Xixuan Liu

Personal Information

Status:	**Undergraduate Student**
Program:	**Computer Science**
School:	**New York University Abu Dhabi**
Website:	**https://www.linkedin.com/in/xixuan-julie-liu-814688148/**
RA Period:	**From 2017-02 to 2018-05**

Biography

I am an undergraduate student at New York University Abu Dhabi and a research assistant in NYU Multimedia and Visual Computing Lab, advised by Professor Yi Fang. I am broadly interested in 3D Computer Vision, Pattern Recognition and Deep Learning.

Research Project: Virtual Touch: Computer Vision Augmented Touch-Free Scene Exploration for the Visually Impaired

1 Description

The Blind or Visually Impaired (BVI) individuals use haptics much more frequently than the healthy-sighted in their everyday lives to locate objects and acquire object details. This consequently puts them at higher risk of contracting the virus through close contact during a pandemic crisis (e.g. COVID-19). Traditional canes only give the BVIs limited perceptive range. Our project develops a wearable solution Virtual Touch to augment the BVI's perceptive power so they can per- ceive objects near and far in their surrounding environment in a touch-free manner and consequently carry out activities of daily living during pandemics more intuitively, safely, and independently. The Virtual Touch feature contains a camera with a novel point-based neural net- work TouchNet tailored for real-time blind-centered object detection, and a headphone telling the BVI the semantic labels. Through finger pointing, the BVI end user indicates where he or she is paying attention to relative to their egocentric coordinate system and we are able to use this foundation to build attention-driven spatial intelligence.

2 Method

In this project, we use our deep learning model to directly classify the object being pointed at. The goal is to have the model automatically focus on the object in the finger-pointed region. The network detecting the fingertip also leads to a region of natural attention, meaning the VI user's attention is mentally focused on the spatial direction around the fingertip. A neural network then extracts image features and detects objects only from the areas on feature maps that contain the finger tip location. The network has default anchor boxes of different aspect ratios at the specific locations corresponding to the finger tip location in several feature maps of different scales. For each default box, we predict both the shape offsets relative to the default box coordinates and the confidences for all object categories. Candidates exceeding a confidence threshold are taken as

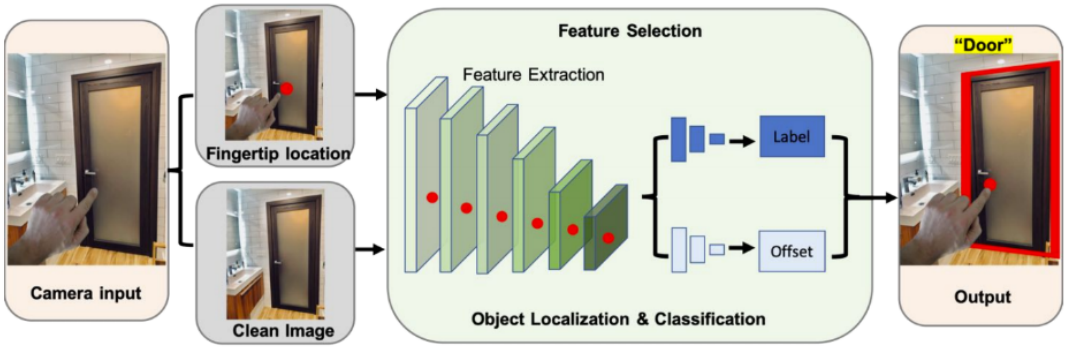

Figure 1: Architecture Overview of the Meta Deformation Network.

positions with objects. The predicted object with the highest confidence score is output as semantic notification to the user. Figure.1 shows the illustration of point-based detection.

3 Results

In this section, we carry out a set of experiments for point-based detection and assess the performance of our proposed TouchNet. Our deep-neural-network model TouchNet for AI-enabled exploration is trained and tested on PASCAL Visual Object Classes Challenge (VOC) 2007 and 2012. Each image contains a set of objects out of 20 different classes. The 20 classes are: Person - per-

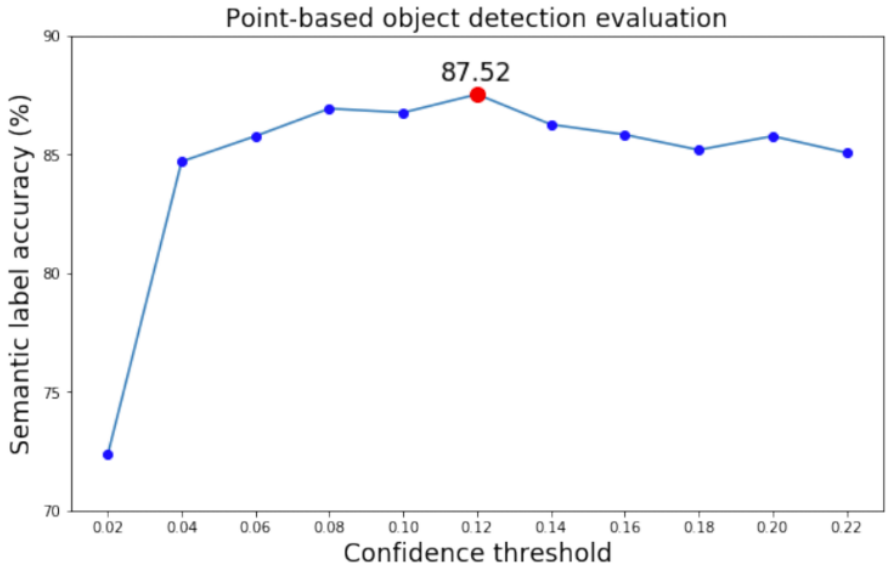

Figure 2: Comparison of performance of the TouchNet320 model on PASCAL VOC 2007, when using different confidence thresholds.

son; Animal - bird, cat, cow, dog, horse, sheep; Vehicle - aeroplane, bicycle, boat, bus, car, motorbike, train; Indoor - bottle, chair, dining table, potted plant, sofa, tv/monitor. Training of TouchNet used training set and validation set from both VOC 2007 and VOC 2012, which include 16,551 images in total. Figure.2 shows the semantic label accuracies for the TouchNet320 net- work with different confidence thresholds. The best accuracy for the Touch- Net320 network is 87.52% when the confidence threshold is 0.12, a performance value higher than the state of art mAPs on object detection. We also con- ducted analysis on false negative predictions, which could assist future improvement. False negative could arise from a finger tip location on the edge of the target object. In this case, the part of feature maps containing the finger tip location and the part of feature maps containing the center of object might not be the same. Predictions made based on the location of finger tip is thus likely to be associated with a low confidence along with the true label of the object. Or it could be a hard-to-see object, for example, an object only partially present in the frame, or an object that is very far and small. This issue could be overcome with more layers in the backbone network, a larger input resolution for the network, or even more data augmentation during training.

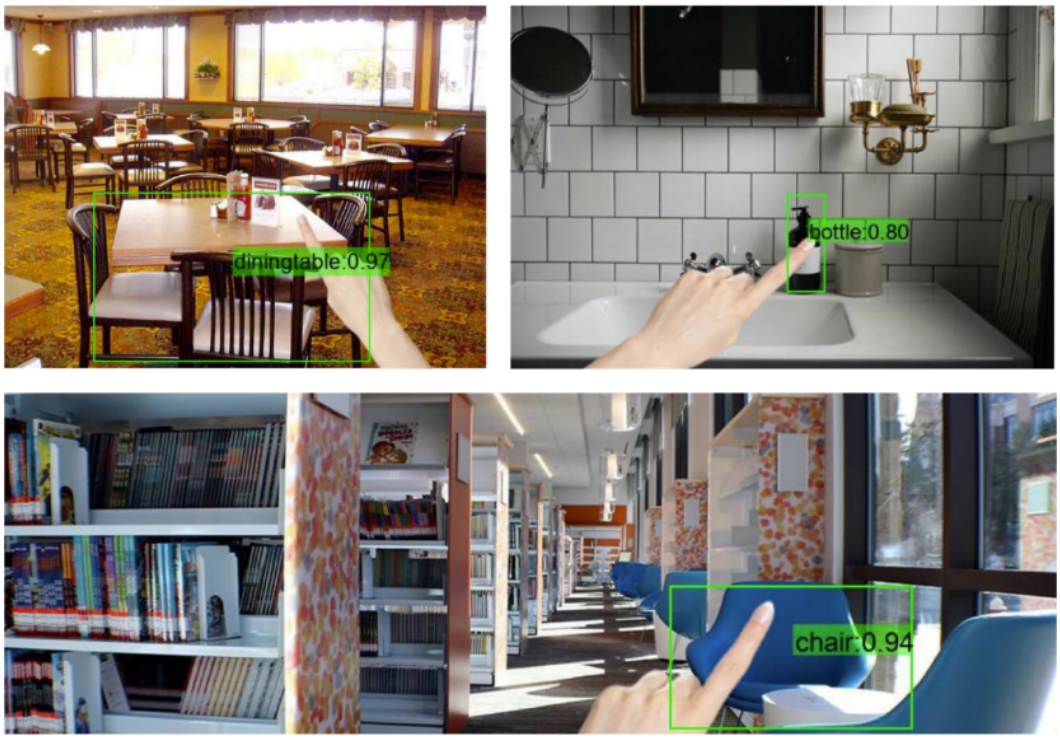

Figure 3: Examples of successful point-based detection in indoor scenes.