UNIVERSITY OF MARYLAND

ACME Robotics Human Detector-Tracker

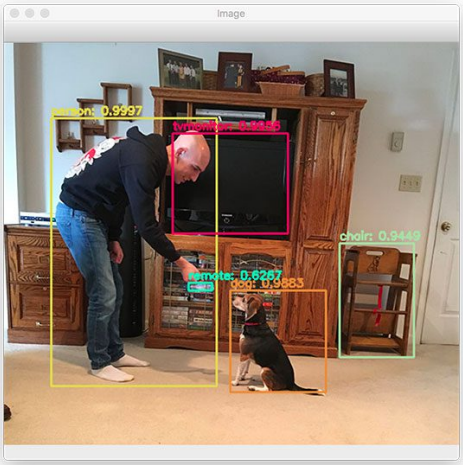
Divyansh Agrawal, Adithya Singh, Rishabh Singh

Abstract: **Human detection and tracking is a significant task in the field of computer vision. It means identifying the presence of a human in all regions of the image by comparing each region of the image to a template or pattern of people. Tracking involves identifying the human in each frame of the video and localizing it in the frame. In this proposal, we present a real-time robust, and accurate human detection and tracking system. In the end, we calculate the location of the human in the robot reference frame.**

1. INTRODUCTION

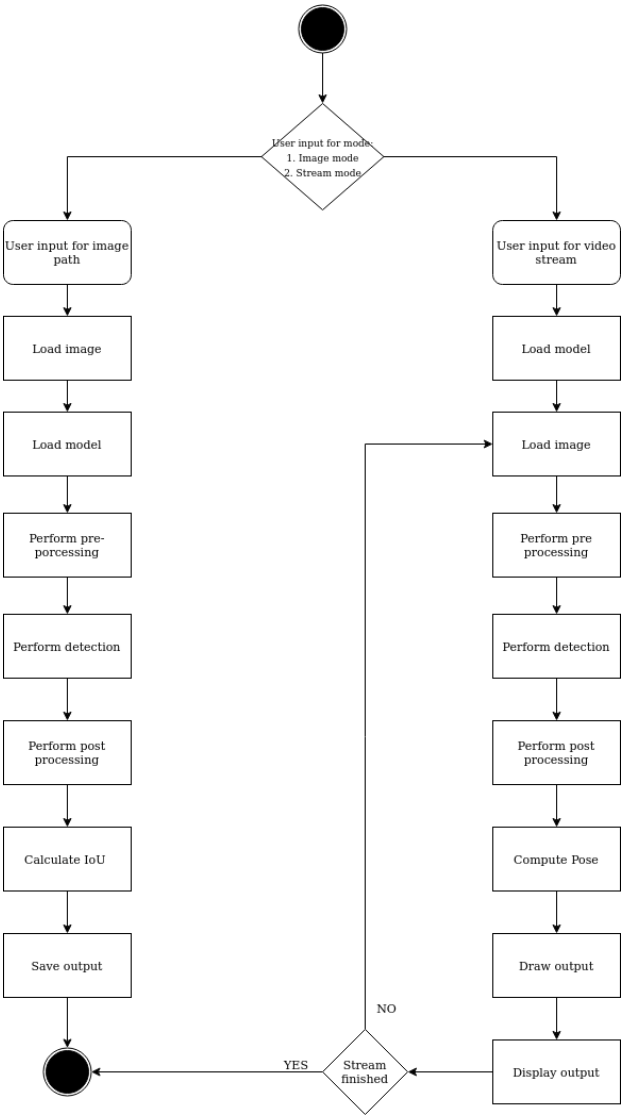
Be it an autonomous navigation system or a surveillance system, obstacle detection is an important aspect of the problem. Dynamic obstacles like humans or animals pose a risk to the system and themselves if they are not identified properly. We propose a human detection and tracking system using computer vision algorithms. The idea here is to generalize the detection pattern based on features. The model will be trained based on features, for example, change in the gradient when moving from forehead to eyes, or identifying a dark region (shadow) on both sides of the nose. The implementation will be done using object-oriented programming on the provided dataset and following the C++ guidelines. Frame transformations will be used to output the final location in the robot reference frame or world frame.

2. IMPLEMENTATION



When it comes to deep learning-based object detection, there are three primary object detectors you'll encounter

- R-CNN and their variants
- Single Shot Detector (SSDs)
- YOLO



Activity Diagram

UNIVERSITY OF MARYLAND

We will be using the YOLO i.e. "You Only Look Once" objects detectors because it is a single-stage detector and hence significantly faster which can achieve 155 FPS on a GPU and is capable of detecting over 9k objects. The code structure will be divided into different classes a Detector for detection purposes, a Tracker for tracking the coordinates of the bounding boxes, Utils for extra functionalities, and a Robot class for final processing and interaction with other classes.

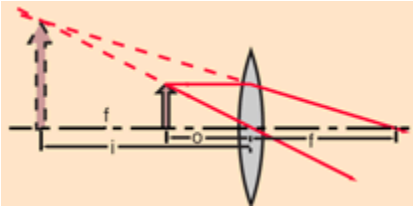
Detection Module:

This module is responsible for invoking the detection algorithm. Ideally, the detection algorithm is to be run on each input frame. However, this will inhibit the system from meeting its real-time requirements. Instead, the detection algorithms in our implementation are invoked every two seconds. The location of the human targets in the remaining time is determined by tracking the detected humans using the tracking algorithm.[4]

Tracking Module:

This module processes frames and detections received from the human detection module, and retains information about all the existing tracks. When a new frame is received, the already existing tracks are extended by locating the new bounding box locations for each track in this frame. If the frame is received accompanied by new detections, the new detections are compared to the already existing tracks. If a new detection significantly overlaps with one of the existing tracks, it is ignored. Otherwise, a new track is created for this new detection. A track is discontinued if the tracking algorithm fails to extend it in a newly coming frame.[4]

Pose Module:



We are assuming the depth of the object in the image to be related only to the focal length of the lens used, which is:

$$\text{ActualDepth} = \frac{(Actualheight * focallength)}{ImageHeight}$$

For evaluation, Intersection over Union (IoU) will be used which measures the overlap between the boundaries of the ground truth of annotation and the predicted boundary. IoU evaluates whether a prediction is "good enough" [2]. The closer the prediction is to 1, the closer to perfect it is. Figure 19 illustrates the graphical view of the equation below.

$$IoU = \frac{Area\ of\ Intersection/Overlap}{Area\ of\ Union}$$

In general, when $IoU \geq 0.5$, we consider the prediction correct. Knowing IoU, we calculate precision and recall. We also calculated the confidence score. The confidence score reflects how likely the box contains an object and how accurate the bounding box is.

Fig. The higher the IoU, the better the performance [3]

References:

1. Davis, J.W., Sharma, V., Tyagi, A., Keck, M. (2009). Human Detection and Tracking. In: Li, S.Z., Jain, A. (eds) Encyclopedia of Biometrics. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-73003-5_35

2. A. Abdulkader & C. Vlahija "Real-time vehicle and pedestrian detection, a data-driven recommendation focusing on safety as a perception to autonomous vehicles", Available at: http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1479957&dswid=-3676 [Assessed: 03/18/21]

3. Object Detection and Recognition Using YOLO: Detect and Recognize URL(s) in an Image Scene (2021)

4. M. Hussein, W. Abd-Almageed, Yang Ran and L. Davis, "Real-Time Human Detection, Tracking, and Verification in Uncontrolled Camera Motion Environments," Fourth IEEE International Conference on Computer Vision Systems (ICVS'06), 2006, pp. 41-41, doi: 10.1109/ICVS.2006.52.