Efficient facial detection on embedded system with CNN for surveillance tasks

UMass Amherst COMPSCI682 Fall 2018
Rafal Bielech
rbielech@cs.umass.edu

BACKGROUND

Goal:
—> Integrate facial detection model for analysis of surveillance footage in real-time in an embedded environment

Motivation:
—> Traditional surveillance systems do not offer “smart” analysis of footage in real time.
—> Huge quantities of data are stored on disks far away despite not providing much use

Approach:
—> Given a video feed from a surveillance camera, a facial detection model can be run on the frames to determine if there are humans in the frames whom could be hailed as intruders

Advantage:
—> System is able to distinguish the interesting from non-interesting frames to minimize storage needs

PROBLEM STATEMENT

Design a facial detection model that can be an integrated with a real-time surveillance system that is running on a Raspberry Pi 3.

Input: Sequence of frames derived from splitting up a real-time camera feed where set of frames (S) is assumed to be in format S = {f1, f2, f3 … fn-1, fn } as time progresses forward

Output: Frames returned that have are found to have faces in them. Frames can be either stored or sent via email.

Performance goals:
FPS >= 2
Accuracy goals:
> 70%

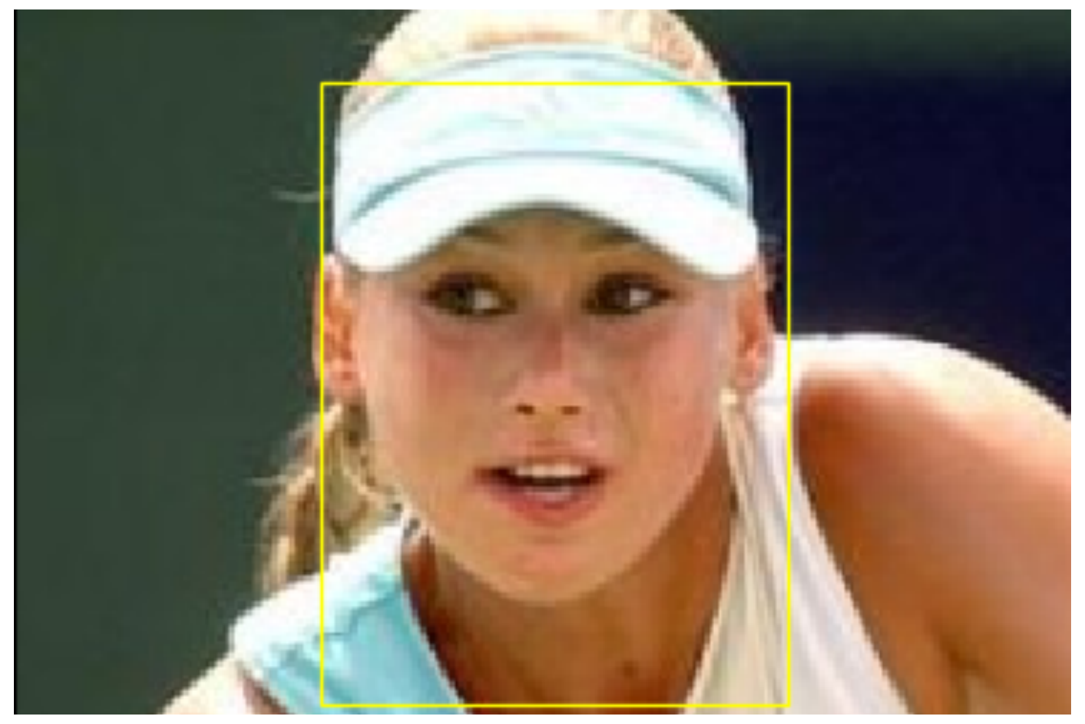


Figure 1; Facial detection model on an image from LFW dataset

DATASET

Dataset used : A Benchmark for Face Detection in Unconstrained Settings [FDDB]
Breakdown :
—> 12,000 Training Images
—> 1,000 Validation Images
—> 1,000 Testing images

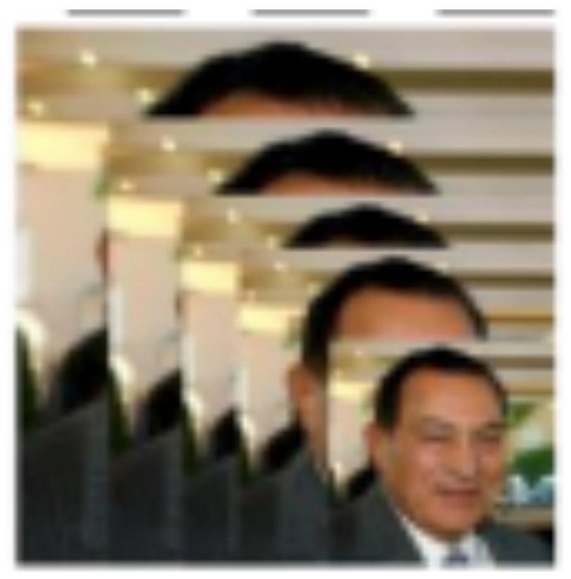
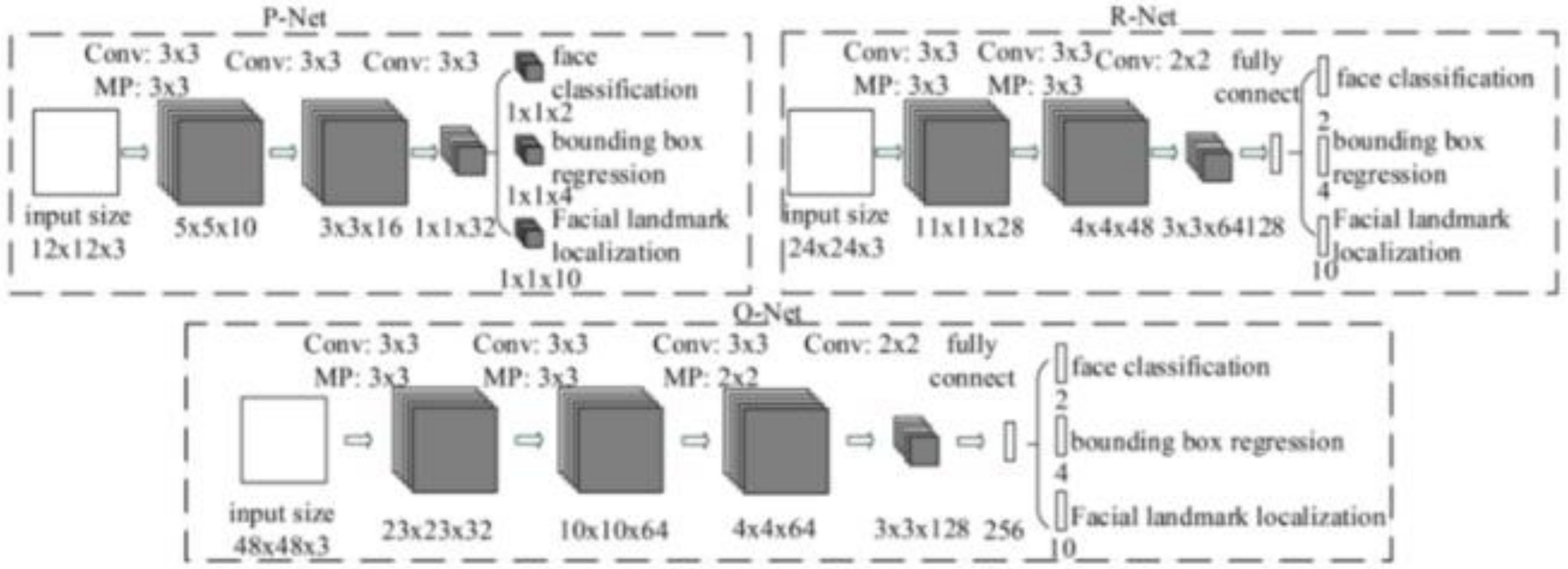


Figure 2; Images were fed to network in form of an image pyramid

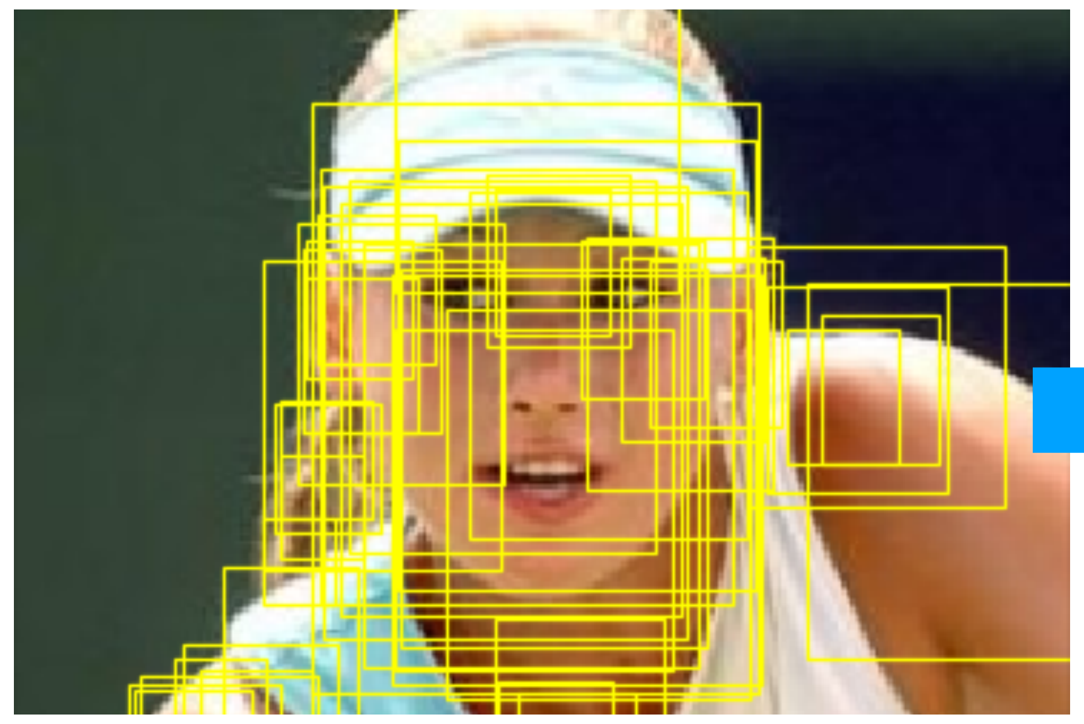
ARCHITECTURE



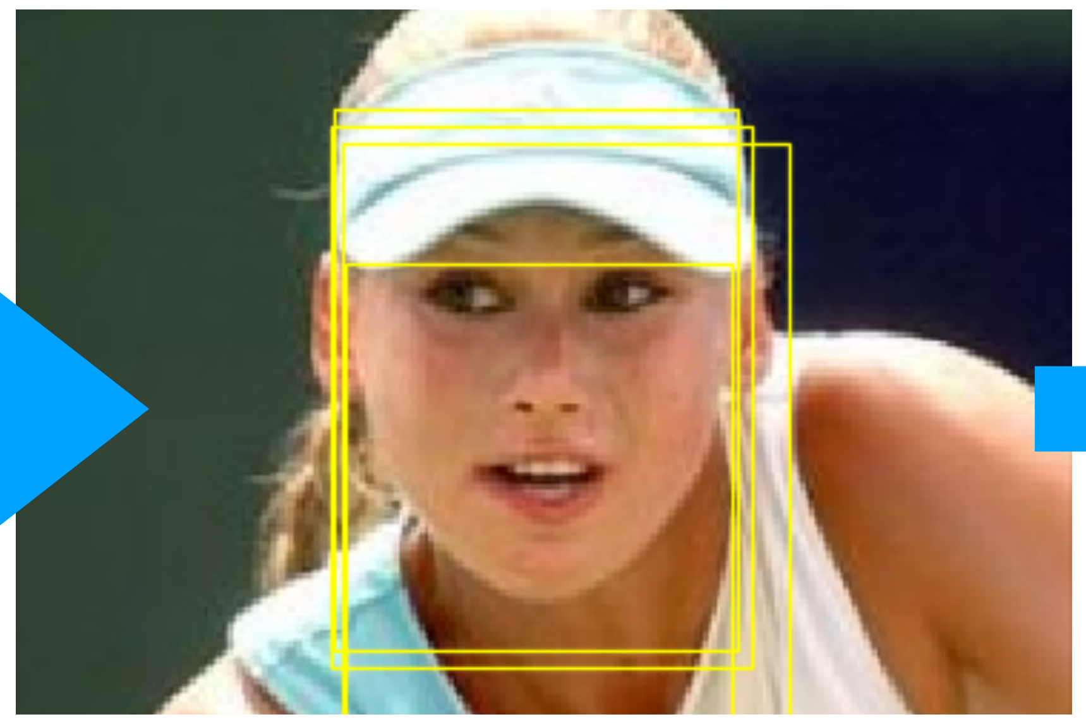
—> Proposal Network (P-NET) : Large number of bounding box proposals to the refinement network

—> Refinement Network (R-NET) : Refine the output from P-NET, drastically reduce number of bounding boxes

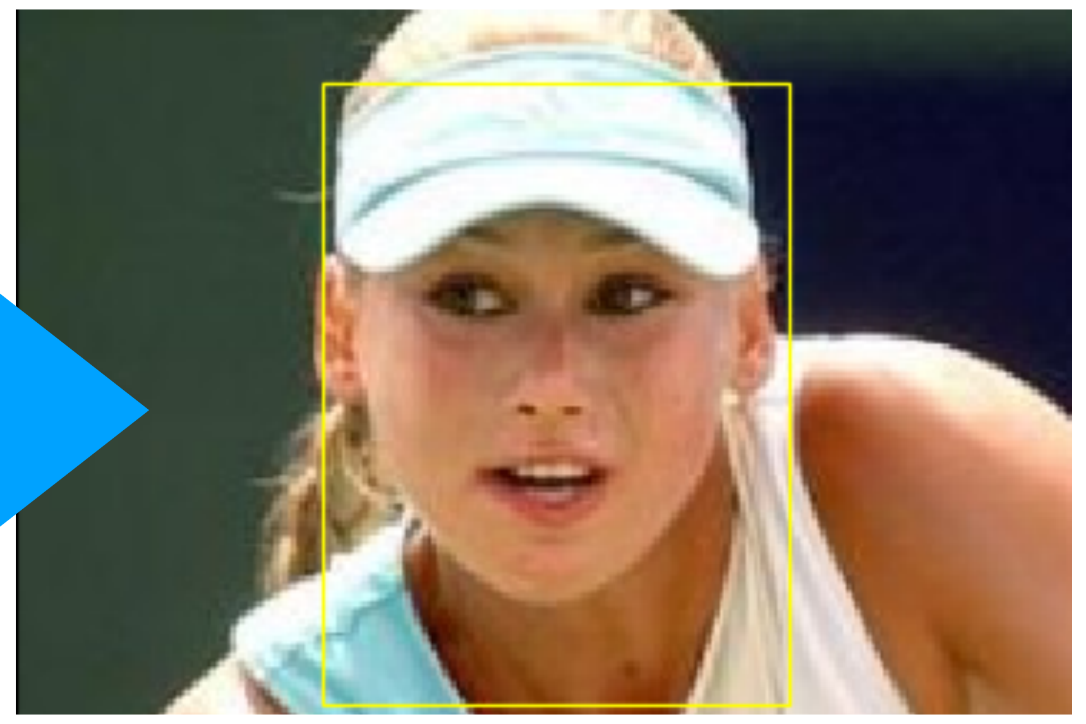
—> Output Network (O-NET) : Further refine the output from previous stage and return one bounding box per person with the highest confidence



P - NET RESULT ON AN IMAGE FROM LFW DATASET



R - NET RESULT ON AN IMAGE FROM LFW DATASET

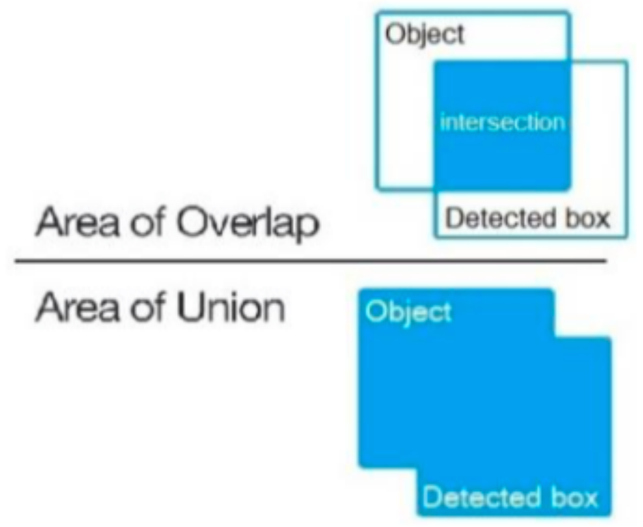


O- NET RESULT ON AN IMAGE FROM LFW DATASET

EVALUATION

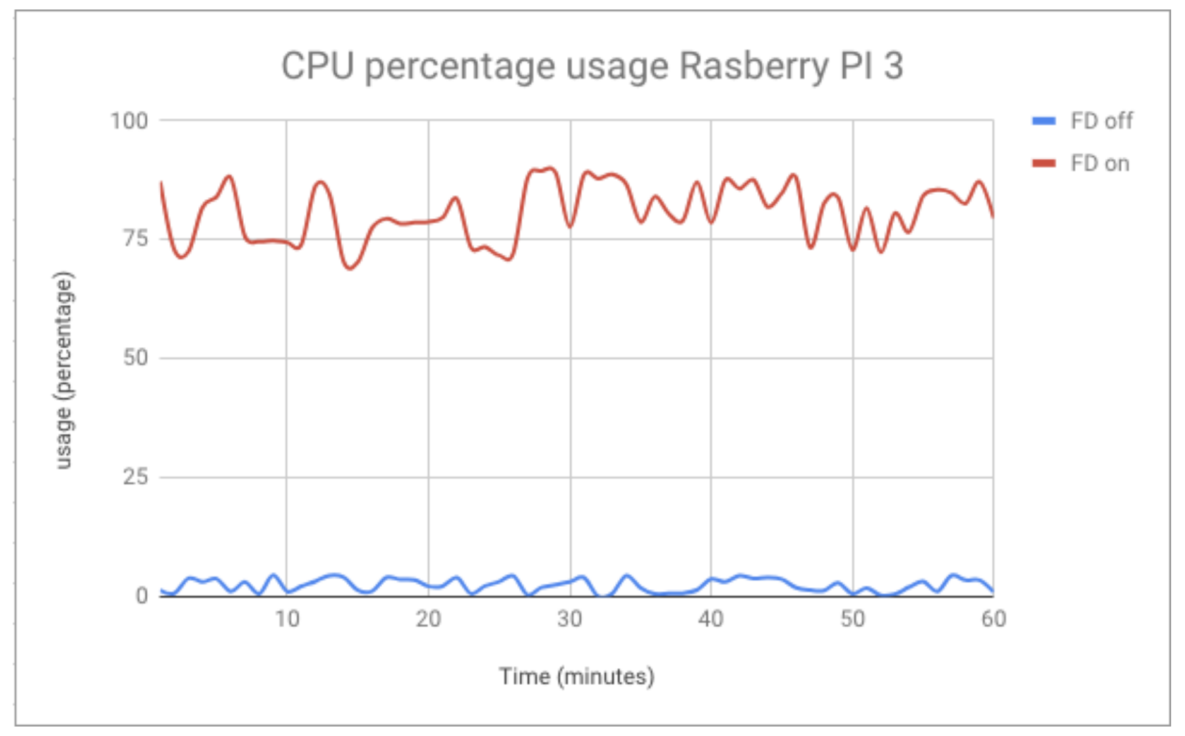
Statistics
Accuracy on finding correct number of faces = 84.7%

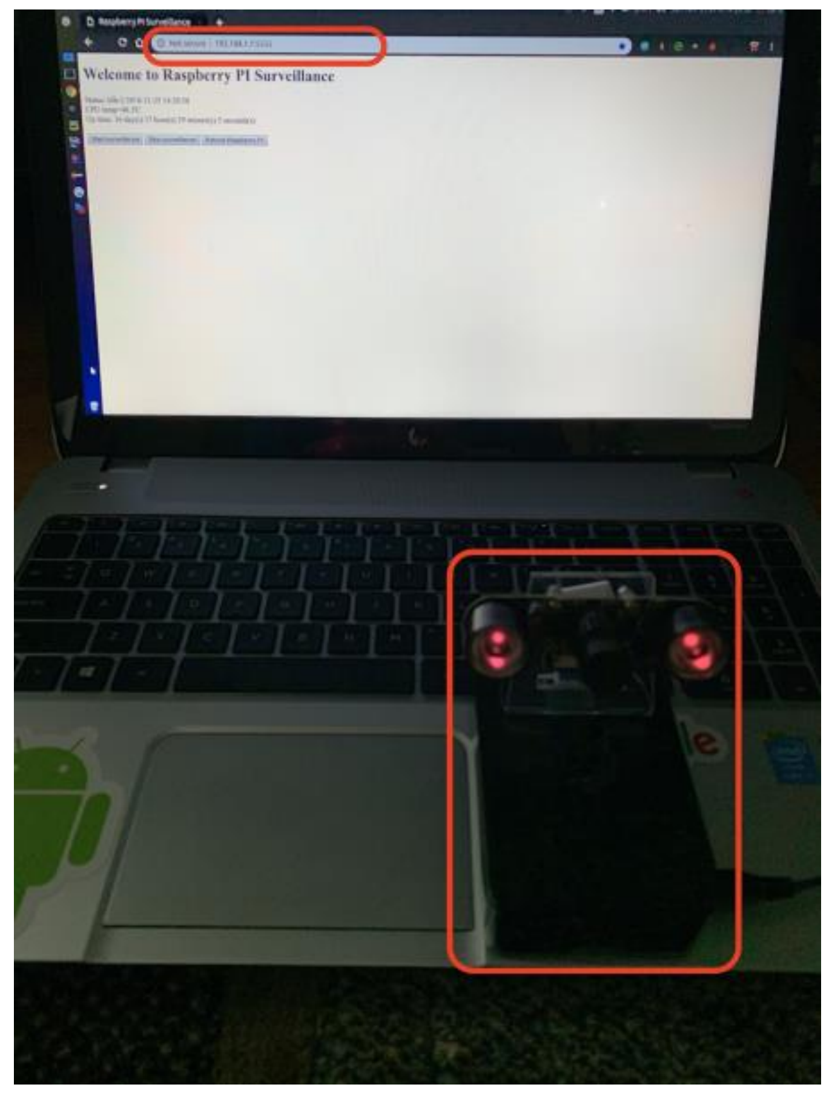
Accuracy IOU= 59.4%



FPS on laptop video feed: ~ 15 frames
FPS on Rpi video feed: ~ 2 frames

Average Proposal Network cost = .157 seconds
Average Refinement Network cost = .096 seconds
Average Output Network cost = 0.017 seconds
** FPS and Performance statistics were taken on small images





INTG. W/ SURVEILLANCE AND FUTURE WORK

—> Surveillance system is integrated with face detection (Shown on the left)

Improvements:
—> Retrain the model and/or model changes with additions to perform facial recognition

—> Object tracking between frames instead of recalculating the model between each frame

—> Better visualization

—> Better performance by simplifying/optimizing the model