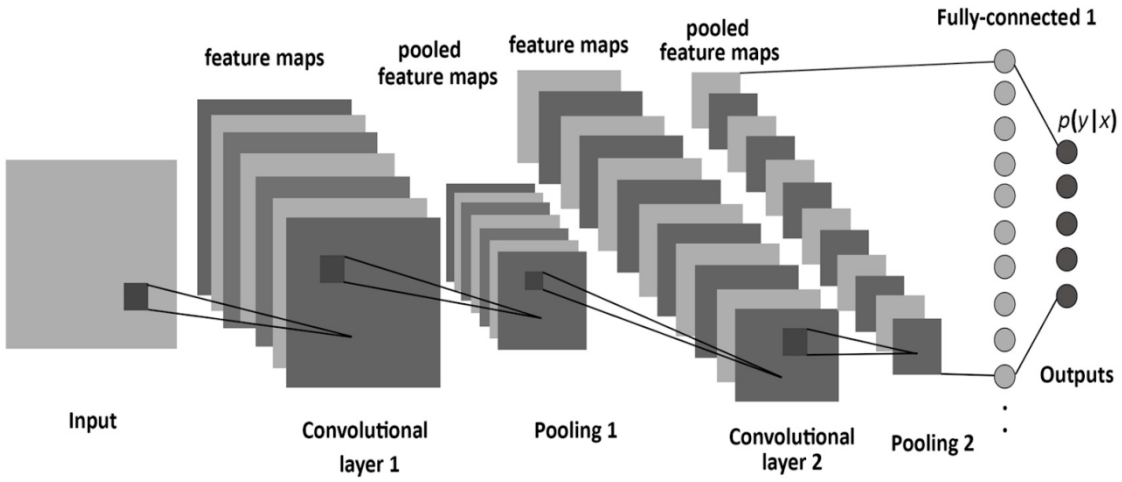

YOLOv3
(Human Detection)
Darknet-53 backend
Trained on COCO
Dataset
Retrieve Region of
Interest of Detected
Human
Neural Network
(Human Pose
Estimation)
Transfer Learning
Trained on LSP
Dataset
Horizontal and
Vertical Flipping
(Data Augmentation)
Total: 8000 Training
Samples
Generate Output
Estimated Skeleton

Convolutional Neural Networks

Deep neural networks used for analysing visual imagery.

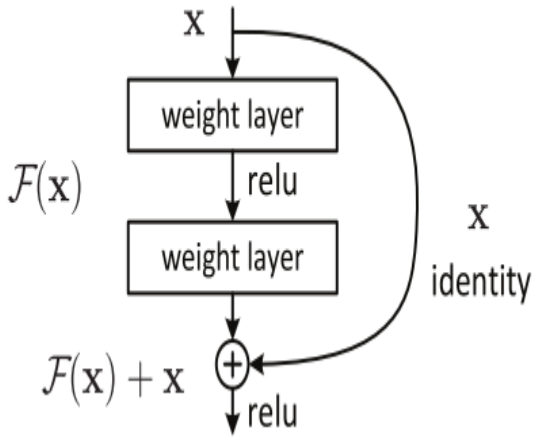
Includes:

- Convolutional layer
- Pooling layer
- Activation layer
- Normalisation layer
- Fully Connected layer



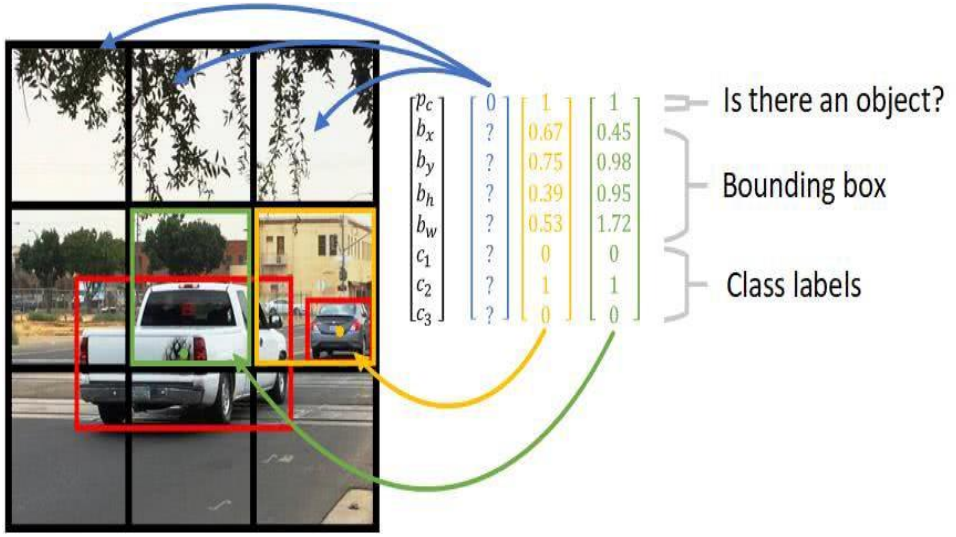
Residual Networks

- Involves residual blocks to skip one or two layers.
- A solution for the accuracy saturation problem in deeper neural networks.
- Works on the idea of “identity-shortcut connection”.



YOLO Algorithm

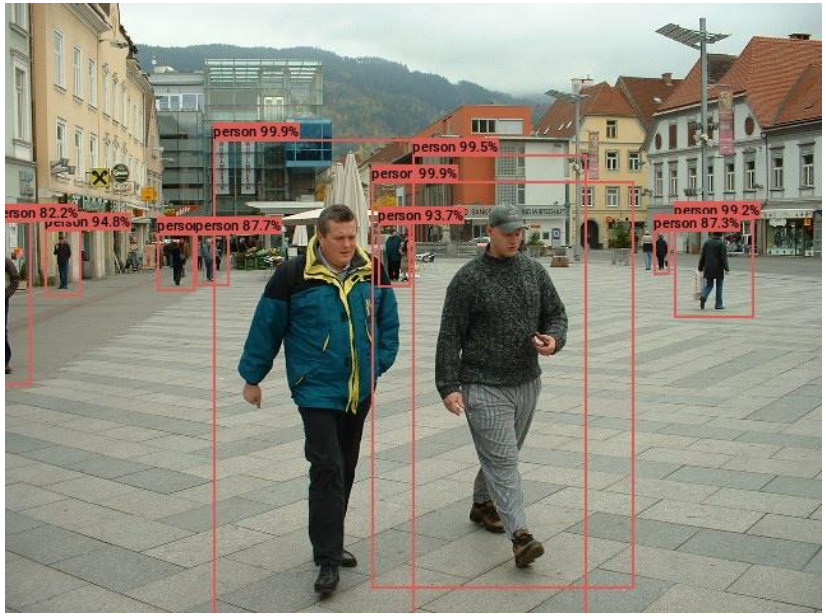
- Object detection algorithm.
- Detects the object in a single pass through neural network.
- Divides the image into grids.
- Predicts the bounding boxes with probability scores for each grid.



Human Detection

Using YOLOv3 with a residual network(DarkNet-53) as backend

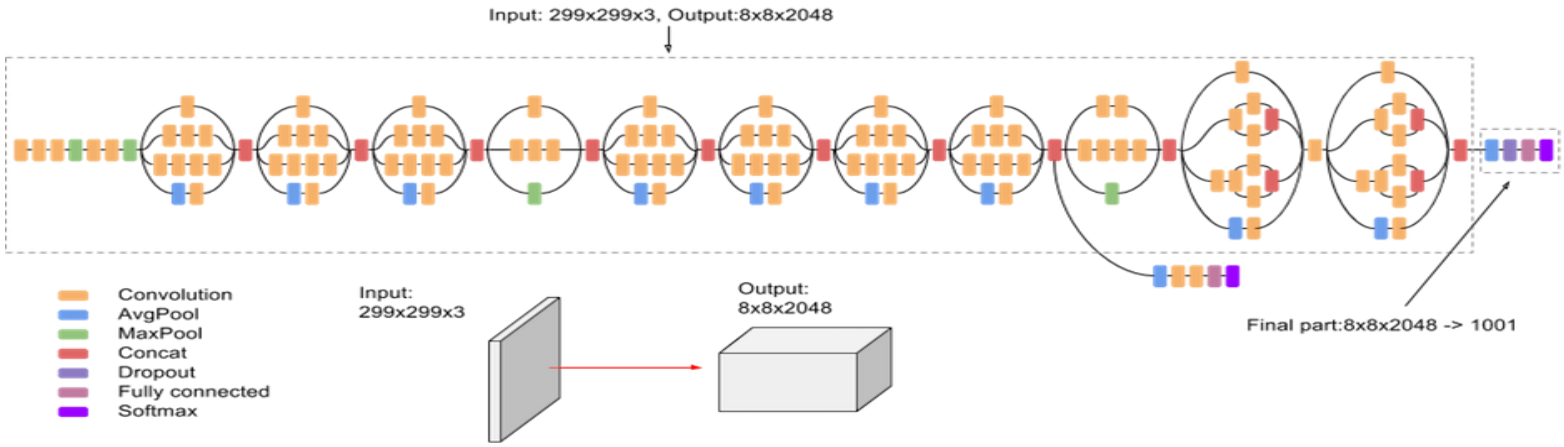
- The network is trained with DarkNet-53 as backend on COCO dataset.
- In this version of YOLO, predictions are made thrice at different locations.



Pose Estimation

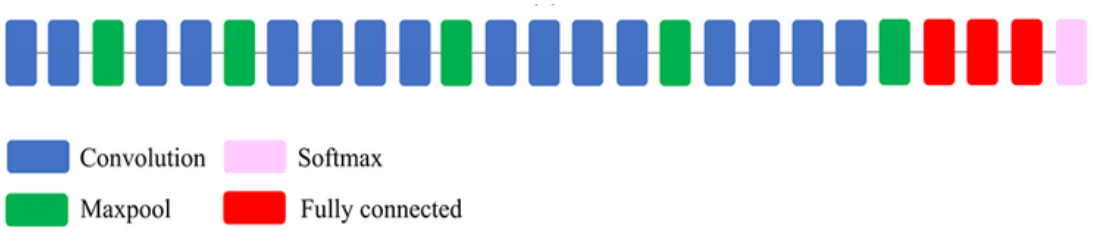
- Generating a skeleton for the detected human
- For each image, labels for 14 joints on the body are predicted:- right ankle, right knee, right hip, left hip, left knee, left ankle, right wrist, right elbow, right shoulder, left shoulder, left elbow, left wrist, neck, head top
- Training is done on LSP dataset and data augmentation is used to increase the size of dataset to 8000 images.

InceptionV3



- In Inception network, we go through different pathways and we concatenate them.
- In InceptionV3, we removed the end layers and modified the structure by adding some layers(Dropout, Dense, RELU activation, Batch Normalisation).
- Total number of parameters: 41,210,146
- Trainable parameters: 20,454,914
- Non-trainable parameters: 20,755,232

VGG-19



- Variant-1
- 23 layers
- Total number of parameters: 27,085,886
- Trainable parameters: 16,497,734
- Non-trainable parameters: 10,588,152

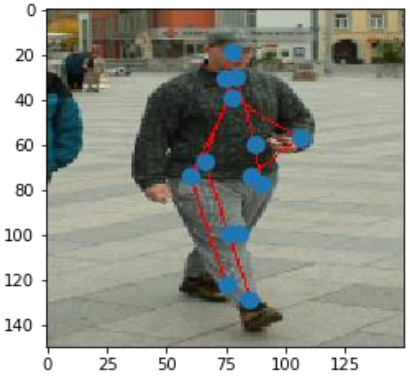
- Variant-2
- 21 layers
- Total number of parameters: 28,945,986
- Trainable parameters: 18,357,762
- Non-trainable parameters: 10,588,224

Adam optimisation

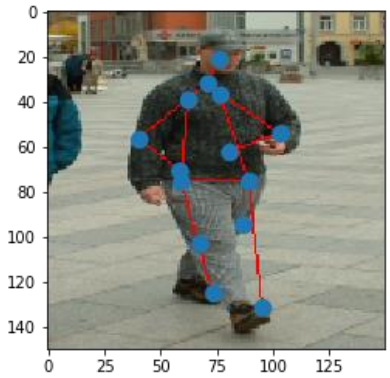
- Optimisation algorithm which combines the use of momentum and RMS prop on mini-batch gradient descent.
- Hyper-parameters used:
  - Learning rate: 0.1
  - Beta1: 0.9
  - Beta2: 0.999
  - Epsilon: 0

Comparison

Variant-1

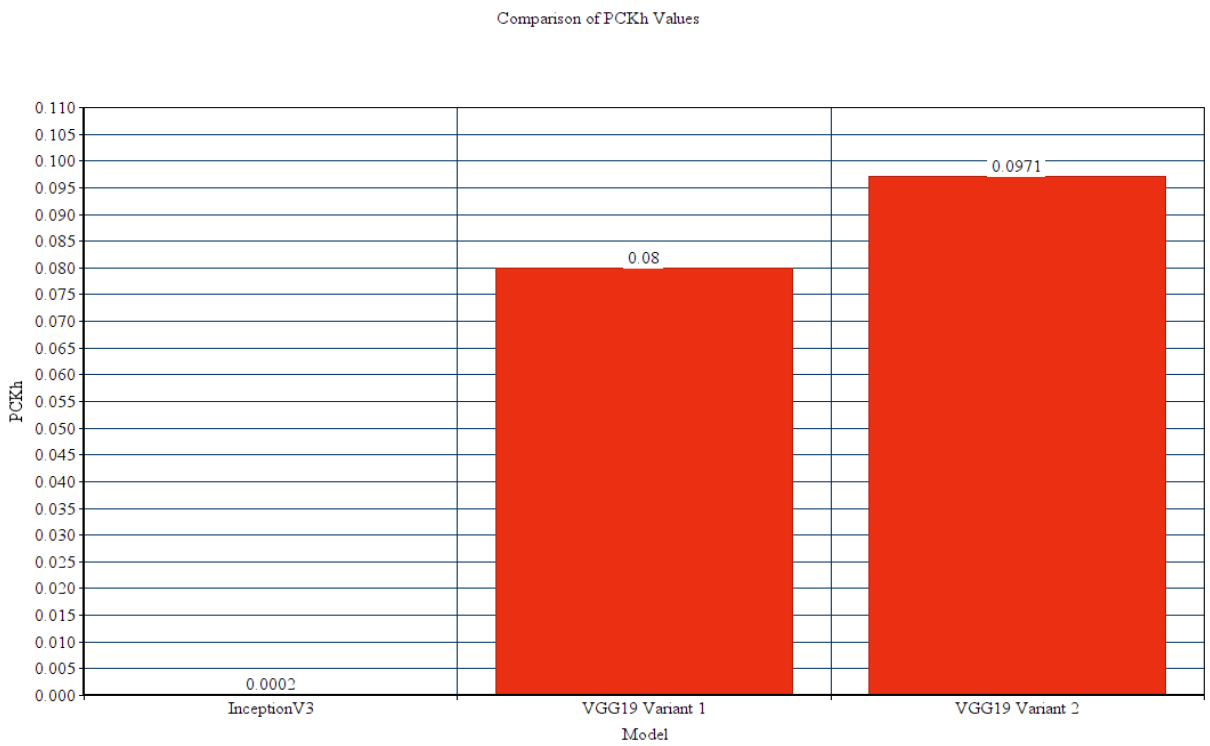


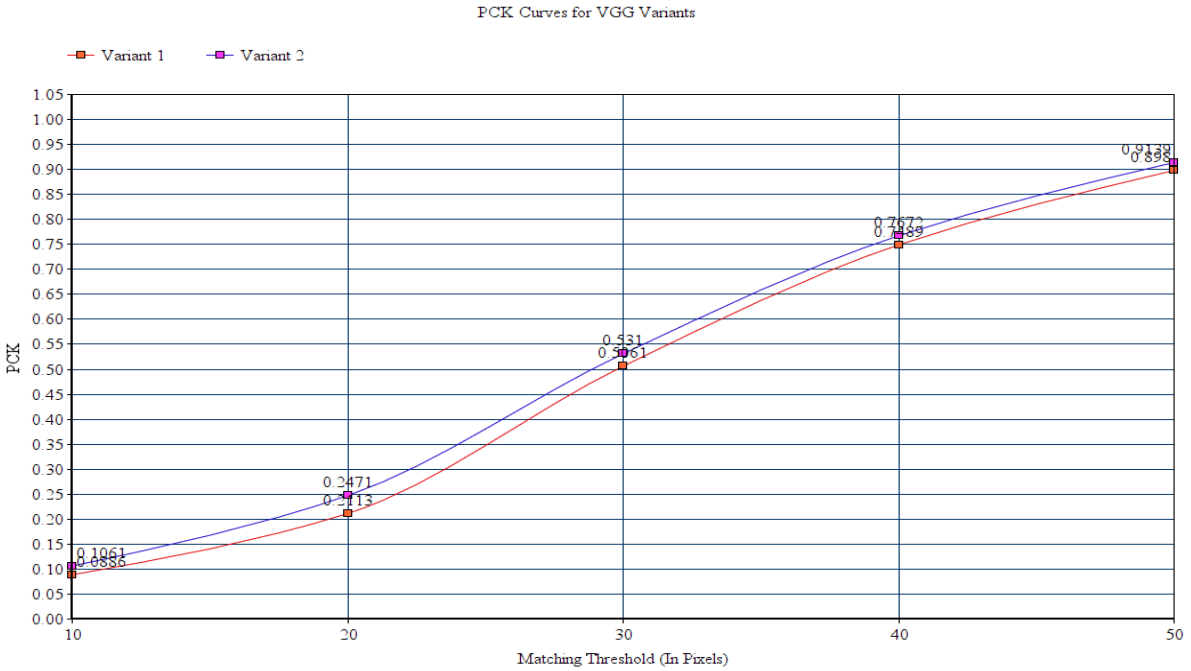
Variant-2

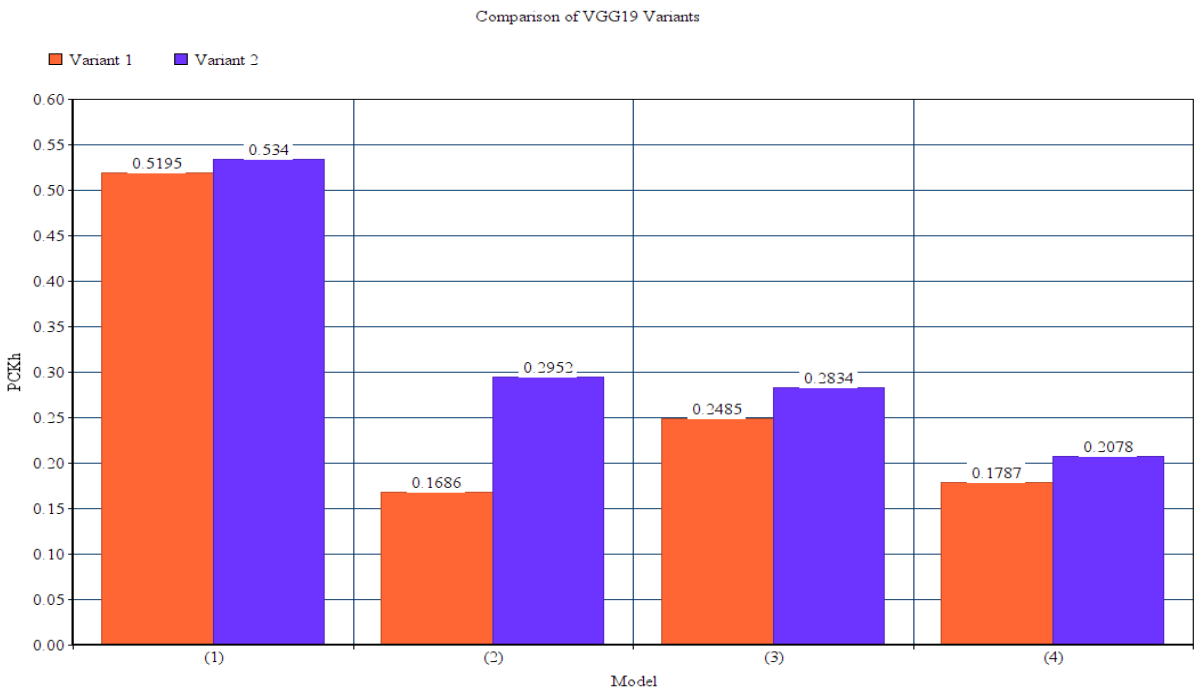


Matching Threshold	PCK Score for Variant 1	PCK Score for Variant 2
10	0.0886	0.1061
20	0.2113	0.2471
30	0.5061	0.5310
40	0.7489	0.7672
50	0.8980	0.9139

Graphical Analysis




PCK Curves for VGG Variants
Variant 1
Variant 2
PCK
Matching Threshold (In Pixels)
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
10
20
30
40
50
0.1061
0.0886
0.2471
0.2113
0.531
0.5061
0.7672
0.7489
0.9138
0.898



X-Axis Label	Meaning
(1)	Detection of head top and neck (face)
(2)	Detection of head top, neck and left and right shoulder joints
(3)	Detection of head top, neck and left and right hip joints
(4)	Detection of head top, neck, shoulder and hip joints (face and torso)

Conclusion

- YOLOv3 is an efficient object detection algorithm.
- After detecting humans in an image, we were able to produce decent pose estimations using deep learning methods.
- Inception Net Version 3 gave a very poor performance when compared to VGG19 Variants.
- Reducing the number of layers and increasing their size resulted in improvement of performance.