CS 6476 Project 5

Safin Salih
ssalih6@gatech.edu
902111076

Part 1: (8 points)

Briefly describe your understanding about the pipeline of mediapipe's objectron detection. Describe the stages and there required input/outputs

<text answer here>

Mediapipe objectron detects objects with 3D bounding boxes in a 2D images and estimates there poses using Neural Network models. This CNN tries to estimate object's size , position and orientation in the world.

The states consist of two-stage pipeline and a single-stage pipeline where single stage pipeline is good at detecting multiple objects and two stage pipeline is good at single dominant object. The first stage uses a object detector to find a 2D crop of that particular object. While the second stage takes the image crop and estimate the 3D bounding box.

Part 1: (4 points)

Is it possible to recover a single 3D point from a 2D point of a monocular image (which means a single image taken by a single camera)?

<text answer here>

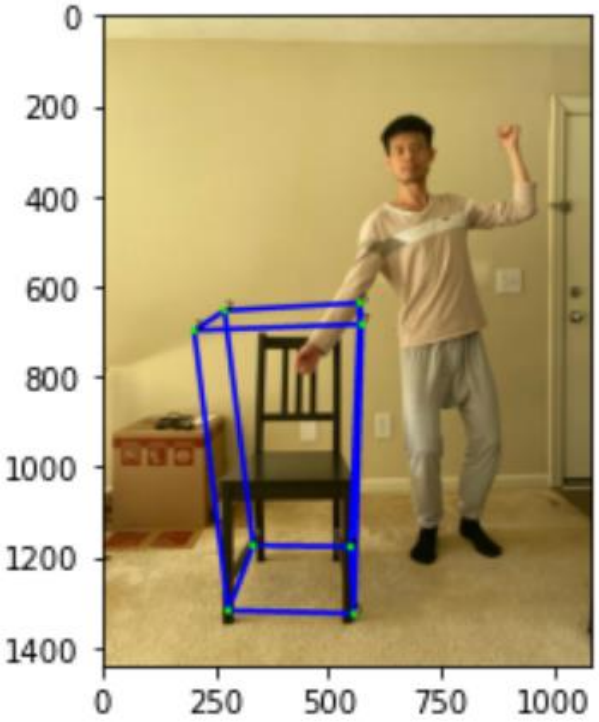
No, We can find the coordinates of a point given a projection of 3D points in two or more images, but not with a single image from a single camera.

Part 1: (4 points)

Why is it possible to estimate a 3D object from a monocular image (like mediapipe's objectron)? What other assumptions or data is needed to accomplish this.
<text answer here>

From the annotated data the CNN tries to initialize the object’s pose.
To predicts the object’s shape the signal depends on what ground truth annotation is available, e.g. segmentation. The assumption is that the annotated data is good representation to the task at hand.

Part 1: (4 points)



Copy and paste the code you fill in “detect_3d_box()” in “my_objectron.py()”
Note: Only paste the code you fill. Do not add the whole function

```
hm, displacements = inference(img=image,model_path=model_path)
```

Part 2: (5 points)

After you did camera calibration, you get a more accurate K, the intrinsic matrix of the camera, can you describe what is the meaning of the five non-zero parameter in K?

<text answer here>

$$K = \begin{bmatrix} \alpha_x & s & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{bmatrix}$$

αy and αx are the focal length of the camera, s is the skew factor, x0 and y0 are principal points expressed in pixels.

Part 2: (5 points)

In the K (intrinsic matrix), there is one value representing fx and another one representing fy, what is the unit of those two values? Why? In practice, when fx is not equal to fy, what does this mean in physical?

<text answer here>

Fx and Fy are focal length of the camera expressed in units of horizontal and vertical pixels where the values would differ if pixels are not perfect square.

Part 2: (10 points)

You also performed the transformation from camera to world by using the equations below. When we set the camera coordinate to world coordinate, what does ctw represent? Using the equation below, can we describe why the P matrix can project 3D points in world coordinate to 2D points on image plane? (Hint: the P matrix achieves two coordinate transform)

$$\mathbf{P} = \mathbf{K}\,{}^{w}\mathbf{R}_{c}^{\top}[\mathbf{I} \mid -{}^{w}\mathbf{t}_{c}]$$

$$z\begin{bmatrix} x_p \\ y_p \\ 1 \end{bmatrix} = \mathbf{P}\begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix}$$

<text answer here>

ctw is the camera center, for translation of the camera in the world coordinate frame. The camera matrix P is a projective mapping from 3D to pixel coordinate which is decomposed into intrinsic parameters and extrinsic parameters. wRtc is rotation matrix, [I | - wtc| is translation, and together they convert points from world to camera coordinate system.

Part 3: (3 points)

Please describe an application situation for pose estimation and explain why it is useful there.

An AI-powered workout trainer would use it’s camera to estimate human pose to recognize if a person completed a workout or properly following the workout instruction/routine.

Part 3: (3 points)

If you are going to do a pose detection project, what kind of pose do you want to detect and explain why these pose are important for you.

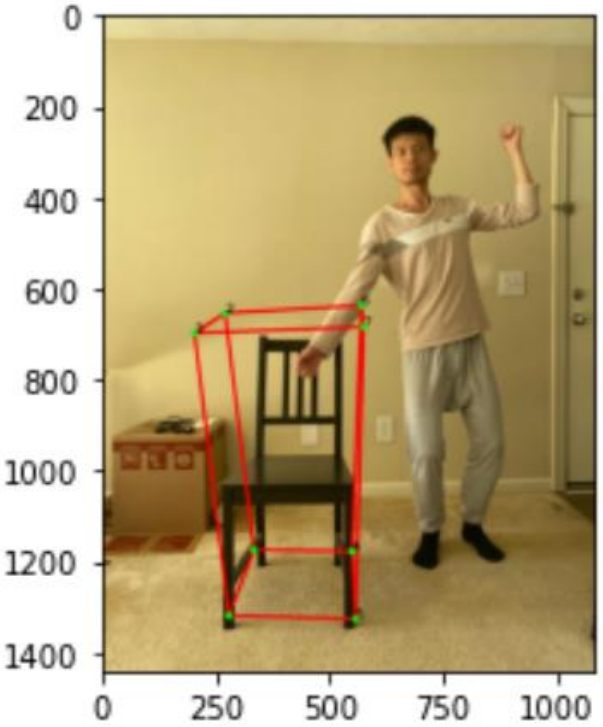
If my pose detection project is to create a chatbot robot that makes small dialogue with people . The pose that I want to are the features of the human faces, like eyebrows, noses ,mouth, lips etc . These could be used as signals/information that the chatbot could use to inform what to say next.

Part 3: (6 points)

What are the two main steps associated with pose detection used in mediapipe (Hints, read the blog post of mediapipe's pose detection)?
Answer:

Using a detector called the BlazePose Detector, which locates the pose region-of-interest within the frame, And the tracker, BlazePose Tracker that predicts the pose landmarks within the region-of-interest using the ROI-cropped frame as input.

Part 3: (4 points)



Copy and paste the code you fill in “hand_pose_img()” in “pose_estimate.py”
Note: Only paste the code you fill. Do not add the whole function

results = pose.process(image)

Part 5: (4 points)

Given the 3D coordinates of eight vertices of a box in space, and one 3D point, describe how do you detect whether this point is inside or outside the box?

Answer:
The 3D coordinates describe the world frame around the bounding box, so we just check if that 3D point is within that bounding box or not.

Part 6: (10 points)

Insert your picture before interacting with the object and other picture after the interaction happens (the bounding box changes color)

Extra Credit: Interaction Video

<Tell us where to access your final video of part 1 and 2, Discuss what you found out. >

Extra Credit: Interaction Video

<Were your results shaky? If so, why/what did you have to do to fix it? >

Extra Credit: Interaction Video

< What kind of factors determined how accurate the intersection detection was?>