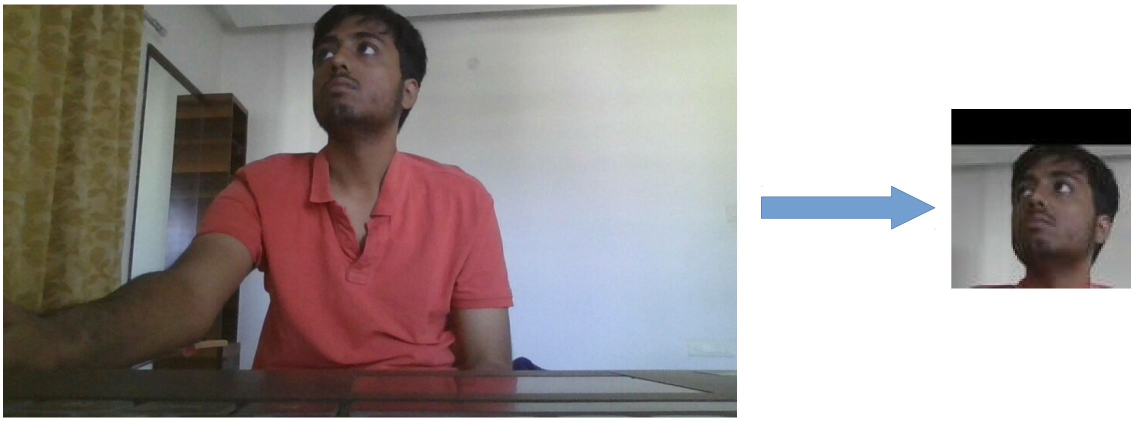
NOSE, EYES AND EARS: HEAD POSE ESTIMATION BY LOCATING FACIAL KEYPOINTS : REPORT

To Train the CNN, BIWI Kinect Database was used. It has over 15K images of 20 people recorded with a Kinect while turning their heads around freely. For each frame, depth and rgb images are provided, together with ground in the form of the 3D location of the head and its rotation angles. The Rotation angles are stored as a rotation matrix attached as metadata for every image in the dataset. The designed CNN takes a heatmaps of an Image as an input and gives yaw,pitch and roll of the head as output. So train the CNN, we need the following training data :

I. Heatmaps of 5 body parts (Nose, Left and Right Ear and Eyes) of every Image :

How this was achieved -

First we have to detect the faces and crop them to avoid noise. I used Dlib's frontal_face_detector for this purpose. Each cropped image was then resized to a size of 96x96 for the purpose of heatmap creation.

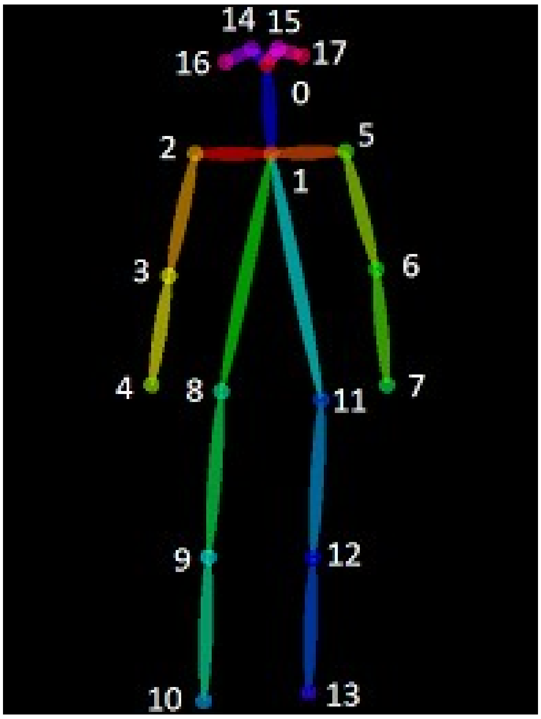


OpenPose was used to create the heatmaps. The following command was used to detect keypoints using COCO's 18 body part model.

```
./build/examples/openpose/openpose.bin --video location/of/image --heatmaps_add_parts --heatmaps_add_bkg --heatmaps_add_PAFs --model_pose COCO --display 0 --render_pose 0 --net_resolution 96x96 --write_heatmaps output_heatmaps_folder/
```

The output was an image of size 96x5472 (18 keypoints + 1 background + 2*19 (number of PAFs: 19 for x-direction and 19 for y-direction) = 18 + 1 + 38 = 57 heatmaps of size 96x96 are concatenated end-to-end and given as an output). The figure below is a sample heatmap.

From this, we have to extact the heatmaps of thes 5 keypoints – Nose, Left and Right Ear and Nose. The Keypoint ordering in the COCO model is as shown below. Clearly, we had to extract the 1st and 15-18th heatmaps from it.

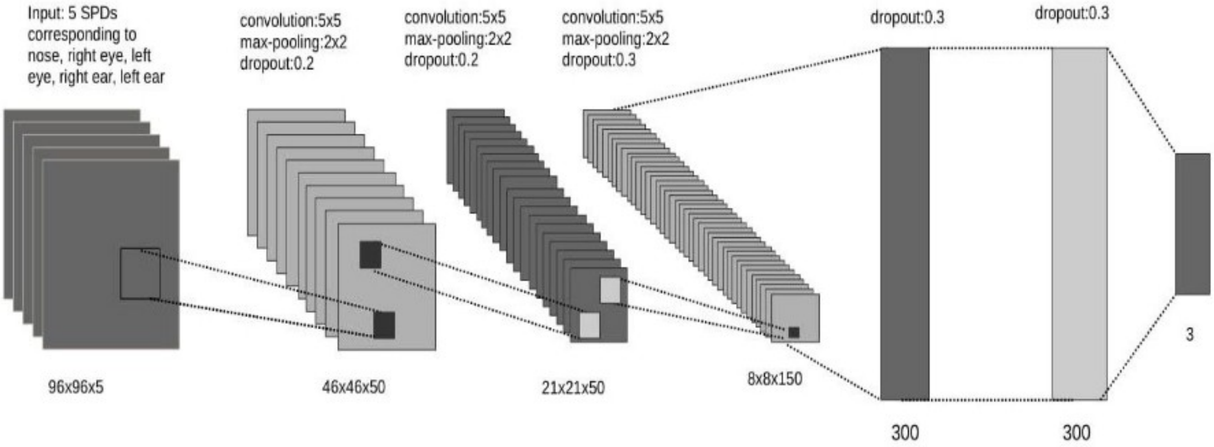


Now we have all the keypoint heatmaps of every image.

II. Yaw, Pitch, Roll of Training Images :

Extract Yaw, Pitch and Roll of every image using the given rotation matrix and convert it to degrees.

CNN Structure :


Input: 5 SPDs corresponding to nose, right eye, left eye, right ear, left ear
convolution:5x5 max-pooling:2x2 dropout:0.2
convolution:5x5 max-pooling:2x2 dropout:0.2
convolution:5x5 max-pooling:2x2 dropout:0.3
dropout:0.3
dropout:0.3
96x96x5
46x46x50
21x21x50
8x8x150
300
300
3

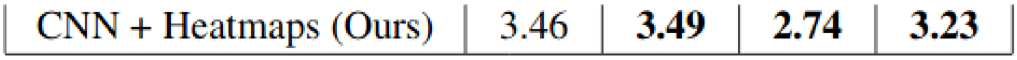
Training conditions :

8-fold cross-validation (21 randomly selected videos for training and the remaining 3 videos for test such that no person appears both in training and test sets). Training is run for 1020 epochs with Adam optimizer and set learning rate of 0.00001. We set the batch size to 128. All the experiments are run on a single Nvidia GTX 1660Ti GPU

Results:

The Errors on average were as follows:

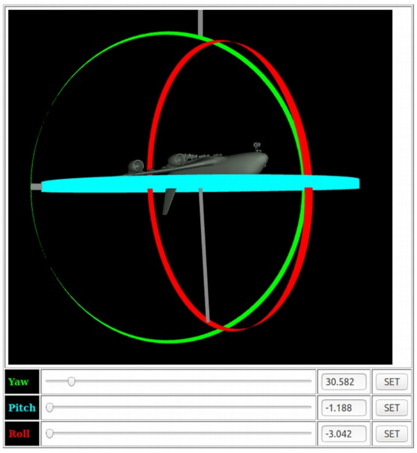
MAE: Yaw 3.00116, Pitch 3.72015, Roll 2.57272, Avg 3.09801



CNN + Heatmaps (Ours)	3.46	**3.49**	**2.74**	**3.23**

Figure: MAE in research paper for reference.

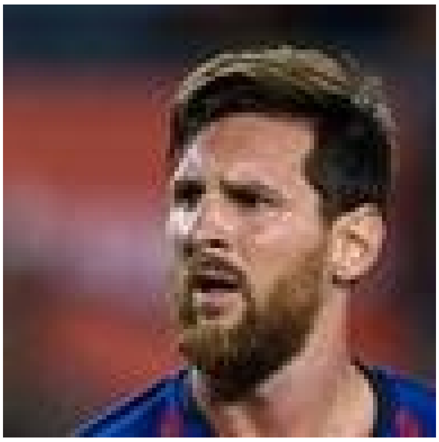
These were the headpose-estimation on a sample of images:

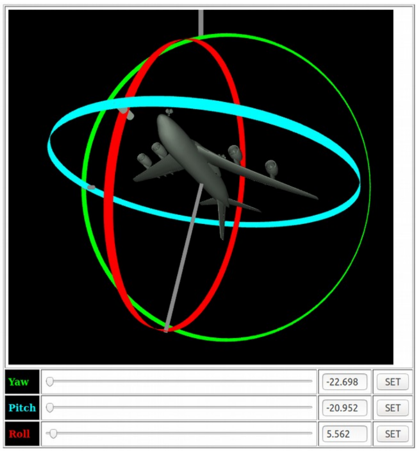


Yaw 30.582 SET

Pitch -1.188 SET

Roll -3.042 SET

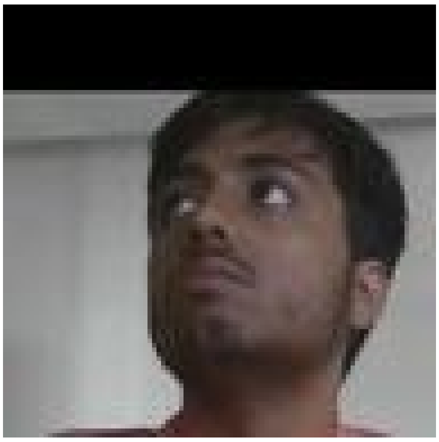


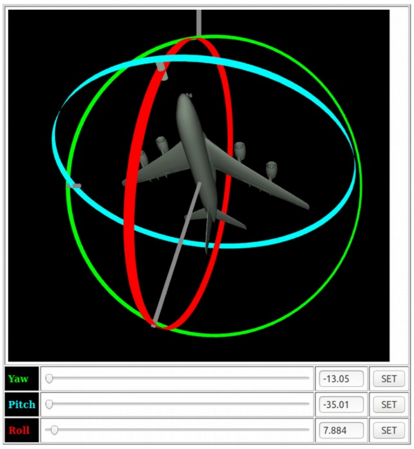


Yaw -22.698 SET

Pitch -20.952 SET

Roll 5.562 SET

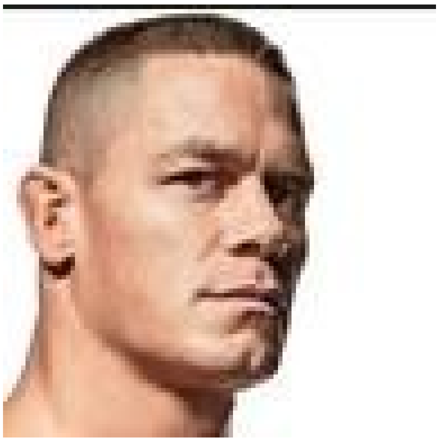


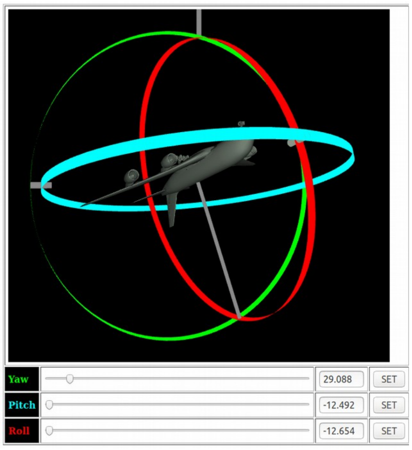


Yaw -13.05 SET

Pitch -35.01 SET

Roll 7.884 SET





Yaw 29.088 SET

Pitch -12.492 SET

Roll -12.654 SET