Estimating distance between pedestrians from street view images using geometric properties

Christopher Salazar

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science in Industrial Engineering

University of Washington

2021

Committee:
Dr. Youngjun Choe
Dr. Linda Ng Boyle
Dr. Joseph Wartman

Program Authorized to Offer Degree:
Industrial and Systems Engineering

University of Washington

Abstract

Estimating distance between pedestrians from street view images using geometric properties

Christopher Salazar

Chair of the Supervisory Committee:
Assistant Professor Dr. Youngjun Choe
Industrial and Systems Engineering

A National Science Foundation (NSF)-supported Rapid Response Research (RAPID) project was granted in the early months of the COVID-19 lockdown for the purposes of tracking Seattle, Washington through the progression of the pandemic. Street view data, reminiscent of Google street view, is generated via a RAPID vehicle with a mounted 360-degree camera as it tours a predetermined route through several locations in Seattle. One major aspect that many locations enacted during the lockdown was the recommendation for individuals to practice social distancing. While there have been studies that look at metrics to quantify social distance compliance, these findings are indirectly measured, focus on indoor environments, and are generally inconclusive to characterize social distance adherence on an individual basis. This unresolved research question, along with the data generated by the RAPID vehicle, presents an opportunity to develop a routine that directly estimates the social distance between pedestrians in outdoor settings. Some approaches to estimating distance between objects from images rely on specialized camera equipment, computationally complex frameworks like *Structure from Motion*, or deep learning networks that require three-dimensional training metadata. These methods either exceed equipment limitations of the RAPID vehicle or are not scalable to the large image dataset being generated. Therefore, I propose an approach that processes 360-degree video data into distortion-free images, utilizes a state-of-the-art pedestrian detection algorithm, and develops a geometric social distance

estimation method. A physical experiment is designed to generate ground-truth data in order to optimize the social distance estimation method and evaluate its performance. This yields a test root mean square error (RMSE) of 1.13 ft, which represents the error when estimating distances between pedestrians. A 95% confidence interval, constructed via bootstrapping, for the true RMSE is determined to be $(1.07, 1.41)$. Furthermore, because of the computational efficiency of the proposed method for estimating distances using geometric properties, the number of required arithmetic operations scale in $O(m^2)$, which is the lower bound and thus optimal for the number of pedestrians m found in any given image. These results summarize an estimation that can be applied to research regarding the estimation of distance between pedestrians when the reported tolerances are acceptable. Such research could include tracking social distance compliance over time, determining size of social groups, or even non-COVID related work like evaluating how bike friendly a city is.

TABLE OF CONTENTS

Page

List of Figures ii

Chapter 1: Introduction 1

Chapter 2: Background 3
2.1 Specialized Cameras and Configurations 3
2.2 Transformation from two dimensions to three dimensions 4
2.3 Deep Learning Frameworks 5

Chapter 3: Methods 7
3.1 Image Processing and Pedestrian Detection 7
3.2 Social Distance Estimation 9
3.3 Ground Truth Experiment 13

Chapter 4: Results 19
4.1 Image Processing Results 19
4.2 Grid Search and Visual Results 19

Chapter 5: Discussion 25
5.1 Applicability 25
5.2 Limitations 27
5.3 Future Work 29

Chapter 6: Conclusion 32

Bibliography 34

LIST OF FIGURES

Figure Number		Page
3.1	360 degree sample image from the RAPID vehicle between Paul G. Allen Center (left) and Bill & Melinda Gates Center (right) at 3800 E Stevens Way NE, Seattle, WA 98195, USA (Lat: 47.653535, Long: -122.305175).	8
3.2	Pedestrian detection with bounding boxes and coordinate axes.	9
3.3	Bird's-eye view of pedestrian detection.	10
3.4	Depth location of pedestrians	11
3.5	Triangular representations of pedestrians	14
3.6	Positioning of test pedestrians for ground truth experiment	16
3.7	Position markings for actors A and B	17
4.1	Corrected images via gnomonic projection.	20
4.2	Bounding box of pedestrians in images.	21
4.3	3D countour plot of Training RMSE	22
4.4	Drawn Social Distance Line Results.	23
4.5	Sample output of Pedestron/distance estimation on unseen images (green lines signify social distances more than six feet and red lines less than six feet).	24
5.1	Field of view for RAPID vehicle of test pedestrians.	26
5.2	Distance estimation with ground truth distance of 6 ft.	27
5.3	Distance estimate of child and their guardian.	28
5.4	Person detected on balcony.	30

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation (NSF) under Grant CMMI-2031119. Special thanks to the principle investigators Dr. Joseph Wartman, Dr. Scott Miles, Dr. Nicole Errett and Dr. Youngjun Choe. Additional thanks Jacqueline Peltier and the rest of the RAPID support crew for their ground support of the street view survey.

Chapter 1

INTRODUCTION

Upon the origination of the COVID-19 pandemic, which was reported as early as November 2019 [18], several countries have faced the challenge of suppressing its spread. The rapid initial spread of the virus along with the growth of the contagion has forced governments to implement safety protocols. While these measures vary depending on the country, these protocols typically recommend people to wear masks and socially distance themselves from others. Though social distancing can simply mean staying at home and limited interaction with others, at a minimum it requires a six-feet physical distance from other people [1]. Some governments have executed lockdowns by closing or limiting capacity of businesses in an effort to enforce social distancing to a larger scale. However, policing the compliance of six-feet separation at the individual level is not practical and requires a certain amount of conformity from its community. This conformity has come into question in a study by Marchiori [16], where they found an alarming amount social distance violations when an unmasked test subject walked in a public environment. This study, though localized to the Venice metropolitan area in Italy, exemplifies a general lack of understanding of social distancing compliance on a mass scale.

There have been some attempts to bring more insight into this problem. Abouk and Heydari [2] in their study to evaluate the effectiveness of the six most common social-distancing policies in the United States, found that stay-at-home orders were the greatest contributor to reduced mobility in the early stages of the COVID-19 pandemic. Though, they could not conclude how such policies affect voluntary social distance practices during the pandemic. In a related study, Sausen et al. [22] proposed the use of electricity consumption measured in households as a means of describing social distancing behavior under government wide

lockdowns. Though this a clever approach that takes advantage of energy data, its approach is not able to show how individuals practice social distancing. Al-Hasan et al. [4] conducted a survey experiment to view how individuals across different countries coped with COVID-19 related recommendations. Though they found significant correlations between individual adherence of social distance guidelines and their perceived threat of COVID-19, it is unknown to what degree adherence is observed since the survey questions were generalized. These studies attempt to bring to light how society has responded to social distancing recommendations set forth by their governments. Though they provide some useful insights, none directly evaluate how social distancing is practiced at the individual level in a public setting. To be even more specific, how social distancing is observed in outdoor environments has yet to be comprehensively studied. Thus, there is a clear need for an approach that evaluates social distancing in outdoor settings in a more direct manner.

The first reported COVID-19 case in the United States took place in the state of Washington. As a response, a National Science Foundation (NSF)-supported Rapid Response Research (RAPID) project was granted to the University of Washington researchers. The project entails conducting street view surveys across a broad cross section of Seattle to collect data over time on the community impact of the pandemic. The street view data comes in the form of high-quality 360-degree video, akin to Google street view. This rich dataset presents an opportunity develop a routine that investigates social distancing adherence on a city-wide scale with high resolution imagery. In the context of understanding social distance practices in outdoor settings, my thesis aims to develop a method that takes 360-degree video data to directly measure the distance between pedestrians at an individual level with an accuracy useful for adjudicating the social distancing compliance.

Chapter 2
BACKGROUND

Measuring distances of objects have primarily taken two approaches: passive and active methods. Active methods send signals to the object (i.e. laser beams, radio signals, ultra sounds, etc.) to measure distances, whereas passive methods do not send signals and use other information like object position or light [23]. Given the equipment limitations from the RAPID vehicle camera, my thesis will mostly focus on passive methods. Detecting distances from images or videos has become a prominent subject of study within the field of computer vision. Recent advancements in image-based distance estimation has played a vital role in the improvement of autonomous vehicles and their use of public roads worldwide. Other use cases have also emerged such as detecting Foreign Object Debris (FOD) and their spatial location in airports or aircraft carriers as a means to prevent perennial hazards [28]. Given the practical applications of image distance estimation, an overview of methods including utilization of specialized cameras, 2-D to 3-D transformations methods, and deep learning frameworks describe the relevant research efforts that guide how my thesis estimation of social distances between pedestrians will be addressed.

2.1 Specialized Cameras and Configurations

If the objective of a certain experiment is to retrieve precise distance measurements, special camera configurations are typically used. Depth, 3D, or RGB-D cameras typically use stereo vision (i.e. multiple lens) to aid in three-dimensional reconstructions of a given environment [11]. Such a setup can be advantageous as described by Bylow et al. [5] where a method is outlined to reconstruct 3D static indoor environments using RGB-D cameras. However, it is not addressed how this methodology performs in outdoor and dynamic environment

which my thesis aims to explore. These types of cameras would typically be the first option for experiments that conduct object distance measurements, but due to their high costs, alternative techniques are often considered. In an effort to circumvent purchasing depth cameras, Adi and Widodo [13] carried out a distance measurement research experiment using multiple cameras to emulate stereo vision. While yielding favorable results, their experiment is predicated on fixed camera locations of known distances for which there is no clear equivalency to the 360-degree camera set up of the RAPID vehicle. An additional approach that has been suggested is via 3D structured light computations. This approach, as demonstrated by Ribo and Brandner [21], combine a camera and light source to recreate an active stereo system used for the reconstruction of 3D environments. As stated previously, these specialized cameras are specifically manufactured for 3D reconstruction of environments which does resemble the RAPID camera set up. Furthermore, these specialized cameras are not intended to be used for capturing high resolution photos while mounted on a vehicle traveling at more than 40 miles per hour.

2.2 Transformation from two dimensions to three dimensions

The crux of my thesis's objective is to find a feasible method of taking two dimensional images and projecting them to a three dimensional space where distances between objects can then be measured. Prematunga and Dharamatne [20] provide a proof that it is possible to recover the position (or coordinates) of an object using a single 2D image, given the size and shape of the image while guaranteeing a certain accuracy. These findings lend way to several techniques to accurately reconstruct three dimensional features.

One method worth exploring is the technique of *Structure from Motion* (*SfM*). At a high level, *SfM* refers to the process of inferring three dimensional structures from two dimensional transformation of their projected images when three dimensional information has not been conveyed [25]. A practical application of this was carried out by Fields et al. [8] where they use a single camera mounted on a ground vehicle to recreate terrain information in the range of 100 meters to several kilometers away. Although this application has parallels

to the RAPID vehicle camera set up, it is only able to recreate static terrain environments such as buildings, grass, or trees in the distance. It not does address how mobile pedestrians would be captured and how they would appear in a fully reconstructed three dimensional environment.

More modern novel approaches have also been explored that specifically address object distance estimation. Chen et al. [6] propose the use of smart phones with accelerometers to back-calculate the orientation and transform pixel ratios to real distances given the position of known objects. Pixel-ratio refers to the conversion of pixel distance to real world distance in an image (e.g. 300 pixels = 20 ft). This approach seems to work for fixed environments where reference objects can be strategically located for calibration purposes. Though this does not address changing environments where reference objects can no longer be utilized as the RAPID vehicle progresses through the route. Additionally, the RAPID vehicle camera is not equipped with an accelorometer, which is problematic since the orientation of camera can change due to driving route inclines of the vehicle. Separately, Jiang and Jiang [12] take advantage of circular properties of known objects in the single image (e.g. compact discs or circular pillars) and use projective geometry for distance measurements. However, this approach is only for a single static image as opposed to dynamic, changing environments that is being explored in my thesis.

2.3 Deep Learning Frameworks

As mentioned before, autonomous vehicle (AV)-based research has come to the forefront of object detection and depth perception studies. Advancements in these fields have a crucial role in the implementation of AVs on roads and their overall public acceptance. This, along with other applications have led to sophisticated methods of using a deep learning frameworks for object distance estimation in monocular images, which are images that do not retain any depth information. One such general approach was conducted by Eigen et al. [7] where they stack two networks: one network that makes a global prediction of the entire image and the other refines that prediction locally. These networks are trained on NYU Depth and the

Karlsruhe Institute of Technology/Toyota Technological Institute (KITTI) datasets, which are common in training depth-estimation deep learning networks since they have associated depth information. However, the results of these trained networks are generalized depth maps of an image with reference to the camera location, rather than discrete distances *between* detected objects. In a similar effort, Liu et al. [15] employ a convolutional neural network that effectively groups pixels of certain categories (i.e. sky, ground, buildings) as "superpixels" or segmentations. These segmentations are then distinguished into separate depth categories learned from ground truth data. This approach is also predicated on learning from the NYU Depth dataset, which does not match the monocular images that are present in this thesis. A slightly different approach explored by Ahmed et al. [3] uses the YOLOv3 overhead view dataset to detect pedestrians and estimate their social distances. The use of the overhead view dataset has the advantage of omitting depth estimation as all pedestrians detected will be on the same ground level. Unfortunately, this limits its generalizability, especially for street view type imagery that my thesis is analyzing.

Zhu et al. [30] likely explore the most sophisticated and relevant approach to this problem. Although this also employs a convolutional neural network with the NYU Depth and KITTI datasets, they specifically detect pedestrians and their corresponding distance from the camera. While this results into quite accurate findings, the estimated difference of pedestrians from the camera cannot be used to calculate the distance between pedestrians unless certain geometric properties are known. These geometric properties include the angle of the detected object in reference to the camera and the pixel-ratio it employs. This information cannot be extracted easily since there is no metadata associated with these values. This approach may have some feasibility, it would still require additional sophisticated method to an already complex problem, where an alternative approach may be preferred. There are several other deep learning framework techniques that have been explored [9, 14, 27, 29] but have similar shortcomings previously discussed in the context of applying their respective methods for addressing my thesis's objective.

Chapter 3

METHODS

This chapter will describe the method proposed for the estimation of distance between pedestrians in images. This will include an explanation of the processing of images, an overview of the equations driving the calculation, suggested physical experiment and calibration used further develop my research objective.

3.1 Image Processing and Pedestrian Detection

The pre-processing methods described in this section were scaled and parallelized by Matthew Martell, a Ph.D. candidate in the Industrial and Systems Engineering department from the University of Washington. The raw data used for this study is generated by the iSTAR Pulsar+ by NC-Tech, a high resolution 360 degree camera that is mounted atop a 2018 Toyota Prius Prime (RAPID Vehicle). As a run commences, the vehicle drives through a predetermined route and records video for the duration of the run. Each run records approximately eight hours of 360 degree video. Upon the completion of a run, the video is processed into frame by frame jpeg files of resolution 11000x5500; a sample output of an image can be seen in Figure 3.1.

Figure 3.1 shows that in order to view the full 360 degree image, distortion effects are present as the image is flattened. This type of image is not ideal for pedestrian detection algorithms as a majority of well-established pedestrian detection algorithms use non-distorted images for training. Therefore, it is imperative to convert the pre-processed images that correct the distortion to achieve the best results. To do this, Mutha [17] implemented a gnomonic projection, which is applied to the images and effectively projects points from a sphere onto a plane. This projection along with a bilinear interpolation results in a suitable

transformation of the images. The result of the conversion are two cropped and corrected images with half the size of the original image resolution at 5500x2750.

Figure 3.1: 360 degree sample image from the RAPID vehicle between Paul G. Allen Center (left) and Bill & Melinda Gates Center (right) at 3800 E Stevens Way NE, Seattle, WA 98195, USA (Lat: 47.653535, Long: -122.305175).

Once the images have been corrected, a pedestrian detection algorithm can be applied to identify all pedestrians in a single image. Pedestron [10] is a state-of-the-art pedestrian detection deep learning framework that is used to identify the locations of pedestrians in an image. The output of Pedestron are four coordinate points for each detected pedestrian that indicate the vertices of a bounding box. The coordinate axis used to reference the vertices are located at top left of the image. An example of applying Pedestron to an image is seen in Figure 3.2. The coordinates of each bounding box is designated by a unique set of points labeled $(x_{min}, x_{max}, y_{min}, y_{max})$. These unique identifiers for each bounding box will be used to develop the distancing estimation described in the subsequent section.

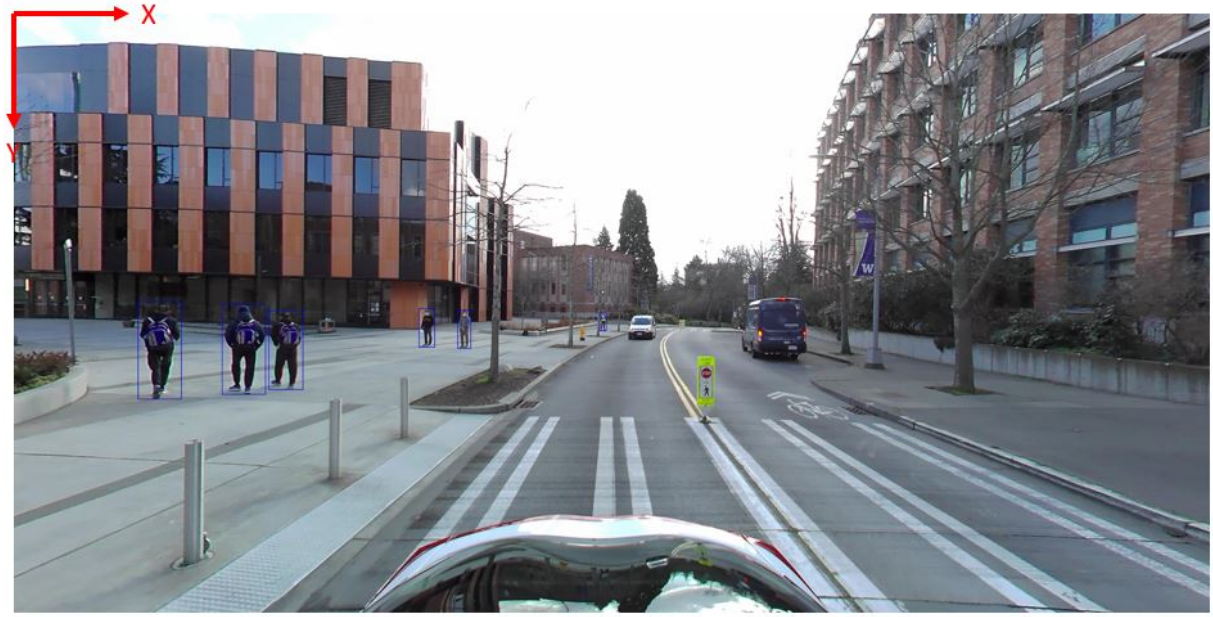


Figure 3.2: Pedestrian detection with bounding boxes and coordinate axes.

3.2 Social Distance Estimation

Chapter 2 discussed several techniques aimed at calculating distances between pedestrians. However, many of those techniques could not be applied given the uniqueness of the image data in this thesis. Thus, the task of developing a method that precisely accomplishes this is non-trivial. However, given that my research objective is to estimate pedestrian's social distance, an approximation method is instead proposed. This approximation enables the use of an approach that takes advantage of geometric properties apparent in outdoor images. This approach effectively uses a distance estimation based on Pythagorean's Theorem to determine the distance between pedestrians. Figure 3.3 shows a bird's-eye illustration of two pedestrians, their triangular positioning, and their corresponding horizontal component, $d_{s,x}$, and the depth component, $d_{s,y}$.

Estimating depth in a single image is perhaps most difficult aspect of this estimation problem. Van Dijk and De Croon [26] completed a comprehensive study for how several

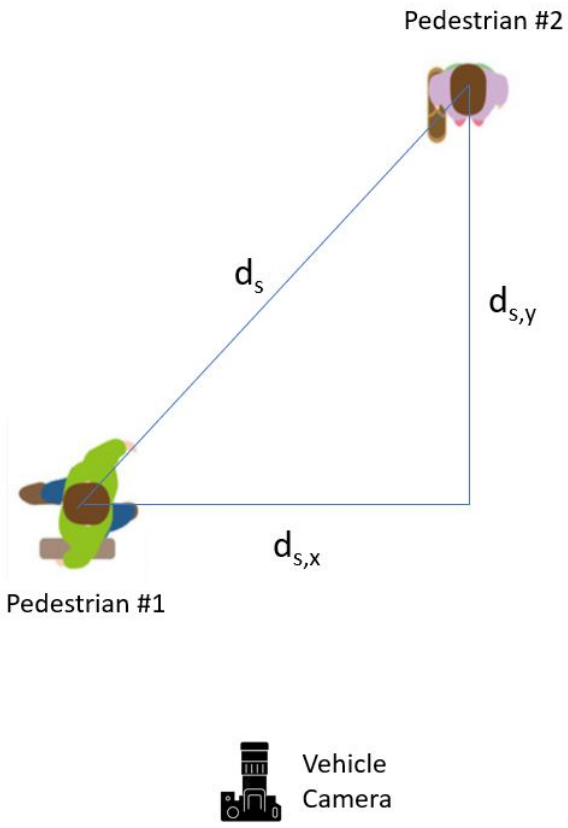


Figure 3.3: Bird's-eye view of pedestrian detection.

prominent neural networks learn depth in single images. Since their study primarily centered around autonomous vehicle object detection, the image capturing environment is exclusively outdoors. This is important to note because outdoor images tend to have a horizon line; the line where the ground and sky are divided. Their study found that the position and size of an object are key indicators of their depth. Particularly, objects that are further away appear higher in an image and closer to the horizon line. Additionally, an object's position on the ground is determined by their vertical position in the image. Evidence of this can be noticed in Figure 3.4 where three test pedestrians are detected with bounding boxes and the furthest pedestrian is both the smallest and highest located.

Noting these observations leads to the following equation to estimate depth in an image

Figure 3.4: Depth location of pedestrians

[26]:

$$d_{object} = \frac{f}{y - y_h} y_{cam} \tag{3.1}$$

The value of f from equation (3.1), which is measured in pixel units, represents the focal length of the image. Typically, this value is gathered from the native settings of the camera. However since the original image has been transformed from its 360 degree field of view with distortion effects to its processed version, the original focal length of the image will be altered. Estimating this value will be presented in the subsequent section. The value y_h, also having pixel units, represents the location of the horizon line. This value can vary depending on orientation of the gnomonic projection and cropping of the original. For this study, the horizon line is estimated to be the midpoint of the vertical image size; Figures 3.2 and 3.4 roughly confirm this. y_{cam} is the vertical position of the camera, in units of feet, which can be directly measured from the camera mounted on the vehicle. Lastly, y is the vertical position on the ground of the pedestrians, measured in pixels. Since the bounding boxes generally

fit the detected pedestrian tightly, the bottom coordinate of the bounding box can used for y; in this case it is y_{max} of the bounding box coordinates. Therefore, equation (3.1) can be written as:

$$d_{s,depth} = \frac{f}{y_{max} - y_h} y_{cam} \tag{3.2}$$

Determining the horizontal component will depend on the ratio between real width to image pixel width. That is, if the real distance between two points in the image is known, then its corresponding pixel distance can be used to calculate a pixel-width ratio. This ratio can then be applied to calculate the real distance between any two points. Unfortunately, a single pixel-width ratio cannot be universally applied given the prominent depth features in images discussed previously. For example, an object will decrease in pixel width the further that object appears in the image, thereby changing the pixel width ratio. One way to handle the dynamic feature of depth is by taking advantage of the bounding box widths. As stated before, the bounding box widths of the detection algorithm tightly fit the pedestrians has an upper and lower bound on its width. Generally, this will no larger or smaller than the actual width of a person. Discussion on determining the ideal pixel-width ratio is presented in the subsequent section. Thus, the following equation is used for determining the horizontal distance between two points in an image:

$$d_{s,x} = \rho_h d_{pixel}. \tag{3.3}$$

where

$$\rho_h = \frac{w_{BB,actual}}{w_{BB,pixel}}. \tag{3.4}$$

The value $w_{BB,actual}$ represents the estimated real width of a bounding box in units such as feet and $w_{BB,pixel}$ is the width of the bounding box measured in pixels. Lastly, d_{pixel} represents the pixel distance between two points of interest in an image. In the context of this thesis, the two points will the be the midpoints of bounding boxes of two pedestrians.

Given the breakdown of horizontal and depth equations, a full calculation of the distance between two pedestrians can be formulated. Figure 3.5 accompanies the formulation as it

shows the geometric logic used to calculate the distance between pedestrians. First, the depth of each pedestrian is calculated using equation (3.2) represented by d_1 and d_2. The difference is then taken to arrive to the following:

$$d_{s,y} = |d_2 - d_1|. \tag{3.5}$$

Next, the horizontal distance between the two pedestrians is then calculated with equation (3.3) using the midpoints of each bounding box. The choice of using either bounding box for $w_{BB,actual}$ of equation (3.4) is chosen to be larger of the two bounding boxes as it considers a more conservative approach in the context of my thesis; it will calculate the a smaller horizontal distance between two pedestrians. Once both values of $d_{s,x}$ and $d_{s,y}$ are determined, the following distance formula, based on Pythagorean theorem, is used to estimate the social distance between pedestrians:

$$d_s = (d_{s,x}^2 + d_{s,y}^2)^{1/2}. \tag{3.6}$$

3.3 Ground Truth Experiment

The methods described in Section 3.3 utilize equations with unknown parameters. Particularly, equation (3.2) requires the value of f, the focal length of the image, and equation (3.3) uses the value of w_{BB_actual}, which represents the estimated width of the bounding box in real units. Typically, the focal length can be retrieved from the camera specifications (as discussed in section 3.3) or estimated with the following equation provided by Dijk and de Croon [26]:

$$Z = \frac{f}{h}H. \tag{3.7}$$

From equation (3.7), the focal length can be calculated if Z, the real depth of an object, and H, the known height of an object are known in an image which can be done with a physical experiment. This generally works well if the object is centered in the image and the field of view is relatively narrow, which is not the case for pedestrians detected at the edge of wide-view images. Equation (3.7) is still helpful to understand general ranges of focal

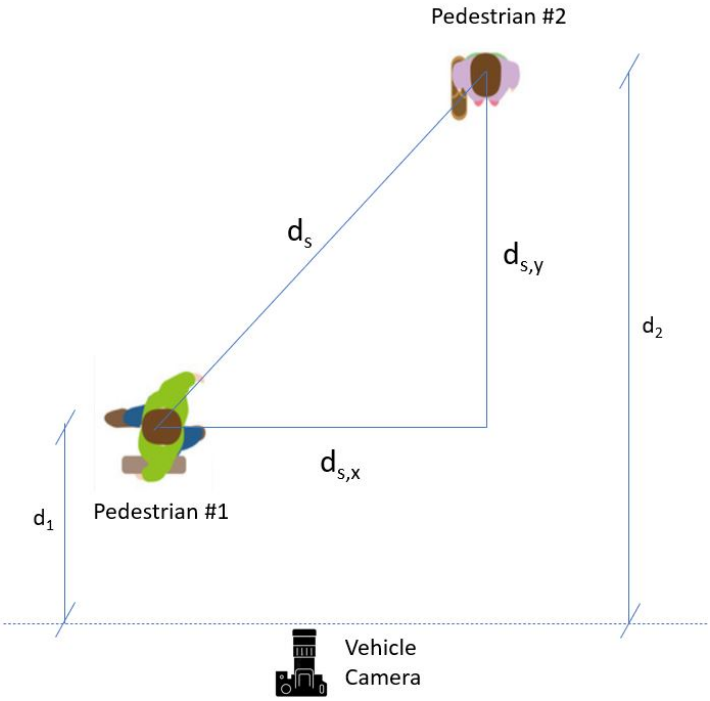


Figure 3.5: Triangular representations of pedestrians

length values that can be tested and calibrated (e.g. using the heights of test pedestrians and their known distances from the camera). The real width of a bounding box can also take a possible range of values limited by the width of an individual. Given that these two parameters have a range of values, a ground truth dataset is necessary to calibrate and optimize these parameters for the distance equations.

In order to generate ground truth data, a physical experiment needs to be conducted where the values of distances between test pedestrians are known. For my thesis, the experiment was designed to capture several factors based on the parameters that could effect the distance estimation and Pedestron's performance under different configurations. These factors include pedestrian distance from vehicle, distance from each other, orientation of pedestrians, stance, and whether pedestrians are standing or walking. A series of runs were composed to consider various combinations of these factors. Table 3.1 displays all the runs and factors that were used for the ground truth data of this thesis paper. Figure 3.6 accompanies this table by showing the layout for position of test pedestrians of the experiment. The image angle is also a possible factor, but given that processed images are taken frame by frame, many angles for the same run will automatically be captured as the vehicle camera drives by the test pedestrians. It is important for the experiment to exhibit accurate positioning for each a run in order to avoid unwanted noise. To mitigate this, the runs were measured and marked prior to the start of the experiment. Figure 3.7 shows the type markings used so that test pedestrians (actors) can position themselves accurately and consistently prior to start of each run. For the walking stance, a rope of marked-out length was used to maintain the fixed length of each pedestrian as they walk based on their positioning.

Upon the completion of the experiment, the images can be processed by the techniques described in section 3.1. This produces ground truth data that includes the image name, associated bounding boxes, and real distance between pedestrians. This gives way to perform a grid search of the adjustable parameters from section 2.3 to find the optimal parameters that will generalize for future unseen images. The grid search is used to scan through the parameters to minimize the difference error of the ground truth values and the estimated

(a) Perpendicular Position

(b) Parallel Position

Figure 3.6: Positioning of test pedestrians for ground truth experiment

distance. Particularly, the root mean square error (RMSE) is suggested for its penalization of outlier values and direct interpretation in the context of estimating social distance. Further description and execution of optimizing the RMSE metric is described in Chapter 4.

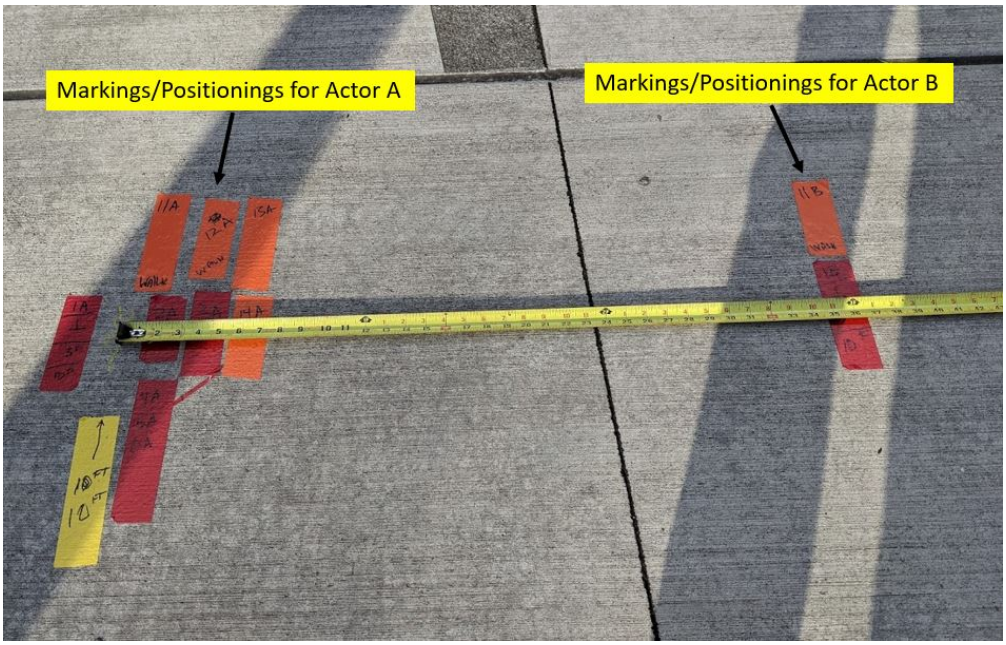


Figure 3.7: Position markings for actors A and B

Table 3.1: Streetview Ground-Truth and Calibration Measurement Experiment

Run No.	Stance	Position	Separation Distance (ft)	Distance from Car (ft)
1	Standing	Perpendicular to car	3	10
2	Standing	Perpendicular to car	6	10
3	Standing	Perpendicular to car	12	10
4	Standing	Parallel to car	3	10
5	Standing	Parallel to car	6	10
6	Standing	Parallel to car	12	10
7	Standing	Perpendicular to car	3	20
8	Standing	Perpendicular to car	6	20
9	Standing	Parallel to car	3	20
10	Standing	Parallel to car	6	20
11	Walking	Perpendicular to car	3	10
12	Walking	Perpendicular to car	6	10
13	Walking	Parallel to car	3	10
14	Walking	Parallel to car	6	10
15	Sitting	Parallel to car	3	15
16	Sitting	Parallel to car	6	15
17*	Standing	Perpendicular to car	6	10
18*	Standing	Perpendicular to car	6	20

*Pedestrians are facing car in run numbers 17 and 18.

Chapter 4

RESULTS

This chapter will overview the results from implementing the methods described in Chapter 3. This includes results from processing of the images, grid search optimization using ground truth data, and visual performance on test data.

4.1 Image Processing Results

Recalling from Section 3.1, two methods were used for the processing of the images. First, the distortion effects are corrected via a bilinear interporlation which effectively separates the 360-degree image into two images with minimal distortion. A result of this is seen with Figure 3.1 being converted into two images as shown below in Figure 4.1. As can be seen in Figure 4.1, the original 360 degree image has been corrected into two images that show the front and rear views, which are more suitable for pedestrian detection algorithms. The second step in pre-processing the images is the application of the Pedestron algorithm to identify the pedestrians in an image. Figure 3.2 shows an example output of this, but this is extended in Figure 4.2 where more variability in the results are shown. These flattened images with detected pedestrians will serve as observations to run the grid search optimization.

4.2 Grid Search and Visual Results

As discussed in the previous chapter, ground truth data is established in order to determine the optimal unknown focal length and $w_{BB,actual}$ parameters. Originally, this amounted into a total of $n = 355$ ground truth observations, however upon closer inspection, four observations were removed because they unduly inflated errors. Further explanation on the removal of these outliers are presented in Chapter 5. With $n = 351$ observations, they are randomly

(a) Front view

(b) Rear view

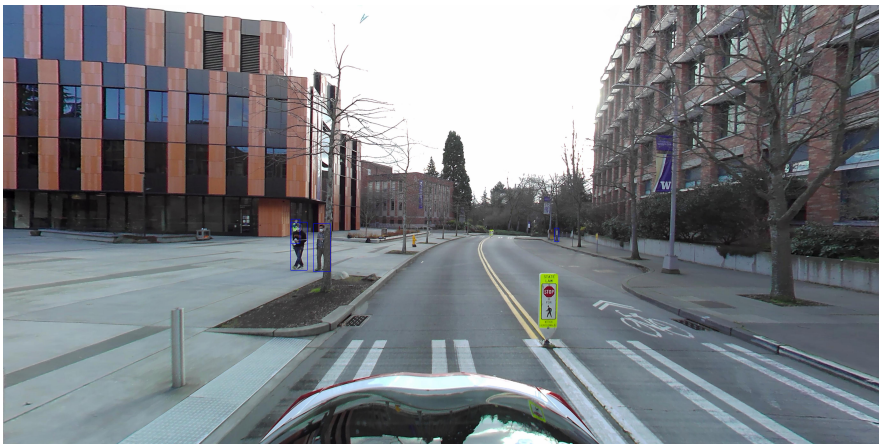
Figure 4.1: Corrected images via gnomonic projection.

split into an 80:20 training/test set. The training set is used to implement a grid search on the unknown parameters with the goal of minimizing the root mean squared error (RMSE) in equation (4.1), where d_{truth} represents the known distance and d_s is the estimated distance.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{T}(d_{truth,i} - d_{s,i})^2}{T}}. \tag{4.1}$$

To perform the grid search, a range for the parameter $w_{BB,actual}$ is considered from 1 ft to 3 ft in increments of 0.1 ft, based on reasonable assumptions of the bounding box widths. That

(a) Variable sizes of bounding boxes.



(b) Bounding box with natural object occlusions.

Figure 4.2: Bounding box of pedestrians in images.

is, based on observations seen in Figures 3.2 and 4.2, the bounding box widths will not be much smaller or larger than the actual width of a pedestrian. The focal length values range from 1000 pixels to 1500 pixels, incremented by 10 pixel, based on the upper and lower bound values given by known test pedestrian values and equation (3.7). For example, table 3.1 can be used with (3.7) since the distance from vehicle (Z), actual height of test pedestrian (H), and the pixel height of the test pedestrian (h) are all known values for several configurations.

Running this grid search yields optimal parameters of focal length = 1210 pixels and

$w_{BB,actual}$ = 2.2 ft with a training RMSE of 1.20 ft. A three dimensional contour plot is constructed to show the general convexity of the RMSE function, which is seen in Figure 4.3. These parameters are then evaluated to see how they perform on test images, resulting

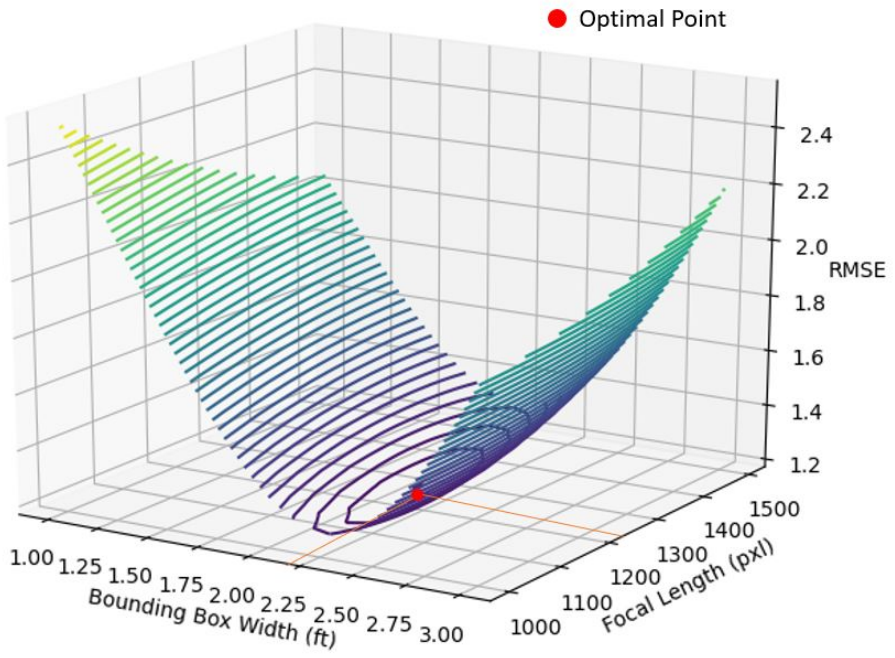
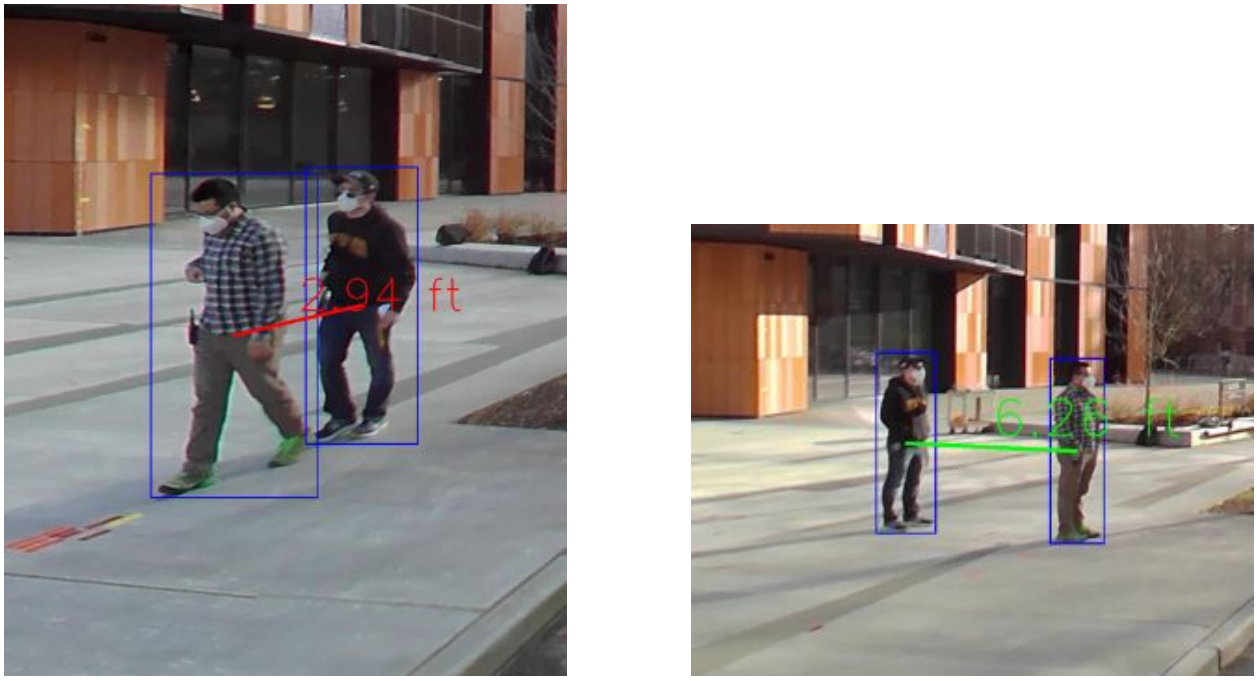


Figure 4.3: 3D countour plot of Training RMSE

in a test RMSE of 1.13 ft. Knowing these optimal parameters, a 95% confidence interval is constructed to estimate to the bounded range of the true RMSE. This is done via bootstrapping where 10,000 bootstrapping samples, each of size $n = 351$, are generated 10,000 RMSE samples. The 2.5, 97.5 percentile values are then calculated resulting in a $[1.07, 1.41]$ 95% confidence interval range, measured in feet. Visually, these parameters are then used to draw social distance lines between pedestrians on the physical experiment images. Samples of these outputs can be seen in Figure 4.4, where two images of differing test pedestrian configurations are annotated with their corresponding estimated distances.

Using ground truth data is the most direct way to measure how well the estimation



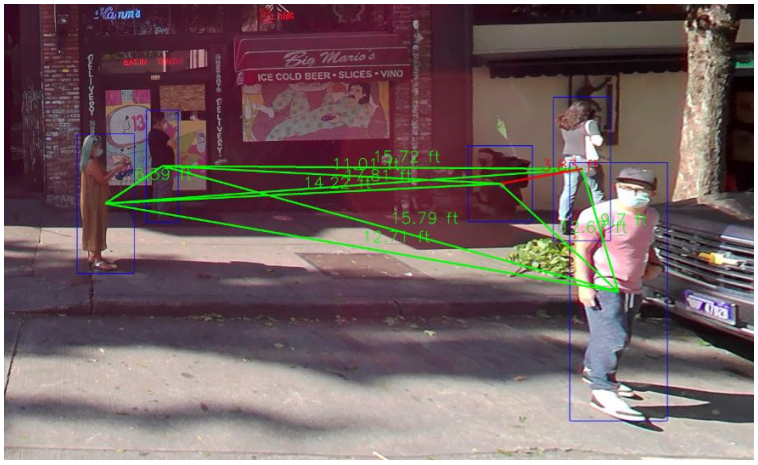
(a) Walking Parallel, 3 ft of Social Distance.

(b) Facing car, 6 ft of Social Distance.

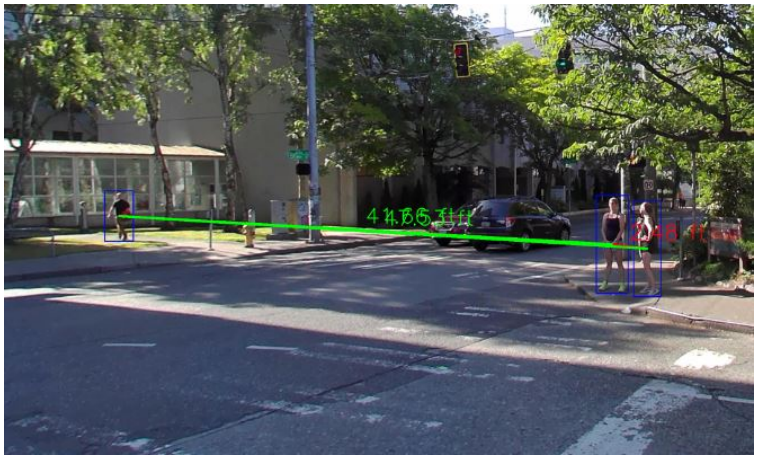
Figure 4.4: Drawn Social Distance Line Results.

method performs. However, another indirect way of measuring performance is to apply the process to a completely unseen set of images. Using a separate set of images from one of the runs done by the RAPID vehicle in Seattle, the image processing and distance estimation is applied which results in the sample outputs of Figure 4.5. The true distances between pedestrians are unknown. However, the knowledge of the approximate sizes of adults and background objects (e.g., trash can) suggests that the estimated distances are reasonable.

(a) Sample output #1.



(b) Sample output #2.



(c) Sample output #3.

Figure 4.5: Sample output of Pedestron/distance estimation on unseen images (green lines signify social distances more than six feet and red lines less than six feet).

Chapter 5

DISCUSSION

This chapter discusses several topics regarding the overall applicability, limitations and future work of the distance estimation method.

5.1 Applicability

Chapter 4 reported the results of the grid search to find the optimal parameters for estimating distances on unseen images. The test RMSE value of 1.13 ft is reported, which at face value, seems reasonable. To be more rigorous, a 95% confidence interval was constructed to understand the true standard deviation quantity of the estimated social distance method. Given that the distribution of the distance estimation is unknown, re-sampling the data via bootstrapping provides a means to construct a bootstrap confidence interval. However, using this technique is contingent on having representative and comprehensive data. Recalling the set-up for the ground-truth experiment in Section 3.3, the objective was to capture as many scenarios that characterise how pedestrians might behave as well as how they appear in images. This includes the stance, positioning, separation distance and distance from car as shown in Table 3.1. One factor that is implicit in the ground-truth experiment is the angle at which the camera captures the test pedestrians. For any given run from Table 3.1, the vehicle will drive by for a fixed test pedestrian configuration and capture a relatively wide field of view. Figure 5.1 demonstrates this by showing the path of travel for the RAPID vehicle capturing frame by frame instances, thereby creating a wide field of view for a fixed configuration. All these factors combined then help construct the 95% confidence interval of $[1.07, 1.41]$. This test of the estimation method on a dataset from different settings supports the validity and potential for generalizability of the method.

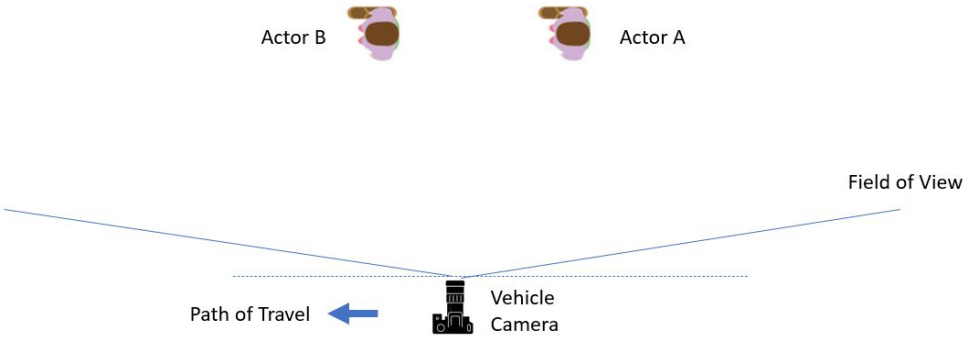


Figure 5.1: Field of view for RAPID vehicle of test pedestrians.

In the context of determining whether pedestrians are violating certain social distancing thresholds (e.g. 6 ft), this type of result can be effective to characterize a majority of pedestrians who are above or below such thresholds. Figure 4.5 showcases what this might look for general images as green lines indicate social distances greater than 6 feet and red lines indicate social distances less than 6 feet. Depending on the research that is being conducted, the test RMSE reported earlier may not be adequate, particularly where pedestrians are social distant near the threshold. A suggestion could be to use the 95% confidence interval from the results to create maximum and minimum estimated distances to determine decision rules that resolve conflicting violations. Another suggestion could to be to create an additional category that collects instances that are near the threshold and develop a certain hypothesis from such results. For example, one study might look at the distribution of pedestrians that are not socially distant (less than 3 feet apart), socially distant (more than 9 feet apart) and a middle category between those ranges. Regardless of its application, one major advantage of this estimation method is the relatively low computational time it requires. For example, if there a total of m bounding boxes detected in a image, then that image will require mC2 total operations of equation (3.6). Compared to the deep learning networks from Section 2.3, this might be more preferable, especially if there are a large set of images.

5.2 Limitations

The original number of ground truth observations was 355, however upon observing some initial results, it was clear that there were some outliers present. Figure 5.2 shows an estimation of distance with the ground truth distance of 6 ft. It is seen that the estimation is

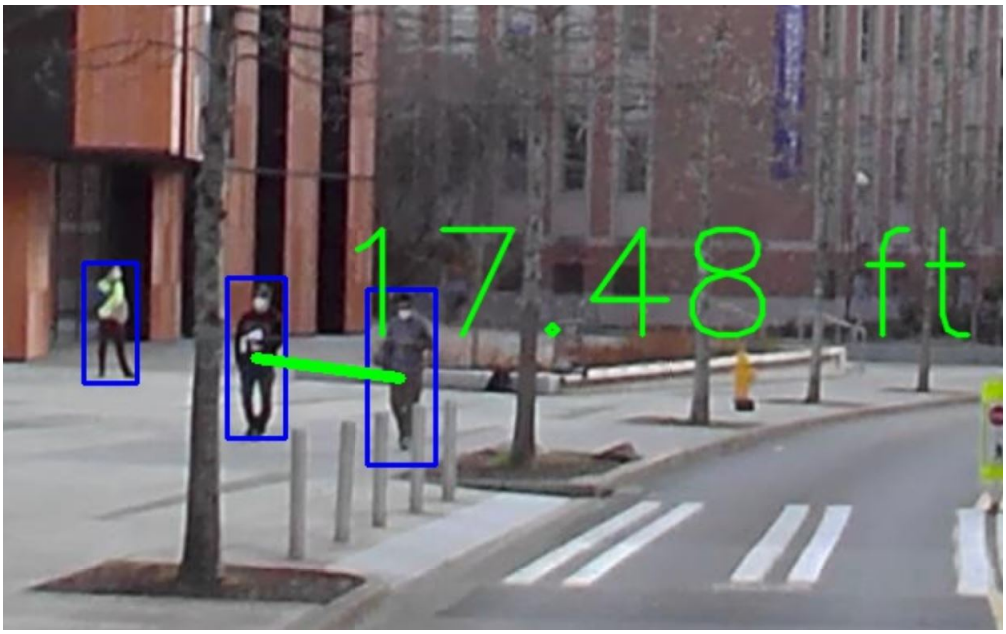


Figure 5.2: Distance estimation with ground truth distance of 6 ft.

17.48 ft, which results in a difference of 11.48 ft from the ground truth distance. Upon closer inspection, Figure 5.2 displays a test pedestrian who is occluded by a bollard, resulting in a bounding box that is larger than the size of the actual pedestrian. Given that the depth component of the estimation from Section 3.3 is sensitive to the bottom edge coordinate position of the bounding box, the distance is overestimated. While this is a possibility that can occur from Pedestron's algorithm, a limited amount of these observations can dominate the RMSE when performing a grid search optimization. The optimization becomes overly conservative due to the outlier-inflated RMSE so the tuned parameters lead to under-estimation of distances for most of the instances. For this reason, these type of outliers were

omitted from the ground truth observation data. This instance may cause concern for similar scenarios where there might be a height differential between two pedestrians (e.g. a child and their guardian). However, recalling that the equations developed in Section 3.3 only rely on the bottom coordinate of the bounding box for its' depth component, this should not be an issue as long as the bounding box reasonably fits a pedestrian. This is demonstrated in Figure 5.3 where a child and their guardian are detected to be close despite their height differential.

Figure 5.3: Distance estimate of child and their guardian.

The distance estimation method proposed in my thesis is heavily dependent on the Pedestron algorithm. Specifically, the methods utilize the output that assume certain geometric properties of bounding box coordinates (e.g. tight fitting bounding boxes). There are other object detection frameworks such as the work by Fields et al. [8] that identifies objects as segmentations rather than rectangular bounding boxes. Since segmentations do not have reliable geometric references, the methods of my thesis can not be applied. Additionally, the

errors associated with Pedestron will be absorbed into the distance estimation. The estimation does not discriminate between true/false positives as it only depends on the presence of a bounding box to implement the estimation. Thus, a certain amount of consideration must be taken to evaluate the performance of candidate pedestrian detection algorithms.

The estimation method is also limited to images taken in outdoor settings. Specifically, the images must capture the presence of a horizon, as the equations from Section 3.3 require a pixel reference of the horizon. This is generally inherent to outdoor images as there is enough depth developed to understand the location of the horizon; even with presence of tall buildings. Along with the outdoor image limitation, the estimation method also relies on detected objects being located on the ground. For example, a driver in a vehicle is not located on the ground and any bounding box associated with their detection may not properly reflect their depth in an image; though the vehicle the driver is operating is on the ground and would be a valid candidate for distance estimation if warranted. Separately, persons detected that are located at high locations above ground (e.g. balconies) will likely be located above the horizon line. Figure 5.4 shows an example of a person on a balcony and their corresponding estimated distance to another pedestrian. Though Figure 5.4 shows a reasonable output for their social distance, the estimation method is not being implemented as intended since the person on the balcony is located above the horizon line. These examples should likely be excluded from any study as they do not fit the proper description of pedestrians. Furthermore, special care should be taken when considering edge cases and how they apply to a particular study.

5.3 Future Work

A natural continuation of the methods described in my thesis is to implement them to the mass set of data that is being collected by the RAPID vehicle. At its most direct application, a study can track the social distancing compliance of inhabitants in Seattle through the progression of the pandemic. Studies can also determine if pedestrians are in groups and what their group size is. These ideas can become even more fine-grained by

Figure 5.4: Person detected on balcony.

separating the data into regions/neighborhood of a city and determining their corresponding social distancing metrics, along with other characteristics unique to those regions. While the RAPID vehicle is primarily focused on capturing 360-degree data during the pandemic, similar 360-degree image data (e.g. Mapillary/Google Street View) can be used. In essence, many research hypotheses can be formed around determining social distancing metrics in a

given city if the appropriate image data is available.

Even though this thesis was initially based on determining social distance in the midst of a pandemic, it can be applied to other similar datasets. This could include determining seasonal changes of social distancing and its relation to other more common ailments like the flu or cold. It could also be used in other catastrophic events (e.g. earthquakes, hurricanes, etc.) in a location to investigate how the movement/distance of persons is altered. The methods can also be extended to objects other than pedestrians. As an example, bicyclist and vehicles are viable objects that could help evaluate how bike-friendly a city is by estimating the average distance between cyclist and vehicle. The work by Tuohy et al. [24] use road markings as a known distance reference to compute distances from vehicles, which can be used for this bike-friendly town evaluation. Theses types of tasks are usually accompanied by massive image datasets, so implementing this distance estimation method can help with returning results with low computational strain. This low computational reliance can be extended to video data where detection algorithms can be used with the distance estimation to provide live feedback of social distancing.

Chapter 6

CONCLUSION

In the midst of the COVID-19 pandemic, locations around the world have enacted safety protocols that include social distancing. Some studies have attempted to quantify how well members in a community have complied to such safety recommendations through survey experiments or tracking publicly available mobility data. Though these methods provide some insight, they indirectly measure social distance compliance, focus only on indoor settings, and are generally inconclusive regarding social distancing practices at the individual level. Thus, a need for a method that directly evaluates the general conformity to social distancing protocols in outdoor environments can be valuable to future research endeavors. With the backing of an NSF-RAPID project, 360 degree video street view surveys of Seattle, Washington have been conducted and present an opportunity to use images to address this research question. Typical methods of estimating distance of objects in images use specialized cameras with stereo vision meant to capture depth or utilize computationally complex ideas such as *structure from motion* that require static environments. Perhaps the most sophisticated techniques implemented are the deep learning frameworks where the training process requires three-dimensional metadata. This type of metadata is not easily attainable and also requires special cameras, which are not compatible with the RAPID vehicle that may be moving at 40+ miles per hour. Therefore, my thesis describes the development of a method for estimating distance of pedestrians from images using geometric properties.

Initially, the 360-degree image data undergoes a projection called biliear interpolation that removes distortion in the images to make it more suitable for object detection algorithms. The distortion-free images are then fed to Pedestron, a state-of-the-art pedestrian detection algorithm that outputs bounding boxes of detected pedestrians. These bounding boxes

in tandem with certain geometric properties of depth are used to develop social distance equations with adjustable parameters. In order to optimize these parameters, a physical experiment was designed to generate ground-truth data, which was then used to evaluate the performance of the methods with test data. This experiment yielded a test root mean square error (RMSE) of 1.13 ft, which represents the mean error that should be expected for a given distance measurement. Additionally, a 95% confidence interval was constructed for the true value of RMSE, which was determined to be $(1.07, 1.41)$. These results can help guide certain hypotheses regarding social distancing metrics so long as the errors and confidence interval are within the acceptable range of a study. This could include tracking the general compliance of outdoor social distancing through the progression of the pandemic, identifying the size of social groups or even non-COVID related studies such as evaluating pedestrian patterns during the flu-season. Lastly, the social distance equations are well suited for assessments of large datasets since the total distance estimation operations on a single image are no more than the total number of object-to-object distances, making it computationally efficient.

BIBLIOGRAPHY

[1] Social distancing. https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/social-distancing.html.

[2] Rahi Abouk and Babak Heydari. The immediate effect of covid-19 policies on social-distancing behavior in the united states. *Public Health Reports*, 136(2):245–252, 2021. PMID: 33400622.

[3] Imran Ahmed, Misbah Ahmad, Joel J.P.C. Rodrigues, Gwanggil Jeon, and Sadia Din. A deep learning-based social distance monitoring framework for covid-19. *Sustainable Cities and Society*, 65:102571, 2021.

[4] Abrar Al-Hasan, Jiban Khuntia, and Dobin Yim. Threat, coping, and social distance adherence during covid-19: Cross-continental comparison using an online cross-sectional survey. *J Med Internet Res*, 22(11):e23019, Nov 2020.

[5] Erik Bylow, Jürgen Sturm, Christian Kerl, Fredrik Kahl, and Daniel Cremers. Real-time camera tracking and 3d reconstruction using signed distance functions. In *Robotics: Science and Systems (RSS), Online Proceedings*, volume 9. Robotics: Science and Systems, 2013. Robotics: Science and Systems (RSS) Conference 2013 ; Conference date: 24-06-2013 Through 28-06-2013.

[6] S. Chen, X. Fang, J. Shen, L. Wang, and L. Shao. Single-image distance measurement by a smart mobile device. *IEEE Transactions on Cybernetics*, 47(12):4451–4462, 2017.

[7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *CoRR*, abs/1406.2283, 2014.

[8] J. Fields, G. Salgian, S. Samarasekera, and R. Kumar. Monocular structure from motion for near to long ranges. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1702–1709, 2009.

[9] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *CoRR*, abs/1609.03677, 2016.

[10] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. Generalizable pedestrian detection: The elephant in the room, 2020.

[11] Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, and Dieter Fox. Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. *The International Journal of Robotics Research*, 31(5):647–663, 2012.

[12] N. Jiang and Z. Jiang. Distance measurement from single image based on circles. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 1, pages I–809–I–812, 2007.

[13] Kusworo Adi and Catur Edi Widodo. Distance measurement with a stereo camera. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, 4(11), 2017.

[14] Yevhen Kuznietsov, Jörg Stückler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. *CoRR*, abs/1702.02706, 2017.

[15] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. *CoRR*, abs/1411.6387, 2014.

[16] Massimo Marchiori. Covid-19 and the social distancing paradox: dangers and solutions, 2020. `https://arxiv.org/abs/2005.12446`.

[17] Nitish Mutha. Equirectangular toolbox, 2017. `https://github.com/NitishMutha/equirectangular-toolbox`.

[18] Jane Parry. Covid-19: Hong kong scientists report first confirmed case of reinfection. *Bmj*, page m3340, 2020.

[19] Mogens Jin Pedersen and Nathan Favero. Social distancing during thecovid-19 pandemic: Who are the present and future noncompliers? *Public Administration Review*, 80(5):805–814, 2020.

[20] H. G. Lochana Prematunga and Anuja T Dharmaratne. Finding 3d positions from 2d images feasibility analysis. In *ICONS 2012 : The Seventh International Conference on Systems*, 2012.

[21] M. Ribo and M. Brandner. State of the art on vision-based structured light systems for 3d measurements. In *International Workshop on Robotic Sensors: Robotic and Sensor Environments, 2005.*, pages 2–6, 2005.

[22] Airam T. Z. R. Sausen, Maurício Campos, Paulo S. Sausen, Manuel O. Binelo, Marcia F. B. Binelo, João M. L. V. Silva, and Moises Santos. Classification of the social distance during the covid -19 pandemic from electricity consumption using artificial intelligence. *International Journal of Energy Research*, 45(6):8837–8847, 2021.

[23] Neha Shukla. A review on image based target distance & height estimation technique using laser pointer and single video camera for robot vision. In *International Journal of Engineering Research and Reviews*, volume 3, pages 128–135, 2015.

[24] Shane Tuohy, D. O'Cualain, Edward Jones, and Martin Glavin. Distance determination for an automobile environment using inverse perspective mapping in opencv. volume 2010, pages 100 – 105, 07 2010.

[25] S. Ullman. The interpretation of structure from motion. *Royal Society*, 203(1153), 1979.

[26] T. Van Dijk and G. De Croon. How do neural networks see depth in single images? In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2183–2191, 2019.

[27] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. *CoRR*, abs/1712.00175, 2017.

[28] Xu Qunyu, Ning Huansheng, and Chen Weishi. Video-based foreign object debris detection. In *2009 IEEE International Workshop on Imaging Systems and Techniques*, pages 119–122, 2009.

[29] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.

[30] Jing Zhu, Yi Fang, Husam Abu-Haimed, Kuo-Chin Lien, Dongdong Fu, and Junli Gu. Learning object-specific distance from a monocular image. *CoRR*, abs/1909.04182, 2019.