Crowd Tracking Techniques for Evaluation of Adherence to Social Distancing Measures: Midterm Report

Computer Vision Final Project Proposal

Jacob Desman, Sandy Shi, Emily Chang, Andrew Cornelio

Github Link: https://github.com/endernac/CV-Final-Project

Progress Update:

Much of the initial work that we have done thus far has revolved around creating a baseline functional system. The first step in developing such a system would be finding a means of person detection. In our case, we looked into popular deep learning person detection systems, YOLO and DETR. These systems perform detection and localization of a number of categories of objects, including people. We have adapted the detectors so that they detect only humans and return the coordinates of their bounding boxes or ground plane positions in the image (Fig. **1**). Then, using the detected people, we then explored two main approaches to gauging social distancing: a generalized bounding box approach and a homography-based approach.

1. **Bounding box approach.** From the bounding boxes found by each detection system, we followed the procedure to determine interpersonal distances as described in Ghodgaonkar, et al (2020). Essentially, by comparing the area occupied by two bounding boxes, we are able to appropriately scale the euclidean distance between them. Given this and the assumption that all boxes average out to the average human height of 5.4 feet, we can get a rough estimate for the amount of distance between two individuals. Using a function to evaluate these distances, we generated videos in which violations of social distancing would be identified by blue bounding boxes, and all other safely-distancing individuals would be identified by green bounding boxes. These can be found in this Google Drive folder. However, though this method provides a functional baseline, we seek to improve on this method by incorporating information on the scene geometry through a homography-based approach..
2. **Homography-based approach.** We intend to use homographies to get a mapping from the image plane to the birds' eye view by using a set of four points inputted by the user. As an initial tool for evaluation, we are using the VIRAT dataset, which includes pre-estimated homographies. In advance of our future pipeline, we have prototyped an interactive system for user inputted homographies (see Appendix, Fig. **2-6**). After adjustments, the finalized homography is then used to map the detections from the deep learning person detection system to the birds' eye view and gauge distance. Here, we are currently making the assumptions that the user can determine an approximate scaling value to convert the birds' eye view units into feet or meters (see Fig. **6**, where the units do not necessarily correspond to feet).

Action Plan:

- **Camera geometry inference.** Looking forward, the largest roadblock involves figuring out a *method of accurately gauging distances between people in the birds' eye view.* Through multiple discussions, we have identified that this involves calculating the angle at which the camera is looking down onto the ground plane. From there, we would be able to use the average height of a person to determine units in the birds' eye view. Thus, the first plan of action will be working out the math for this problem and implementing it to refine the results we have from our baseline implementation. For this, the assumptions that we will be making are that:
    - People appear to stand up-right, or 90 degrees, relative to the ground plane.
    - The surveillance camera's horizontal orientation is parallel to the horizon.

  These assumptions will be used to simplify the relationships between the image space and real-world coordinates to help with the above calculations. Once the camera angle is inferred, we will be able to use simple trigonometric relationships to determine the scale in the homography space.

- **Tracking.** Concurrently, we are looking to get better bounding box saliency, as there have been some cases in which the deep learning person detection systems have identified overlapping boxes for the same object. To better help discern between overlapping boxes that are tied to the same object and overlapping boxes that represent an object directly in front of another, we hope to further tune the model parameters and/or implement helper functions to improve the person detections. One way we seek to do this is through non-maximal suppression based on the objectiveness score of overlapping bounding boxes.
- **App development.** As a bonus stretch goal, we hope to expand upon the existing interactive point selection interface and integrate the functionalities into an app – however, this should only be done once we are satisfied with the performance of the entire system.

Appendix

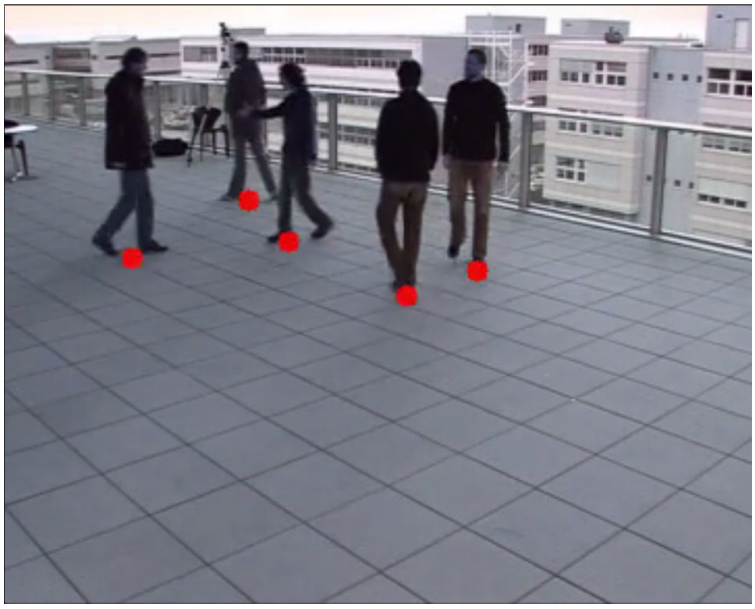

Figure 1. We use detectors YOLO and DETR to detect and localize people as they move through an image. We take the center of the bottom edge of the image as a person's position on the ground plane.

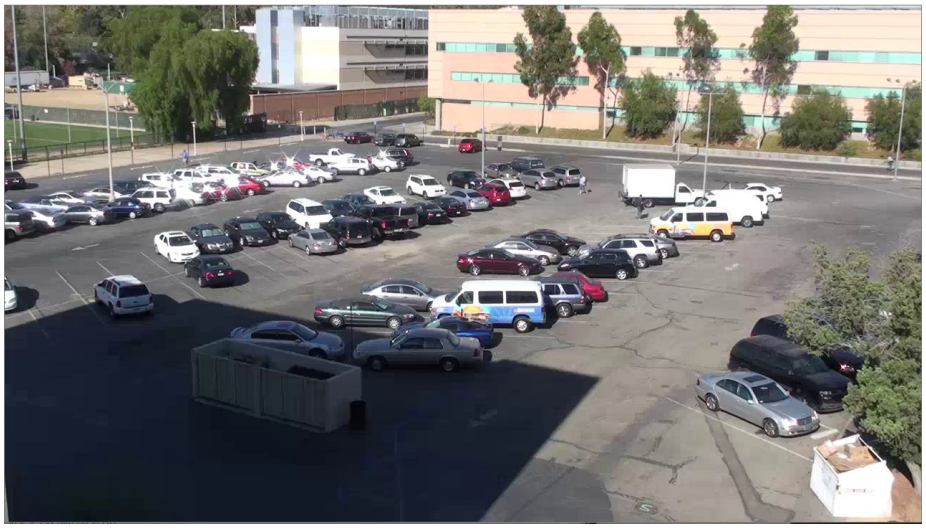

Figure 2. A scene taken from a single image frame before homography fitting.

Figure 3. An initial user point placement. As can be seen in this figure, the initial placement of the square is often inaccurate and requires some adjustments. In this case, the region is actually rectangular as well.

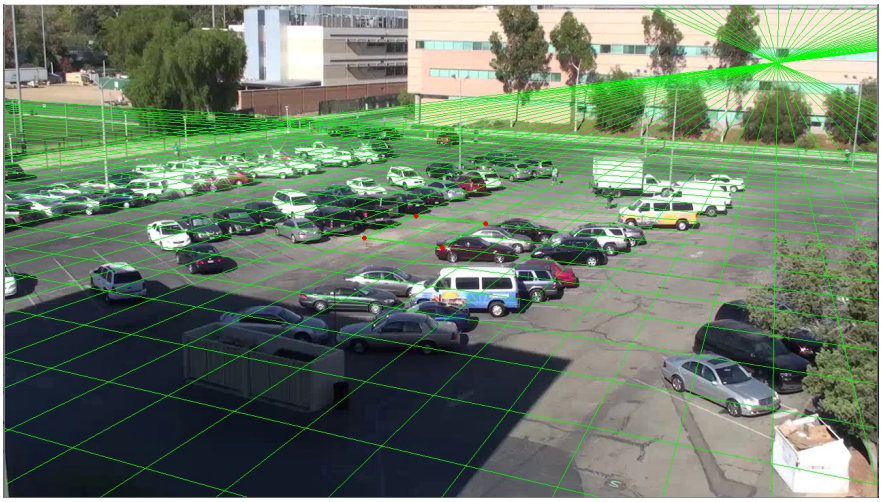

Figure 4. After user adjustment, this represents a significantly better homography for fitting the ground plane to a birds' eye view.

Figure 5. One the homography is found, for this test case, three points were selected in one test scenario. There is a blue dot over one corner of the building shadow, an orange dot over the other corner of its shadow, and a green dot close to the street light in the back right corner of the image.

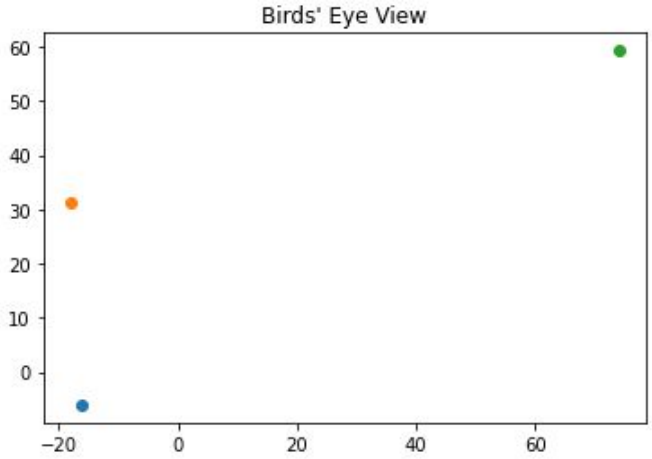

Figure 6. The homography can then be used to determine the relative locations of these points on the ground plane. Note the differing scales of the axes, showing the orange and blue markers much closer together than the green marker. Additionally, based on the homography gridlines (Fig. 4), the orange and blue markers appear that they should be on relatively the same axis, while the green should be offset and further away. These relationships have successfully been captured in this representation.