

Enhancing Design, Management, and Operation of Social Infrastructure Through Deep Learning Methods

by

Gabriel Draughon

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Civil Engineering)
in the University of Michigan
2023

Doctoral Committee:

Dean Jerome P. Lynch, Co-Chair, Duke University
Professor Cynthia Finelli, Co-Chair
Professor Robert Goodspeed
Professor Branko Kerkez
Professor SangHyun Lee

Gabriel T. S. Draughon

draughon@umich.edu

ORCID iD: [0000-0002-3557-6451](https://orcid.org/0000-0002-3557-6451)

© Gabriel T.S. Draughon 2023

Dedication

This dissertation is dedicated to my dearest Fiancée, Anna Stuhlmahcer, who I am deeply inspired by, and treasure most of all. Also, to my cats who are indifferent to most anything.

Acknowledgements

I would like to start by thanking my co-advisors, Dean Jerome Lynch, and Professor Cindy Finelli for their constant support and guidance. Dean Lynch was especially patient and flexible with me, allowing me time to find a topic I was passionate about. Together, I believe, we carved out a unique and deeply impactful dissertation topic that has been a joy to work on. Furthermore, he encouraged the branching of my work and supported my time spent outside the dissertation topic in engineering education – enabling me to find other passions and build a broader skillset. Professor Finelli was also key in my professional development, especially as an educator, providing numerous opportunities and connections in engineering education research. Under her guidance I grew in areas outside of my comfort zone and am a better teacher, engineer, and person. Working with her was a pleasure and I will always be grateful of her involvement. It is abundantly clear and evident my co-advisors were dedicated to my success in and outside of the program, their guidance, advice, and support went above and beyond any expectation and for that I am forever thankful.

I would also like to thank the rest of my dissertation committee, Professor Robert Goodspeed, Professor Branko Kerkez, and Professor SangHyun Lee. They bring a diverse range

of expertise and perspectives to the topic, and their feedback has only made this work stronger. Their time, effort, and flexibility has made my life easier.

The Detroit Riverfront Conservancy (DRFC) is heavily featured throughout this document. Their team and their spaces have made a significant impact on the dissertation. Working with Rachel Frierson, Mark Wallace, and Mac McCracken has been a pleasure and I look forward to any continued work with them and am thankful for all the opportunities they afforded this research.

Now I would like to thank family and friends, who were cornerstones to my success. My mother and stepfather, Jan and Kenneth Hammonds have been a part of my journey every step of the way, their unending support and love has carried me through difficult moments and cannot go unmentioned. Of course, my Fiancée, Anna Stuhlmacher, must also be thanked, for her support, encouragement, and guidance which was frequently needed and relied upon. Andrea Ventola's friendship has been a treasure and has provided much distraction as well as professional guidance. Yearly shenanigans from lifelong friend, Ryan Schrock, has helped keep the PhD a joy, and I am lucky to have him as a close friend. I would like to thank the LIST Lab, I carry immense pride coming from LIST. A big part of that is due to the impressive accomplishments of my lab mates and the culture they helped create. I also have pride for my involvement and connections to the Finelli lab which was a major piece of my final years. Aaron Appelle has been a wonderful travel companion, who is always up for an adventure or night out exploring new areas. Additionally, his expertise and feedback has helped shape parts of the dissertation. Patrick Sun, was a wonderful mentor and a better friend. Other LIST and Finelli lab

members, Katheirne Flanigan, Kidus Admassu, Wentao Wang, Rui Hoa, Hao Zhou, Omid Bahrami, Andrew Burton, Leah Marlor, Gracie Judge, Xiaping Li, Nolgie Oquendo-Colon, and Caroline Crockett provided fellowship, academic support, and inspiration along the way.

Table of Contents

Dedication	ii
Acknowledgements	iii
List of Tables	xi
List of Figures	xii
List of Appendices	xvi
Abstract	xvii
Chapter 1. Introduction	1
1.1 Goals & Objectives	3
1.2 Relevant Work	4
1.2.1 Sociability frameworks and taxonomies	4
1.2.2 Urban sensing in public open spaces	10
1.3 Sensing Framework	12
1.3.1 Privacy	15
1.4 Dissertation Organization	18
Chapter 2. Detecting, Tracking, and Mapping Human Use of Social Infrastructure Through Deep-Learning Methods	22
2.1 Methods	23
2.1.1 Object Detection	23
2.1.2 Object Mapping	32
2.1.3 Object Tracking	35

2.1.4 Secondary Classification.....	44
2.2 Performance Results	46
2.2.1 Object Detection	46
2.2.2 Object Tracking	52
2.3 Experiments and Field Deployment.....	55
2.3.1 Detroit Riverfront Camera Network	55
2.3.2 Activity Mapping.....	57
2.3.3 Continuous People Counting with Mobility Direction	61
2.4 Conclusion	67
2.4.1 Mask R-CNN vs YOLO	68
Chapter 3. Extended DeepSORT Framework	70
3.1 Background.....	73
3.1.1 DeepSORT and Multi-object tracking.....	73
3.2 Methods.....	75
3.2.1 DeepSORT Extensions	75
3.2.2 Mass Balancing.....	80
3.2.3 Tracker Challenges	82
3.3 Results.....	85
3.3.1 Association Recall Improvement.....	87
3.3.2 Association Precision Improvement	89
3.3.3 Evaluation of specific extensions.....	89
3.4 Conclusion	91

Chapter 4. Exploring Sociability in Public Open Spaces via in-depth Interviews with Park Managers and Operators	93
4.1 Goals and Objectives	93
4.2 Interview structure	96
4.3 Results.....	97
4.3.1 Background and importance of sociability	97
4.3.2 Exploring sociability in Public open spaces	98
4.3.3 Data on park use.....	103
4.3.4 Web dashboard.....	105
4.4 Conclusion	106
4.4.1 Revisiting interview objectives.....	106
Chapter 5. Real-Time Tracking and Visualizing of Sociability in Public Open Spaces	113
5.1 Methods.....	115
5.1.1 Sociability module	115
5.1.2 Real-time implementation and web-dashboard	125
5.2 Performance	128
5.2.1 Evaluation on live data.....	128
5.2.2 Evaluation on choreographed sequences	132
5.3 Experiments	134
5.3.1 Cullen Plaza	135
5.3.2 Freight Yard.....	139
5.3.3 Archer greenway	142
5.3.4 Valade Park.....	145

5.3.5 Real-time server	148
5.4 Conclusion	148
Chapter 6. Quantifying Sociability in Public Open Spaces	151
6.1 Methods.....	152
6.1.1 Activity Index	153
6.1.2 Social index.....	155
6.2 Experiments	156
6.2.1 Freight Yard.....	157
6.2.2 Valade Park.....	158
6.2.3 Cullen Plaza	160
6.3 Conclusion	163
6.3.1 Validation.....	165
Chapter 7. Discussion.....	166
7.1 Extensions of the framework	166
7.1.1 Methods.....	167
7.1.2 Results.....	171
7.2 Broader Impacts	174
7.2.1 Model bias and accessibility	175
7.3 Challenges and limitations	178
7.3.1 Sociability interviews and community feedback	178
7.4 Conclusion	179
Appendices.....	181

Bibliography200

List of Tables

Table 2.1 Object detector performance on OPOS validation set	51
Table 2.2 Tracking performance on custom video.....	55
Table 3.1 Evaluation of the tracking performance of various tracking-by-detection algorithms applied to three custom tracking challenges using HOTA-challenge Metrics and ID count.....	86
Table 3.2 Evaluation of the tracking performance of DeepSORT (with various static λ s) and ..	90
Table 5.1 SM performance results on footage from Cullen Plaza and Valade Park.....	129
Table 5.2 SM results on the controlled verification dataset.....	133
Table 7.1 WAR + COCO training results.	171
Table 7.2 WAR position sub-classifier training results.	171
Table 7.3 GPS Matching	173

List of Figures

Figure 1.1 Examples of public activities mapped on the optional/necessary spectrum.....	6
Figure 1.2 An overview of the sensing framework and its various components.	14
Figure 2.1 (A) Mask R-CNN with CNN + FPN as the backbone structure. (B) Feature Pyramid Network.....	26
Figure 2.2 Pinhole camera model, WCS, CCS, and PCS.	35
Figure 2.3 High level overview of the DeepSORT tracking-by-detection algorithm.....	36
Figure 2.4 DeepSORT tracking algorithm.	37
Figure 2.5 Face mask detection using secondary CNN (WideNet) classifier.....	46
Figure 2.6 Performance of Mask R-CNN detector with manually annotated ground truth images (A, C, E) with different weather conditions (sunny, cloudy, rainy) and detector results (B, D, F).	47
Figure 2.7 Examples at Dequindre Cut (Camera 87) of (A and B) original images with sparse and dense patron distributions; (C and D) processed images with segmentation; (E and F) processed images with 3d bounding boxes; (G and H) processed images with 2D maps.	49
Figure 2.8 (A) Detroit Riverfront parks camera locations; (B) Location of Camera 29 at the southeast corner of Cullen Plaza; (C) Camera calibration with checkerboard: and (D) Calibration reference points.	57
Figure 2.9 Activity classifications in the pixel coordinate system (A and C) and corresponding mapping in the world coordinate system (B and D).	59
Figure 2.10 Scatter maps and Density plots of detected users on a cloudy weekday (A, D), a sunny weekend (B, E), and total over a week (C, F). Workers trimming trees (G), patrons seeking shade (H), and patrons sightseeing along the fence (I).....	61
Figure 2.11 Examples at the Dequindre Cut (Camera 87): (A and B) pedestrian tracking and continuous counting with directions; (C and D) cyclist entering the field of view; (E and F) cyclist approaching two pedestrians and about to exit viewing; and (G and H) pedestrian exiting the field of view.	63

Figure 2.12 Daily pedestrian count and cyclist count from 11:00 to 13:00. (A) at the Dequindre Cut in fall 2020; (B) at the Dequindre cut in summer 2020; (C) at the riverfront in summer 2020. Grey segments represent overcast weather, segments with diagonal stripe pattern represent rainy weather.	66
Figure 2.13 Examples of object detection using YOLO framework trained on OPOS	69
Figure 3.1 Cosine difference between crops of the same person (n=16,000) and crops of different people (n=83,000).	78
Figure 3.2 (A) Detection error – detecting merged persons as a single person. (B) Precision tracking error due to poor illumination and distance from camera – matching two different persons to same ID.	80
Figure 3.3 Examples of tracklets that entered and exited (A), entered but was lost (B), appeared then exited (C), and appeared then lost (D).	82
Figure 3.4 Overview of tracklet balancing and stitching.	83
Figure 3.5 Scenes from the three tracker challenges. A) Mixed-use plaza. B) Pedestrian greenway. C) Riverfront walking path -sunset. D) Riverfront walking path – night.	84
Figure 3.6 Recall errors in the mixed-use plaza challenge. Using Default DeepSORT (A) and a dynamic lambda strategy with mean of the appearance feature gallery (B).	87
Figure 4.1 Affiliated organizations/parks and listed job duties of the fifteen respondents.	98
Figure 4.2 Goals and desired outputs of social programming and park assets.	99
Figure 4.3 List of activities respondents were given and asked to rank on a scale of 1-5 on how “sociable” they believe them to be.	101
Figure 4.4 Compiled responses for each item in figure 4.3 showing the mean and standard deviation.	102
Figure 4.5 Types of data on park users and usage patterns respondents would like to have.	104
Figure 5.1 Example of data output by the tracking module (with information on locations in the WCS from the mapping module).	115
Figure 5.2 Diagram of the Sociability Module and its three layers: movement, social, and location.	116
Figure 5.3 Movement layer decision tree for assigning dogs and strollers to their owners.	117
Figure 5.4 Social layer decision tree of the SM.	121
Figure 5.5 Examples of the SM socialization categories around a park bench.	122

Figure 5.6 Location demarcations of a scene (B) with mapping visualization showing individual locations in the WCS of a calibrated FoV (A).....	124
Figure 5.7 Example JSON encoded activity report for ID_1 at Valade Park on 8/1/22.....	126
Figure 5.8 Web dashboards for Campus Martius (A) and the DRFC park spaces (B).....	127
Figure 5.9 Example footage from Cullen Plaza (A), and Valade Park (B) used for evaluating SM performance.....	129
Figure 5.10 Examples of two observed fleeting interactions at Valade Park.	131
Figure 5.11 Two example choreographies from the controlled SM verification dataset. A-C are different frames from the same sequence, as well as D-F.	134
Figure 5.12 Heatmaps of social behavior at Cullen Plaza during July and August of 2022.....	136
Figure 5.13 Cullen Plaza social interactions by location during July and August 2022.....	137
Figure 5.14 Traffic and sociability statistics of weekday mornings at Cullen Plaza in August and July 2022.	138
Figure 5.15 Heatmaps of social behavior at Freight Yard entrances during July and August of 2022.....	139
Figure 5.16 Traffic and sociability statistics of Sunday afternoons at the Freight Yard in August and July 2022.....	140
Figure 5.17 Average time on site at the Freight Yard by socialization category.....	141
Figure 5.18 Location heatmaps of the socialization categories at Archer greenway and examples of independent field gatherings (A-C).....	142
Figure 5.19 Sociability stats at teh Archer greenway on days with and without field events. ..	143
Figure 5.20 Daily visitation by location at the Archer greenway during August-November of 2022.....	144
Figure 5.21 Sociability heatmaps at Valde Park during a summer Wednesday afternoon (A-D), and a summer Wednesday evening (E-H).	146
Figure 5.22 Visitation statistics for Wednesday mornings at Valade Park during June and July of 2022.	147
Figure 6.1 Scenes from the Freight Yard camera on a 7/24/22 (A), 7/17/22 (B), 7/22/22 (C), and 7/29/22 (D).	158
Figure 6.2 Traffic count (TC), Social index (SI), and Activity index (AI) of specific times in July of 2022.....	159

Figure 6.3 SI and SI examples on data from Valade Park in July of 2022.	160
Figure 6.4 Images from Cullen Plaza during. (A) 7/20/22 8:00am. (B) 7/11/22 8:00am. (C) 7/7/22 10:00 am. (D) 7/16/22 4:00 pm.	162
Figure 6.5 SI and SI examples on data from Cullen Plaza in July of 2022.	163
Figure 7.1 Examples from WAR image gallery.	167
Figure 7.2 (A) GPS heatmap of six subjects jogging through MCity street - using GPS estimates from mapping module. (B) GPS heatmap of two subjects patrolling MCity street and crossing paths – using GPS estimates from mapping module.	172

List of Appendices

Appendix A: Sociability Interview Protocol.....	182
Appendix B: Sociability Interview Results.....	191

Abstract

Social infrastructure (*e.g.*, public parks, squares, markets, etc.) plays a pivotal role in bolstering community quality of life, promoting economic prosperity, public health, and community resilience. Building on the work of Jane Jacobs, research has shown that walkable, mixed-use neighborhoods with a higher concentration of social gathering places and public space encourage the development of social capital and place attachment through an increase in social interaction. In short, the built environment and social organization of people are intimately connected. While autonomous sensing systems are ubiquitous in the modern world for monitoring and managing other forms of critical infrastructure, none yet exist for social infrastructure. Furthermore, despite its profound impact on the social health and resilience of the communities it supports, social infrastructure is often underfunded and underutilized.

While data surrounding social infrastructure use and how communities interact with and derive benefits from it would certainly improve design, management, and operation of social infrastructure and address issues surrounding funding and utilization such data is hard to come by. Autonomous solutions are too simple, yielding only traffic estimates of people and cyclists. In depth datasets are too expensive and tedious to collect relying on pen and paper approaches to manually observe and annotate how social infrastructure is being used.

This research aims to push a complete paradigm shift in how we design, manage, and operate social infrastructure. This document presents a first of its kind fully autonomous sensing system to detect, track, and map persons as they engage with social infrastructure all the while classifying their social behaviors and quantifying the sociability of the space. With this system, stakeholders (social infrastructure managers, and operators) can rigorously assess performance of their space and comprehensively analyze the impact of

investments, programming, and design choices on the social health and utilization of their spaces.

The sensing framework presented utilizes image streams from surveillance cameras and includes a multi-object detection component reinforced with a mapping module which projects pixel coordinates of detected individuals to a world coordinate system. Additionally, a tracking module, using data from the object detector and mapping module, traces individual trajectories and tracks time on site and interaction with park infrastructure. A sociability module then analyzes each trajectory and annotates the social interactions and behaviors observed for each tracked person, classifying their behaviors according to a novel schema. The social classification schema builds upon existing classification systems in urban sociology with modifications coming from insights gleaned from a series of in-depth interviews with critical stakeholders on the concept of sociability and how it relates to public spaces. From these interviews desired measurable outcomes and social behaviors were identified, leading to the development of activity and social indices'', calculable from outputs of the sociability module, to capture and quantify the performance of a space.

The sensing system was designed, developed, and implemented on a series of community parks along the Detroit Riverfront. These parks were chosen to serve as the primary research sites due to the diverse park social programming run in the spaces as well as the rich infrastructure (*e.g.*, food carts, carousels, playgrounds, etc.) featured throughout the spaces. The system generated detailed reports on space use and enhanced decision-making processes of critical stakeholders showcasing the impact it will have on community health and how shared spaces are designed and managed.

Chapter 1. Introduction

In [1], Whyte introduces the term “triangulation” which describes a phenomenon in which a stimulus brings together strangers in a public place and creates social bonds. The stimulus is often temporary and external to the space – such as street performers or musicians. However, the stimulus may also be a physical feature of the site, such as in Lonsdale Quay, Vancouver, where a baby blue piano sits on a pier often drawing small crowds as amateurs stop, play, and interact with the unique piece. The piano is a social catalyst, sparking conversations and interactions between strangers as they take pictures and play with the instrument. The piano is just one example of urban design and the built environment being intimately connected to the social organization and capital of the surrounding community.

The design, management, and operation of social infrastructure (e.g., parks, community centers, libraries, public markets, transit hubs) directly impacts the social sustainability and resilience of the communities they support [2]. Social sustainability refers to the community’s capacity to support the individual and collective well-being of its inhabitants. It’s a metric of the community’s viability, health, and ability to function. Someone concerned with a community’s social sustainability would seek answers to the question: How well do community members relate and interact with each other? How do they utilize their physical environment and organize to function as a community? Community resilience describes how inhabitants leverage community resources to adapt in an environment characterized by change, uncertainty, unpredictability, and surprise. As explained in [3], “members of resilient communities

intentionally develop personal and collective capacity that they engage to respond to and influence change, to sustain and renew the community, and to develop new trajectories for the communities' future". Socially sustainable resilient communities care for each other and their built environment in transformative ways, enabling them to sustain and recover from disaster. Social infrastructure significantly enhances a community's quality of life by offering space for recreation and social interaction while simultaneously reinforcing community resilience, economic prosperity, and improved public health. Social infrastructure advances community resilience and social sustainability by building the social capital essential for community members to organize and support one another.

By offering spaces to gather and organize, social infrastructure develops and strengthens social networks, increasing access to employment opportunities and financial resources [3], [4]. Furthermore, deeper social networks and ties to one's own neighborhood increase well-being, attachment to place, sense of community and belonging, and sense of security [5]. It is through these strong social networks that communities are able reduce disaster vulnerability factors [6], assist in disaster mitigation and response activities, and increase knowledge of resources that facilitate preparedness [7]. Additionally, social infrastructure improves physical health and fitness of the community, increasing access to recreational fields, equipment, and exercise programs [8]–[10].

Despite the numerous health benefits and impacts on community resilience and sustainability, social infrastructure is often underfunded and underutilized. Unfortunately, under investment disproportionately impacts communities living below the poverty line. In [7], the authors explored the deadly 1995 Chicago Heatwave, in which 739 people above the norm died – twice as many as in the infamous fire that ravaged Chicago in 1871. Research [7] into the deaths

showed that beyond lacking access to air conditioning (which increased risk of death by 80%) one of the biggest impacts on heat-wave mortality rates was the level of isolation of individuals. Poor but tight-knit communities such as Auburn Gresham fared better than most, having only 3 deaths per 100,000 despite not having access to air conditioning. In contrast, poor isolated communities that had been mostly abandoned such as Englewood suffered the worst with 33 deaths per 100,000. Furthermore, social infrastructure offers some of the most affordable forms of leisure and social gathering available, leaving poorer communities without access with fewer viable outlets than wealthier communities. Given its importance, it is imperative to advance scientific understanding of social infrastructure. Issues of underutilization can be addressed by advancing knowledge and understanding of how people utilize and derive benefits from social infrastructure. Furthermore, deeper understanding can help mitigate issues surrounding funding by increasing the impact of available funds, investing in furnishings and equipment which have the most impact.

1.1 Goals & Objectives

The overarching goal of this work is the advancement of cyber-physical-social systems (CPSS) to rigorously assess the performance of social infrastructure and to create multi-stakeholder frameworks based on data and computational models that can lead to resilient and equitable social infrastructure design. While smart cities and cyber-physical systems have introduced sensors that measure the performance of physical systems like transportation networks and physical infrastructure, there has been comparatively less emphasis placed on monitoring the social systems that use these physical systems and derive benefits from them. *Hence, this work aims to break new ground in leveraging computer vision and machine learning*

methods to quantitatively model user behaviors, including the social interactions these spaces support. By measuring and analyzing community sociability to inform and empower stakeholders (park managers, and operators), this sensing framework (see Section 1.3 ‘Sensing Framework’) aims to improve equitable access to the benefits derived from social infrastructure investments. The research utilizes community parks as the primary research site for deploying and testing our framework because they offer a rich set of unstructured social interactions to observe and study. However, the implications of the research on other social infrastructure types are explored by also considering a city center square. In brief, the key objectives of this research are as follows:

- Create a computer-vision based sensing framework to measure and track how people interact and move through public spaces
- Operationalize urban sociology frameworks to autonomously capture and infer social activities within public spaces
- Develop a calculable index to quantify sociability in public spaces and measure targeted positive outcomes of social programming and design interventions.
- Work with community partners in Detroit to implement and verify the framework

1.2 Relevant Work

1.2.1 Sociability frameworks and taxonomies

While I did not find any taxonomies or structures for explicitly assessing performance of social infrastructure, I found numerous frameworks for studying the public behavior within and how that links back to the quality of a space and the social capital and cohesion of the community. In general, I found two different “styles” or “focuses”. One sect focused on the

physical activities taking place, how people moved through a space, and how long they spent in the area. Other works focused heavily on documenting and classifying the social relationships supported by and taking place in the space. These frameworks are presented and discussed below.

1.2.1.1 Studying public life and activity

Danish architect, Jan Gehl, is perhaps one of the most influential voices in human centered urban design in the last 60 years. His large body of works [11]–[13] advocate cities designed around the pedestrian and cyclist instead of the car and focus on understanding the relationship between urban design and social wellbeing of the community. Gehl and his studio have produced frameworks and taxonomies, as well as numerous practical guides on how to systematically document and measure the influences of urban space on public life.

In *How to Study Public Life* Gehl and Svarre urge those studying public life, and specifically impacts on it from the built environment, to explore these questions: how many, who, where, what, and for how long? Counting traffic of pedestrians, cars, and cyclists and thus answering the question of how many, is one of the most basic tasks for observing public life. More specifically, by delineating how many people are passing through (pedestrian flow) versus how many are staying (stationary activity) researchers can better understand if design improvements are working. Are more people staying in the public square after this design change than before? And how does that compare when controlling for other factors such as weather, time of day, day of week, and season. The next question researchers should focus on, is who? Visitor demographics can reveal if the space is adequately serving different groups, are mobility device users frequently present in the scene? Children? Women? Or Families? Understanding at a basic level the demographics of users of the space enables researchers and planners to explore

their accommodations, if and where they are lacking, and how to better plan to be more accessible and better serving to the community.

Researchers should also explore the ‘where’ and observe how people are moving through space, where they are staying, and where they are partaking in certain activities. In doing so planners can better understand and plan the layout and positioning of furniture, gates, entrances, and other assets. From here, Gehl and Svarre suggest researchers extensively document what activities are taking place. Additionally, to get a better measure of the quality of the space activities should be mapped on a spectrum reflecting the necessity of said activity (Fig. 1.1). For instance, certain activities will happen in space regardless of the quality, such as those needing to pass through to commute to work. While other activities, such as leisurely strolling through the area, are much more likely to happen under good weather conditions and in well designed places.

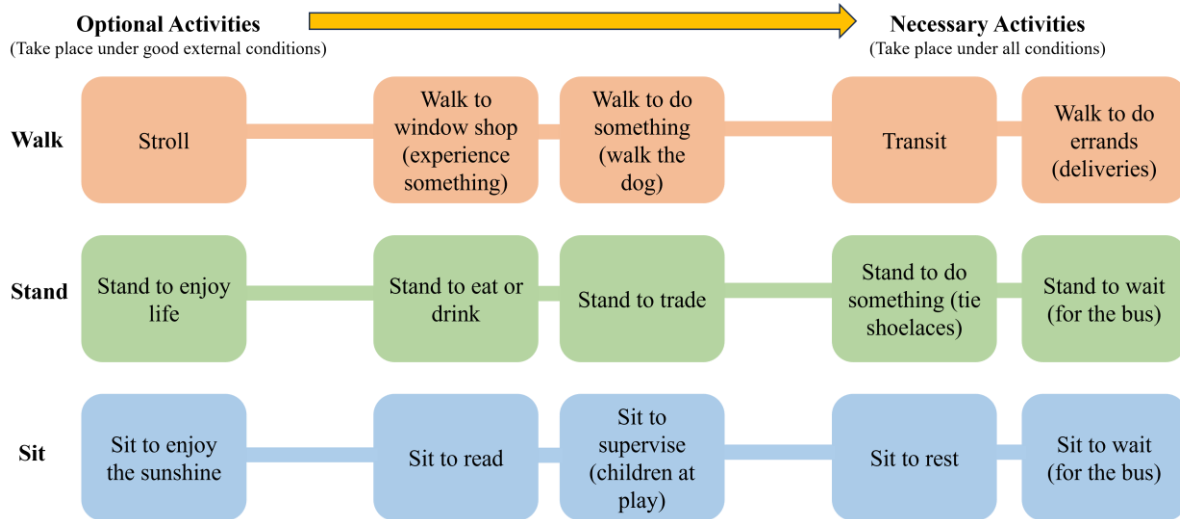


Figure 1.1 Examples of public activities mapped on the optional/necessary spectrum.

Lastly, researchers should observe the “how long” and record the amount of time people spend staying on site and partaking in various activities. A strong indicator of the quality of a space is how long people spend in it, even walking speed is expected to be a function of the quality of the space, people slow down and take their time enjoying the sights, sounds, and other features of high-quality spaces. Furthermore, understanding precise timings of how long people are willing to walk to get to transport, or how long they spend on certain activities are crucial to improving design and planning of public spaces.

All the above questions are important to study to get a better grasp of the quality of a space and how the built form impacts the function, use, and quality of a space. In *Life Between Buildings*, Gehl further refines his structure for documenting “the what”. Building on the optional to necessary spectrum, Gehl adds a “social” category. Using this taxonomy Gehl postures outdoor activities in public spaces can be divided into three categories: necessary activities, optional activities, and social activities, noting each of these categories place very different demands on the physical environment. Necessary activities are those that occur almost regardless of environmental quality: commuting to work or school or standing and waiting at a crosswalk would be considered necessary. Optional activities, such as going for a stroll or sunbathing, occur when environmental conditions are optimal. Social activities are dependent on the presence of others: conversations on a bench and children playing are examples of social activities observed in lively spaces. Each of these categories is influenced by the quality and design of the space. While necessary activities may happen regardless, Gehl notes the level of optional and social activities reflect the quality of the space and are given a chance to develop when investments are made to improve the design of the space.

1.2.1.2 Studying social relationships in the public realm

Since the mid twentieth century, researchers have shown the social organization of people and the built environment are intimately connected. In the 1960's, Jane Jacobs drove discussion around mixed land use, exploring the influence of the built environment on social networks [14]. Jacobs' early work argues that the mixture of commercial and residential space encourages walk-able spaces leading to the social integration of mixed races, incomes, and ages as they linger and interact in these multipurpose spaces.

These ideas are also present in New Urbanism, a neo-traditionalist design movement by urban planners and architects originating from the 1980's [15]. New Urbanists stress the importance of community and belonging to the health and wellbeing of residents, necessitating the consideration of how the built environment impacts residents' psychosocial sense of community [16]. How often community members interact with each other, deepen their social bonds, and engage spontaneously with unknown community members is directly related to a community's social capital, cohesion, and overall community resilience. In [17], the authors identified and listed the level of these interactions between community members and strangers as a key fundamental for neighborhood resilience after reviewing 22 academic papers from the health and community development literature.

In these works, we see the number of social interactions, with both known persons and strangers is linked to the quality of a space and resilience of a community. To systematically assess the performance of social infrastructure and measure it's impacts on the social health of the community we will need to understand and track the social interactions the space is supporting. Other researchers have developed taxonomies for classifying social relationships in the public realm and offer frameworks to build from.

In [18], Anderson describes four subtypes of stranger and categorical relationships which are important to understanding social life and interactions in the public realm: fleeting, routinized, quasi-primary, and intimate secondary. Fleeting relationships occur between people who are personally unknown to one another, stopping to ask a stranger the time or if a certain bus services a stop would be classified as a fleeting interaction. Routinized relationships describe interactions between persons that are categorically known to each other, examples would include interactions between vendors and purchasers, or patrolling officers and citizens. Quasi-primary relationships are relatively brief encounters lasting several minutes to hours between strangers - chatty dog owners at a dog park or talkative seat mates on a bus would be an example. Lastly, intimate secondary relationships are relatively long-lasting (weeks, months, or years) relationships between persons whose primary contact occurs in the public realm. Examples of intimate relationships include racetrack buddies or riders of a commuter bus who regularly sit and talk with each other.

One framework for classifying sociability in public streets [19] does build from the work of Gehl and Svarre while incorporating a higher focus on the social relationships. In [19], Mehta expands on Gehl's three categories by incorporating ideas presented in [18]. Mehta draws from [18] to add a relationship dimension to the categorization of public activities. Building on the classifications for human activities in public open spaces (POS) put forward in [11] and the classification of social relationships in the public realm in [18], Mehta develops an expanded taxonomy of human activities which fall under three social classifications (fleeting, enduring, and passive) to study and observe the sociability of city streets. *Fleeting* is similar to the type described in [18] and describes short interactions between strangers. Friends and family

engaging in a space together would be classified as *enduring* and those that are alone, either silently observing or partaking in a solitary activity in the space would be classified as *passive*.

The frameworks presented offer great insights into the necessary dimensions to capture when studying public life and the impact of urban form. While no above framework will be explicitly chosen for this research, ideas from all will be explored. A final structure, which builds from and leverages the existing bodies of work, to study public life and assess the performance of the social infrastructure is presented in the chapters which follow.

1.2.2 Urban sensing in public open spaces

Various taxonomies and classification schemas have been used to annotate observations of public life in POS (Section 1.2.1), however most studies on sociability use the traditional pen and paper approach. Traditionally, researchers sit in a space, manually recording and annotating activities and behaviors observed in a journal. Alternatively, researchers may review camera footage and record observations without being physically present.

Whyte [20] was one of the first researchers to use cameras to record how people use open public spaces such as streets. By positioning cameras on rooftops and in elevated windows, he was able to study how people moved and socially interacted on city streets, informing new design concepts for people centric public spaces.

Surveys and interviews have also been used to study sociability and the impact of urban form and design. In [21], residents in a two-block radius of a public square were surveyed before and after community led renovations to measure the impacts of the new design on the community's wellbeing, sense of community, and use of the space. A mixed-methods case study in Dehli, India [22] explored resident's social interactions and participation in neighborhood activities and how they were influenced by the neighborhood urban form and design of the POS.

The study looked at three separate sites across two suburbs. In two sites the residences were set around POS, giving residents quick access to diverse spaces to gather. The third site faced paved roads and lacked easily accessible POS. The researchers used numerical surveys as well as open-ended interviews to assess the differences in the three communities, their overall wellbeing, pride and attachment to place, and sense of community and belonging. Additionally, the researchers supplemented the data with manual observations on the communities' social behaviors and use of the surrounding POS. While these methods gather detailed insights and data on the social behaviors of communities and how they relate to urban form they require extensive time investment of researchers to conduct.

The cost and tedium of such manual methods prevent them from being widely used in a variety of social infrastructure contexts, such as studies tracking changes in social behaviors over prolonged periods of time or vast spaces.

For these reasons researchers have developed a variety of automated solutions for person tracking and counting. Industry available pedestrian counting systems utilize discrete sensors (passive infrared, geophones) placed along pathways to give counts of people passing through. When distributed across a space, counts of pedestrians (and their direction traveled) can be estimated. Additionally, readily-available software has been developed for use on cameras which generates foot traffic and direction estimates through blob detection and centroid tracking, by drawing a line in the frame and counting every time a detected object's centroid passes through the line. However, to be effective the camera angle needs to be aimed directly downwards, limiting the field of view and other uses for the camera. In these methods only counts and direction are estimated. Similar approaches [23], [24] install cameras along doorways to count the people entering and exiting, but suffer the same downsides. Other methods utilize smart-phones [25]

and WiFi signals [26] to count and localize pedestrians carrying a smart phone. Additionally, commercially available passive infrared sensors (*e.g.*, ECO Counters) are commonly deployed in urban areas and parks to count every time a patron crosses through. However, with recent advances in deep-learning based computer-vision models for object detection, more complex and capable methods for object tracking have been developed and implemented and will be discussed in the research approach. While advancements in urban sensing technologies have led to robust autonomous person counting and tracking tools, they lack social dimensions, yielding only counts and traffic patterns of visiting patrons leaving still a need for an autonomous, easily deployable framework for measuring and tracking social behaviors and interactions in POS.

1.3 Sensing Framework

Machine learning methods such as deep neural networks are at the center of the sensing framework (Figure 1.2) which consists of multiple modules; a high-level diagram is shown in figure 1.2 with references to the chapters where each module is discussed. First, a convolutional neural network(CNN)-based **object detector** ingests images from a camera feed and identifies relevant objects in the scene, such as people, dogs, cars, bikes, and scooters. From there a multi-**object tracking** module utilizes a Kalman filter, CNN, and mass-balancing heuristic to track objects as they move through the scene and keeps track of their time on site. An **object mapping** module is used to project pixel coordinates of tracked individuals to a world coordinate system, allowing interactions with furniture and various park assets to be easily recorded. Next, a **sociability module** analyzes trajectories of each individual to infer and classify social relationships and behaviors, enabling the capture of spontaneous interactions, number of users with family/partners, and other social trends observed in the space. Lastly, a set of **social and**

activity indices' purposefully built to capture desired outcomes of park programming and pro-social behaviors, give a measure of the park's current performance. The indices' enable park managers and operators to quickly assess the impacts of various design interventions and park programming on patron use and behaviors. The **sociability module** and **social and activity indices**' were heavily informed and shaped by a series of in-depth **interviews** with critical stakeholders on the concept of sociability (in the context of POS) and how it relates to social behaviors and park outcomes.

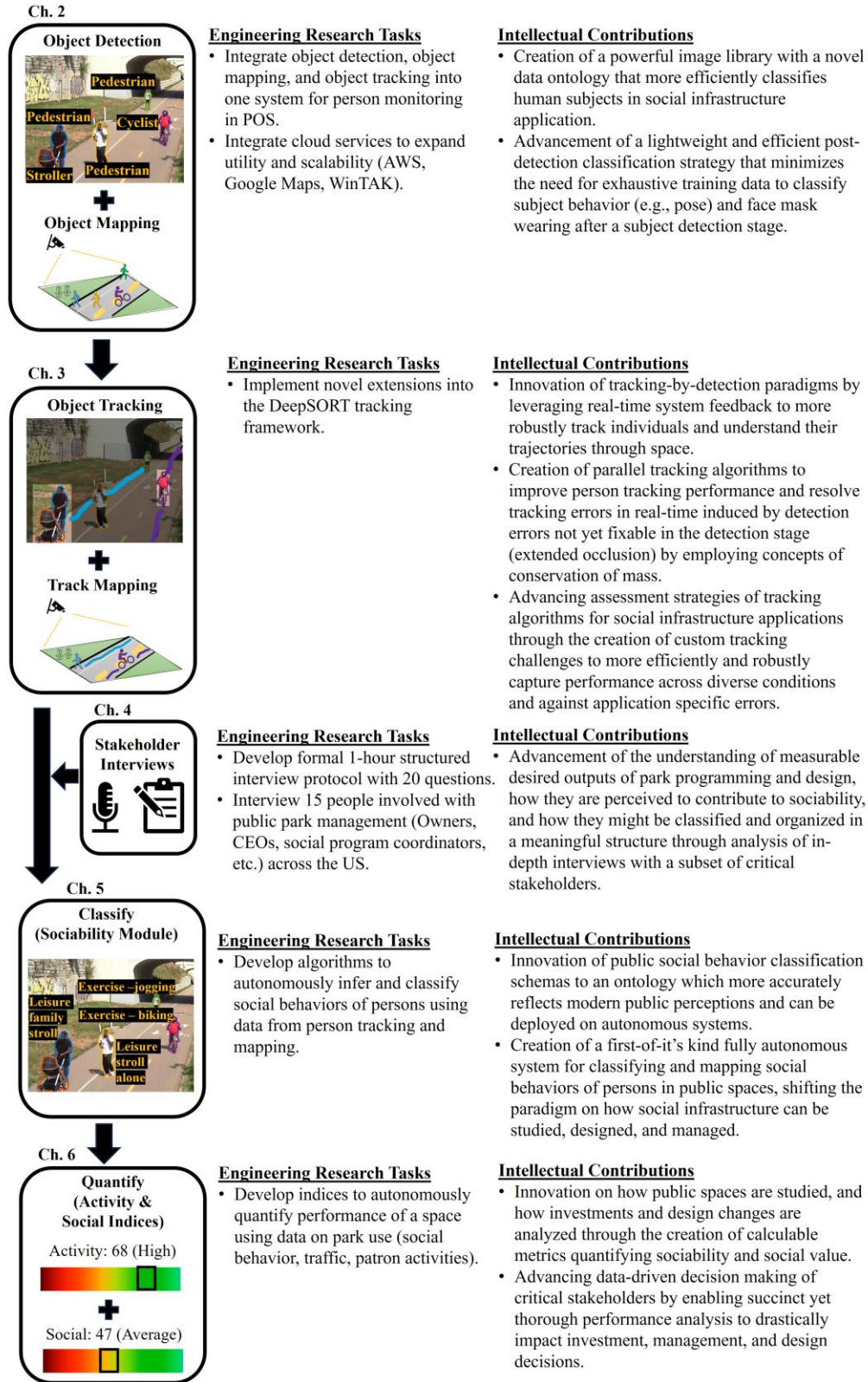


Figure 1.2 An overview of the sensing framework and its various components.

1.3.1 Privacy

One of the biggest concerns surrounding the framework is privacy. Given the sensing framework is built using image feeds from cameras, it is important to address community concerns and attitudes around surveillance. Additionally, it is important to explicitly discuss how the framework handles privacy and sensitive data.

First, it is important to acknowledge surveillance in general is a controversial topic. While most would agree to an assumption of privacy within the confines of personal property, people have various responses to surveillance in public spaces. It is unsurprising to find people's levels of comfort and attitudes towards surveillance in public spaces is closely connected with the purpose of the surveillance and type of data being collected [27], [28]. In, [28], 164 individuals took part in a 10-day study in which they were shown examples of different forms of video surveillance and purposes in public and where polled on their levels of comfort for each scenario as well as their level of surprise on the type of data being collected in its uses. Additionally, participants were asked if they would prefer to be notified whenever such data was being collected on them and if they would allow or deny it if given the option. Various examples included anonymous people counting, facial recognition to replace ID cards, and anonymous face detection to estimate a customer's race to offer tailored deals and coupons. Anonymous methods with more generic purposes such as people counting were more widely accepted with 13% answering they were very uncomfortable and 33% answering they were somewhat uncomfortable, which is much fewer than the 33% and 35% that felt very uncomfortable and somewhat uncomfortable with targeted ads based on inferred race and gender and the 32% and 39% that felt very uncomfortable and somewhat uncomfortable with facial recognition for ID verification. However, that is still 46% of the respondents who are at least somewhat

uncomfortable with surveillance for person counting and traffic monitoring purposes. Even with anonymous data collection practices and less intrusive purposes surveillance is a sensitive issue in public spaces and should be heavily considered when designing and deploying such sensing systems.

1.3.1.1 Privacy and data policies of the framework

Edge Computing

While images of specific applications and demos of the sensing framework are shared through this document, it is important to note the sensing framework does not store any images or video feeds. During the development and testing, images and videos were saved for research purposes - to visually check performance and showcase the toolchains capabilities. However, when deployed the framework does not store any images or video feeds. By performing all computations on site, or on “the edge”, no images are transferred outside of the closed surveillance network. By doing so, all collected data is anonymous, no data on a track’s movement, social interactions, and activities is tied back to an image. Each track is identified with a number alone with no indication of age, gender, or ethnicity.

The framework is able to generate anonymous results as it was designed to operate in real-time. No footage needs to ever be stored or recorded, instead at the end of each day the framework generates a daily activity report and updates the social and activity numbers on the web-dashboard.

Additionally, the framework is designed to ingest image feeds from already deployed surveillance cameras. The framework does not need high-resolution images nor highly tailored angles to accurately collect data. During this research, the framework was only deployed in areas

where surveillance cameras existed and the decision on cameras was already made and accepted in the community.

Digital Transparency

The author believes in transparency and building digital trust. The author believes a good practice for deploying the sensing framework is to involve the community as much as possible and be transparent with the data collection and privacy policies. I was fortunate enough to work with pre-existing camera infrastructure in communities where the decision to have cameras was already made and their presence is made clear to the park patrons, however that will not always be the case and so these conversations and considerations are paramount to have when presenting this type of work.

With the addition of our sensing technology our partners, the Detroit Riverfront Conservancy, are making moves to increase the transparency of all data collection and further involve the community. As this research continues, we are working with Helpful Places, a social impact enterprise, that helps ensure smart city solutions are transparent, inclusive, participatory, and in alignment with community goals. One way of doing so is placing signs and symbols around the park places where data is being collected which informs park users specially the type of data being collected, how it is being used, and how it is being stored. Impactful research is being done in this area to work on and develop universally recognized symbols which can be easily recognized and understood throughout the world as smart city solutions and data collection is becoming more rampant. Other practices include tabling near data collection sites and surveying patrons, or distributing QR codes around the park which poll users on their thoughts and beliefs about potential data collection practices before they are implemented.

1.3.1.2 Final statement on privacy

This framework was deployed and tested on a series of pre-existing surveillance cameras in Detroit, Michigan. All images saved and shown throughout this document have been cleared for research demonstration purposes under the IRB that governs the research. These saved images, as stated earlier, are only for demonstration purposes and to manually assess early iterations of the framework. The final deployed product does not store any video or image of any kind. Furthermore, per our IRB, we are not allowed to perform any form of facial recognition. The presented sensing framework does not need facial recognition capability and I recommend any similar researchers in the area to follow suit.

The community should have a voice in the decision to deploy surveillance systems or not. Some areas with greater trust in the management and parks are willing and comfortable for their data to be collected. However, some spaces without established community trust may do more harm than good installing surveillance systems. Attitudes around public surveillance and impacts of surveillance on distrusting communities is outside the scope of this research, but nevertheless should be addressed when discussing challenges and limitations of the framework.

1.4 Dissertation Organization

This dissertation consists of seven chapters. Chapter 2 focuses on the foundations of the sensing framework, introducing the deep learning methods used to sense, map, and track people. Chapter 3 dives into multi-object tracking and presents novel tracking algorithms to extend the performance of person tracking when faced with challenges such as occlusion. Chapters 5 and 6 focus on quantifying the sociability of public spaces leveraging insights from stakeholder interviews captured in chapter 4. Chapter 7 focuses on the flexibility of the framework and its

extensions to other applications while also concluding with a discussion on the toolset and its broader impacts.

In **Chapter 2**, we explore state of the art CNN based object detectors for person detection, as well as existing frameworks for object tracking and mapping. Additionally, novel approaches to add secondary classifications of detected persons such as their pose and if they are wearing a facemask are presented. The tools presented are also tested and verified on footage from partnered spaces along the Detroit Riverfront.

In **Chapter 3**, we address challenges in person tracking, namely those from occlusion, with novel additions to the DeepSORT tracking algorithms. We take a dynamic approach, choosing to tune the influence of appearance vs motion information in our association calculations based on real-time information from the image feed. Furthermore, we introduce a novel mass-balancing heuristic to stitch together broken trajectories of tracked persons in real-time. Our novel tracker is tested against other state-of-the-art tracking algorithms on three custom tracking challenges, representative of various camera angles and population densities common for our application.

In **Chapter 4**, we momentarily step away from algorithm writing to further explore the concept of sociability within the context of public open spaces. Through a series of formal structured interviews with various public space professionals we seek to connect sociability to measurable targeted outcomes and goals of park spaces and social programs. Additionally, we explore how specific activities and behaviors impact a space's perceived sociability and how best they might be classified. A breakdown of the interview results are presented with details and analysis of respondent answers for each question. Insights gleaned from the interviews are used

to inform the sociability classification schema presented in chapter 5 and the social and activity indices’ presented in chapter 6.

In **Chapter 5**, building off the insights from chapter 4, we introduce a novel sociability module to our sensing framework. The module allows for the toolchain to infer and classify social interactions. The classification schema builds on existing taxonomies from urban sociology informing users on the social relationships and behaviors of park patrons. The sociability module is tested and verified on manually annotated footage across various spaces along the Detroit Riverfront parks. Additionally, in this chapter, cloud infrastructure, powered through amazon web services (AWS), is implemented to enable real-time data visualizations and analytics. For experimentation, the sociability module and real-time framework is deployed on four different cameras in the Detroit Riverfront park spaces during the summer and fall of 2022 to analyze patron social behaviors and patterns at times of day and during organized social programming.

In **Chapter 6**, we introduce two novel indices’, the sociability index and activity index, which utilize data from our sensing framework to quantify sociability data and provide simpler metrics to quickly analyze performance of park spaces and various social programs. In chapter 4, through a series of in-depth interviews with various park managers and operators throughout the US we identified various measurable targeted outcomes and goals of park spaces and social programs. The indices’ presented in this chapter seek to capture these outcomes and measure performance. The indices’ were used to measure performance of three different park spaces which hosted various social programs throughout 2022-2023.

In **Chapter 7**, we conclude with discussions on broader impacts as well as other relevant issues and limitations of the sensing framework – such as privacy concerns, transparency, and

building digital trust with the communities where our tools may be deployed. We briefly address the history of racial and gender bias in deep learning-based image processing algorithms as well as discuss other potential forms of bias – such as detection of individuals with mobility equipment. Additionally, we move away from sensing patron use patterns and social behaviors in public open spaces and explore the capabilities of the toolset in other domains. More specifically, we utilize the toolchain for human health and performance monitoring in urban warfare settings. We look to merge visual data from images with biometric sensing data of warfighters to add context and provide more complex warfighter health monitoring.

Chapter 2. Detecting, Tracking, and Mapping Human Use of Social Infrastructure Through Deep-Learning Methods

This chapter focuses on the detection, mapping, and early tracking modules of the sensing framework (Figure 1.2). The algorithms presented in chapters 3, 5, and 6 are entirely novel and heavily build upon the framework established in chapter 2. However, before we can move into the innovations of the next chapters we must describe the foundational components of the sensing framework, which leverage existing computer-vision tools to accomplish object detection, mapping, and initial tracking (tracking is significantly improved in chapter 3). In this chapter I extensively test object detection, tracking, and mapping methods and select the best options for the specific application of human monitoring. I then integrate these tools into a unified platform to track and map persons as they move through POS and run experiments on real image feeds at Detroit Riverfront parks. To enhance performance, I manually curate custom training datasets. Additionally, I introduce a novel secondary real-time classifier to add complexity to object classifications (*e.g.*, are they wearing a face mask?) without needing large training data.

With Section 2.1: Methods, I provide background information on the available tools and describe in detail the algorithms and models chosen for my specific application space. In Section

2.2, I present custom datasets and tracking challenges to evaluate the accuracy and performance of the chosen detection models and tracking algorithm. And in Section 2.3 Experiments, I leverage partnerships with the Detroit Riverfront Conservancy (DRFC) to test and verify these modules on image feeds from their camera network and provide insights and visualizations on patron usage patterns and behaviors. The chapter also presents a novel secondary classification stage (Section 2.1.4) to provide further context to detected persons (such as their pose and if they are wearing a face mask).

For object detection, two state-of-the-art models are rigorously tested and deployed on DRFC camera streams. I compare and contrast these two models in the conclusion section, which also provides additional commentary on the modules and their limitations.

2.1 Methods

2.1.1 Object Detection

The first computational module for the proposed sensing framework (Figure 1.2) is a detector trained to automate the processing of camera images to yield segmentation of people with their activities labelled. While other human activity mapping strategies exist (Sec. 1.2.2), I primarily explored computer vision-based toolchains due to their increased accuracy and capabilities. Object detection strategies answer the question: Given an image; taken from a surveillance camera in the park, what relevant objects can be found in the image (person, car, bicycle, dog, stroller, etc.)?

Various computer vision methods exist for object detection. While CNN architectures are the most popular choice for object detection today, there are other well-known methods within the computer vision field. More traditional methods such as the Viola-Joins detector [29],

which is popular with face detection, utilize handcrafted feature extractors (kernels) such as Haar or histogram of oriented gradients (HOG) [30]. These features are then passed into a cascaded detection algorithm [29] or into a State Vector Machine (SVM) [30] for object classification.

However, advances in deep-learning have led to a sharp rise in CNN-based architectures for object detection. When given sufficient training data, CNNs are able to learn kernels for feature extraction. Furthermore, CNN architectures enable multilayered feature extraction by learning kernels which convolve over feature maps to produce richer and deeper feature maps, yielding high degrees of object detection accuracy. Popular CNN architectures can be described as either single stage [31], [32] or two-stage detectors [33], [34]. My research explored the use of both single-stage and two-stage detectors to find the optimum architecture for the application.

2.1.1.1 Two-Stage Detectors

Region-based convolutional neural network (**R-CNN**) [35] was one of the first two-stage object detection models - it adopted selective search methods based on a region proposal tool (*i.e.*, a selective search process that generates a large number of bounding box proposals). In R-CNN, an image is first passed to the region proposal module, which selectively searches the image to identify 2,000 bounding box candidates that may contain an object for detection. The bounding box candidates are each reshaped into a fixed square and fed to a deep CNN that is trained to first extract a 4,096-dimensional feature vector that can be used by a support vector machine (SVM) to classify an object (if any) in the candidate bounding box.

To overcome the limitations of R-CNN including the fixed nature of the region proposal stage and the high computation overhead of processing 2,000 bounding box candidates through a CNN, **Fast R-CNN** [36] was proposed. Fast R-CNN speeds the original R-CNN framework using a single CNN to first extract a global feature map from which regions of interest (RoIs) are

identified by selective search. A technique called RoI pooling takes the local RoI feature vectors extracted from the CNN output to simultaneously determine the object classification (by using a softmax classifier layer) and optimal bounding box parameters (by regression). The Fast R-CNN has a smaller inference time than the original R-CNN because it only requires the execution of a single CNN. When compared to R-CNN implemented with the same CNN backbone (*i.e.*, VGG-16 [37]) and identical computational hardware, Fast R-CNN increases the inference speed 15-fold.

To address the computational bottleneck of selective search, **Faster R-CNN** [34] was proposed. Faster R-CNN is similar to Fast R-CNN but uses a regional proposal network (RPN) to do selective search for identification of regions of interest. The RPN is a small neural network that slides over the global feature map to classify objects and to perform a regression that produces the optimal bounding box that fits the object. The elegance of the RPN is that it is trained as part of the training of the Faster R-CNN framework. RPN reduces the inference time of the detector to a point where it can be used in real time [if the video speed is less than 10 frames per second (fps)]. Specifically, Faster R-CNN has an inference speed 10 times faster than Fast R-CNN.

Mask R-CNN [33] is the newest member of the two-stage object detector family. Mask R-CNN also uses an RPN prior to detecting objects in a second stage but offers improved object detection capabilities by outputting an instance segmentation with improved detection precision. In my research, I explored the use of Mask R-CNN because of these improved performance features, especially the segmentation of detected people.

2.1.1.2 Mask R-CNN

The Mask R-CNN framework is shown in Figure 2.1(A). It begins with the use of a CNN to process input images to extract a rich set of global feature maps based on a feature pyramid network (FPN) architecture that permits features to be defined on different spatial scales (Figure 2.1(B)). The input images used in this study consist of 1,280 by 720 pixels, with each pixel

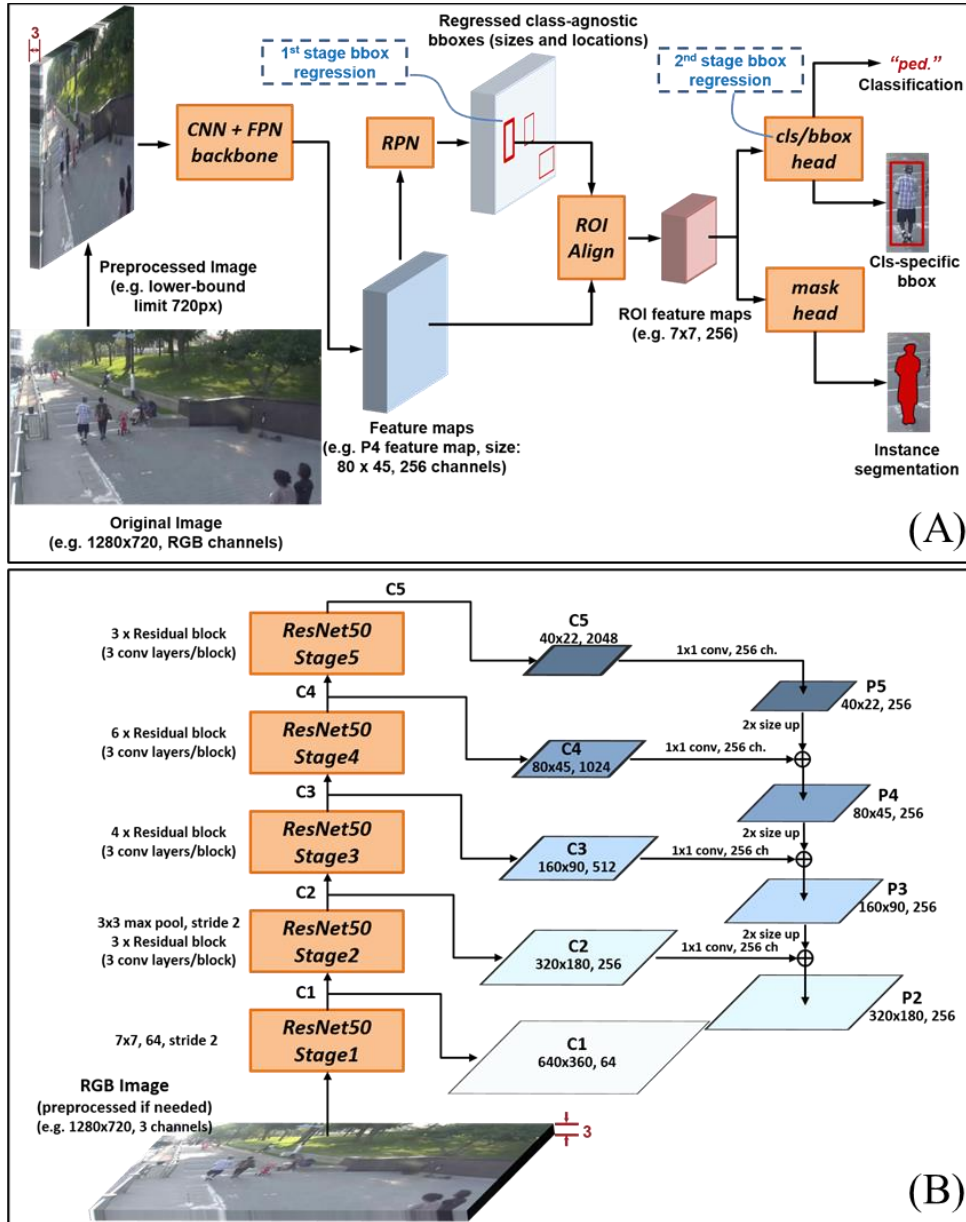


Figure 2.1 (A) Mask R-CNN with CNN + FPN as the backbone structure. (B) Feature Pyramid Network.

having RGB values (i.e., three integers). During training or testing, any image of arbitrary resolution was rescaled through linear interpolation to the standard $1,280 \times 720$ pixel resolution before passing through the CNN pipeline. The image was then fed into a CNN + FPN backbone to generate a set of global feature maps. Multiple predefined CNN backbones have been explored for image processing applications including AlexNet [38], VGGNet [37], and ResNet [39]. For my research, I adopted the CNN backbone termed ResNet with 50 layers (i.e., ResNet50) due to the use of residual blocks that alleviate the issue of overfitting of the network during training. A residual block is a group of CNN layers that include a shortcut (identity) connection that feeds forward the input vector x to combine with the residual mapping, $F(x)$ to yield $F(x) + x$ (He et al. 2016). I chose ResNet50 over shallower (e.g., ResNet18) or deeper (e.g., ResNet101) ResNet backbones due to the balance achieved between inference accuracy (which comes from more layers) and speed (which comes with fewer layers).

ResNet50 is partitioned into five stages as shown in Fig. 2.1(B). Stage 1 takes the original image and processes it through one convolutional layer with a 7×7 , 64-channel filter with a stride of 2 pixels, effectively outputting a $640 \times 360 \times 64$ feature map defined as C1. Stage 2 takes C1 as the input and applies a 3×3 maximum pooling operator (maxpool) with stride 2. The output of the maxpool is then passed through three residual blocks, each consisting of three convolutional layers with varying filter sizes to produce Feature Map C2 with dimensions $320 \times 180 \times 256$. Stages 3–5 similarly apply varying numbers of residual blocks with different convolutional filter sizes to output Feature Maps C3–C5 with sizes $160 \times 90 \times 512$, $80 \times 45 \times 1,024$, and $40 \times 22 \times 2,048$, respectively. The global feature maps C1–C5 contain semantic and spatial information from the original image at varying spatial scales.

An FPN [40] is added to the CNN framework to gather information from the differently scaled feature maps (C1–C5) in a top-down strategy. As shown in Fig. 2.1 (B), the P5 feature map was processed from C5 using a 1×1 convolutional layer to form a 256-channel-deep feature map. P4 comes from the element-wise combination of a scaled-up version of P5 (by a factor of 2 in both width and length) and a shortened version of C4 in depth (shortened from 1,024 to 256 channels). Likewise, P3 and P2 come from the element-wise combination of a two-times scaled-up P4 and P3, respectively, and a depth-shortened C3 and C2, respectively.

P4 is selected as the input into an RPN designed to detect objects based on a sliding window applied to the P4 feature map. A 3×3 convolutional layer was applied to the P4 feature map with the output fed into two analytical branches: one branch consists of a 1×1 convolutional layer fed to a sigmoidal function to detect if an object is present (outputting an objectness score); the other branch uses a 1×1 convolutional layer to predict class-agnostic bounding boxes. In this second branch, a total of nine bounding boxes are produced combining three different areas (128×128 pixel², 256×256 pixel², and 512×512 pixel²) with three different width-to-height aspect ratios (1:1, 1:2, and 2:1) at each pixel on the feature map. Using P4, a total of 32,000 bounding box candidates were generated but the candidate list was pruned by removing candidates with either a high amount of overlap with another bounding box [using a 70% overlap threshold for nonmaximum suppression (NMS)] or based on portions of the bounding box falling outside the feature map perimeter. Using the objectness score associated with each bounding box, the top candidates were retained during training (e.g., 2,000 bounding box) and testing (e.g., 1,000 bounding box).

The RoI Align method is used to extract a regional feature map from P4 for each class-agnostic bounding box output by the RPN. In my study, the regional feature maps were reshaped

into a defined size of 7×7 with 256 channels before being input simultaneously to two functional heads. First, the regional feature maps produced for each bounding box candidate were fed to a fully connected neural network whose output was passed to two branches, one representing a softmax classifier (assigning the object to a class) and a bounding box location regressor (outputting an optimal bounding box for the object). This is referred to as the class/bounding box head (Fig. 2.1(A)) outputting a class-specific bounding box. The second, termed the mask head, takes the bounding box candidates from the RoI Align step and inputs the candidates into a series of 3×3 convolutional layers followed by a linear rectifier function (i.e., ReLU) to yield a vector representing segmentation of the objects in the feature map space. Finally, deconvolution is used to up-sample the vector into the original image space, representing a segmented mask of the detected object.

2.1.1.3 Single Stage YOLO Architecture

You only look once (YOLO) [31] is a single-stage architecture widely adopted for computer vision tasks due to its inference speed and high degree of accuracy. Two-stage detectors may be slightly more accurate, but they are slower and more computationally expensive. For these reasons, I also explored the YOLO framework for the task of object detection. The YOLO architecture has seen several changes and improvements [41], [42].

My work specifically uses YOLOv3 [42], a single CNN which segments images into an $S \times S$ grid of cells and predicts bounding boxes and class probabilities for each cell simultaneously. Each cell uses three anchor boxes centered in the cell to make three predictions. Anchor boxes, or bounding box priors, are bounding boxes with predefined heights and aspect ratios. The sizes of the bounding box priors were chosen by the YOLO authors based on a cluster analysis of the bounding boxes present in the COCO [43] dataset. Each prediction consists of the objectness score, bounding

box coordinates, and class probabilities for each class used in model training. The objectness score determines the confidence of an object present in the box, which should be 1 if an object is present and 0 otherwise. The bounding box coordinates are the offsets of the bounding box of the object in relation to the bounding box prior. In a CNN, each convolutional layer extracts semantically richer feature maps, but at the cost of resolution. As a result, large CNNs may have trouble detecting small objects, to combat this, YOLOv3 adopts ideas from the feature pyramid networks [40] featured in Mask R-CNN and uses earlier feature maps in the CNN to predict additional bounding boxes at two other scales.

2.1.1.4 OPOS Training Dataset

In order to learn, CNN architectures utilize loss functions and large amounts of pre-annotated training data. There are a vast number of open-source annotated image libraries available to use for training [43], [44]. COCO, common objects in context, is one of the most popular datasets containing over 200,000 annotated images labeling objects from a set of 80 categories (i.e, person, phone, fire-hydrant, chair, etc.).

To optimize performance for our intended application, low-resolution and distant surveillance cameras in parks, we developed the Objects in Public Open Spaces (OPOS) dataset [45] during this research. OPOS is a private dataset owned and curated by the Laboratory for Intelligent System Technology (LIST) and designed specifically for human detection in public open spaces using a dense network of security cameras along the Detroit Riverfront Conservancy (DRFC) parks. The library is not open to the public to ensure the privacy of the community using the DRFC spaces. OPOS consists of 7,826 fully annotated images collected from surveillance footage from different seasons and various lighting conditions. All the footage used for the data set are from the years 2018 and 2019. The detection data set includes object instances under

various weather conditions (e.g., sunny, cloudy, and rainy) with 6.7% of instances occurring during rainy weather. Furthermore, the data set includes object instances during different phases of the day (i.e., morning, afternoon, and evening) with different sizes and from various cameras with different fields of view. The diversity in season, weather, time of day, and lighting was intentional and when used for training yields more robust models. The OPOS taxonomy includes four super categories (i.e., people, vehicles, accessories, animals) grouping 11 object types (i.e., pedestrian, cyclist, sitter, dog). With a total of 18,000 class instances manually segmented from DRFC images, the library is one of the largest specialized for people detection in public spaces.

2.1.1.5 Mask R-CNN and YOLO training

Loss functions - metrics on a network's error between prediction and ground truth used to find optimal solutions which minimize error - are needed for CNNs to learn kernels and candidate weights for the model.

The loss function used in YOLO training is comprised of three parts: *localization loss*, *confidence loss*, and *class loss*. The localization loss utilizes the sum of squared errors function and computes an error on the bounding box coordinates of the predicted versus ground truth. The confidence loss is broken into two components, *objectness loss* which evaluates objectness scores on predicted and ground truth, and *no-objectness loss* which evaluates no-objectness scores on predicted and ground truth. Combined, these two components, measures how well the model predicts the presence or absence of objects. Both the localization and confidences loss are computed using the cross-entropy loss function. Lastly, *class loss* is a measure of the discrepancy between the predicted class probabilities and the ground truth class labels. It is calculated using cross-entropy loss.

The loss function used in training for Mask R-CNN is comprised of three parts: *RPN loss*, *R-CNN loss*, and *Mask loss*. *RPN loss* has two parts: *anchor classification loss* and *anchor regression loss*. Anchor classifications determine whether an anchor box contains an object or is a part of the background, *anchor classification loss* measures the discrepancy between predicted anchor classifications and ground truth labels and uses the binary cross-entropy loss function. *Anchor regression loss*, is a measure of the error between coordinates of the predicted anchor box and ground truth and uses the mean square error function.

R-CNN loss, similarly, has two parts: *bounding box regression loss*, which evaluates error between predicted bounding box coordinates and ground truth using the mean square error function, and *classification loss* which measures the difference between predicted class probabilities and ground truth class labels using the cross-entropy loss function. Lastly, *Mask loss* evaluates the errors of the segmentation, computing differences in the pixel-wise segmentation mask and ground truth masks using the binary cross-entropy loss function.

I used pre-trained weights, instead of random initialization, on both object detectors before training on the OPOS dataset. The Mask R-CNN architecture was pretrained with the weights of the ResNet50 backbone pretrained using the general-purpose ImageNet-1K data set [44] and the weights of Stages 3–5 of ResNet50, and RPN and RoI Align were pretrained using the COCO_2017 data set [43]. YOLO was also pretrained using weights from ImageNet.

2.1.2 Object Mapping

With object detection complete (i.e., people detected and their activities classified by Mask R-CNN or YOLO), the next step in the computational framework is object mapping (i.e., mapping people into a 3D coordinate system using a pinhole camera model for monocular cameras [46]). In this model, a point in the 3D world coordinate system (WCS) defined as $\{X, Y,$

Z } can be projected onto the image plane of the camera, which is defined by the two-dimensional (2D) pixel coordinate system (PCS) $\{u, v\}$ as shown in Figure 2.2.

Given the intrinsic properties of the camera (i.e., focal length f_x and f_y , and principal point location of the camera lens, c_x and c_y) and identification of the camera coordinate system (CCS) $\{X_c, Y_c, Z_c\}$ relative to the WCS, the projection of a point in the WCS to the PCS can be achieved through a perspective transformation

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (2.1)$$

$$s\mathbf{m} = \mathbf{A}[\mathbf{R}|\mathbf{t}]\mathbf{M}$$

where s is the scaling factor converting pixels to length units; \mathbf{A} is the camera intrinsic matrix that includes intrinsic camera properties; and \mathbf{R} and \mathbf{t} relate the WCS and CCS through a rotation matrix and translation vector, respectively. The vectors \mathbf{m} and \mathbf{M} are referred to as homogenous coordinates of the PCS and WCS, respectively. Finally, the CCS and the WCS are related by

$$\mathbf{M}\mathbf{c} = \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = [\mathbf{R}|\mathbf{t}]\mathbf{M} \quad (2.2)$$

where $s = Z_c$. The coordinates $\{X_c, Y_c, Z_c\}$ in the CCS can also be normalized by s as $\{X_c', Y_c', 1\}$. Hence, given s , \mathbf{A} , \mathbf{R} , and \mathbf{t} , equations 2.1 and 2.2 can be used to easily convert m from an image to \mathbf{M} and \mathbf{M}_c .

While in theory the mapping of the PCS to the WCS is analytically straightforward, image distortion introduced by the camera lens must first be accounted for [46]. Consider the distorted location point in the PCS as u' and v' . If the camera distortion is not severe, radial

distortion (defined by coefficients k_1 , k_2 , and k_3) and tangential distortion (defined by coefficients p_1 and p_2) can be utilized to correct the images by the Brown-Conrady model [47]

$$\begin{aligned} u' &= (1 + k_1 r^2 + k_2 r^4 + k_3 r^6)u + 2p_1 uv + p_2(r^2 + 2u^2) \\ v' &= (1 + k_1 r^2 + k_2 r^4 + k_3 r^6)v + 2p_2 uv + p_1(r^2 + 2v^2) \end{aligned} \tag{2.3}$$

where $r^2 = (u - u_c)^2 + (v - v_c)^2$; and u_c and v_c are coordinates of the distortion center. Often the distortion center is assumed to be zero.

To spatially map the people detected by the Mask R-CNN (or YOLO) detector, each monocular camera must be calibrated. For our purposes I adopted the calibration process proposed by Zhang in [47] and implemented in OpenCV. The calibration method aims to estimate the intrinsic matrix (\mathbf{A}) and distortion coefficients (k_1, k_2, k_3, p_1, p_2) of the camera as well as the extrinsic properties (\mathbf{R} and \mathbf{t}) through a process using a large checkerboard with known dimensions. The large black-and-white checkerboard has a 10 x 7 check grid with each grid square measuring 5 cm (2in) x 5 cm (2in) and is moved at various angles and distances through the camera's field of view. The intrinsic matrix, extrinsic matrix, and distortion coefficients of a camera may be stored after following the steps in [47].

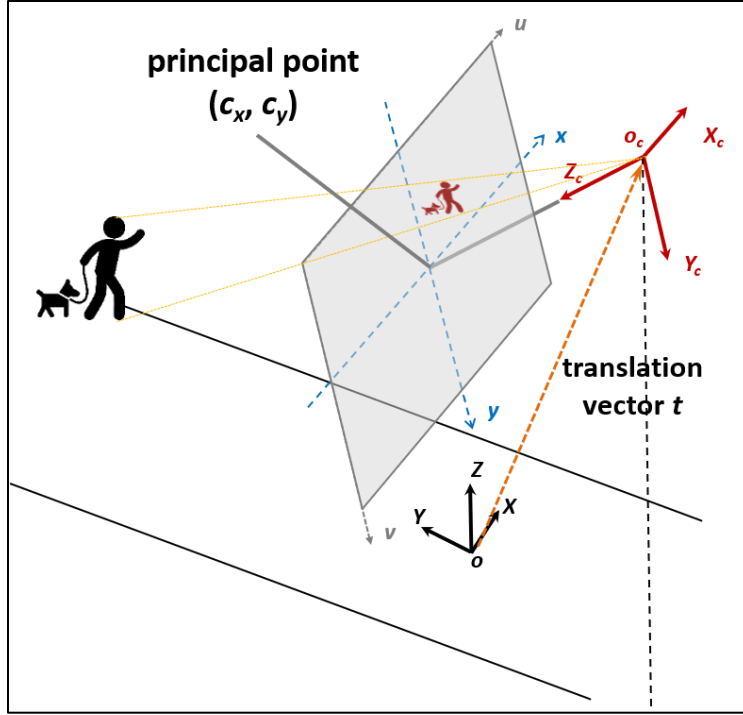


Figure 2.2 Pinhole camera model, WCS, CCS, and PCS.

Once a camera is calibrated 3D bounding boxes for detected persons can be generated. To estimate a 3D bounding box, the lowermost pixel (i.e., $\{u_{bot}, v_{bot}\}$) of the segmentation of the detected person is assumed to be connected to the ground (which was defined as $Z=0$ in the WCS) while the topmost pixel (i.e., $\{u_{top}, v_{top}\}$) is assumed to be associated with the top of the head. Given equation 2.1, the location of the person on a map can be extracted to yield $(\{X_{bot}, Y_{bot}\})$ in the WCS with $Z_{bot} = 0$. The area of the 3D bounding box is assumed to be 60 by 60 cm for pedestrians and 60 by 160 cm for cyclists; these dimensions were conservatively set based on the average shoulder width of adults being 40 cm [48]. The height of the 3D bounding box is then found by using $\{u_{top}, v_{top}\}$ to determine the height of the detected person.

2.1.3 Object Tracking

I adopted a tracking-by-detection paradigm to build a person tracking and counting module. Tracking is fundamentally a complex reidentification (Re-ID) problem that aims to

reidentify objects in sequential image frames to track them over time. Here, I adopted a tracking algorithm based on a modification of the DeepSORT algorithm [49] to build the tracking module using spatially mapped detected objects. The tracking problem is formatted as an assignment problem with objects detected in a given image either assigned to existing tracks (based on the tracking results from previous frames) or assigned to a new track.

The proposed detection module relies on the Hungarian algorithm [50] to solve this combinatorial optimization assignment problem. The tracking algorithm utilizes both motion information (based on Kalman filtering) and object appearance to define data association metrics to be used by the Hungarian algorithm as shown in Figure 2.3. I adopted DeepSORT because of its robust tracking performance (especially under the condition of a high number of object occlusions) compared with other real-time tracking methods previously proposed in the computer vision field [e.g., SORT [51]].

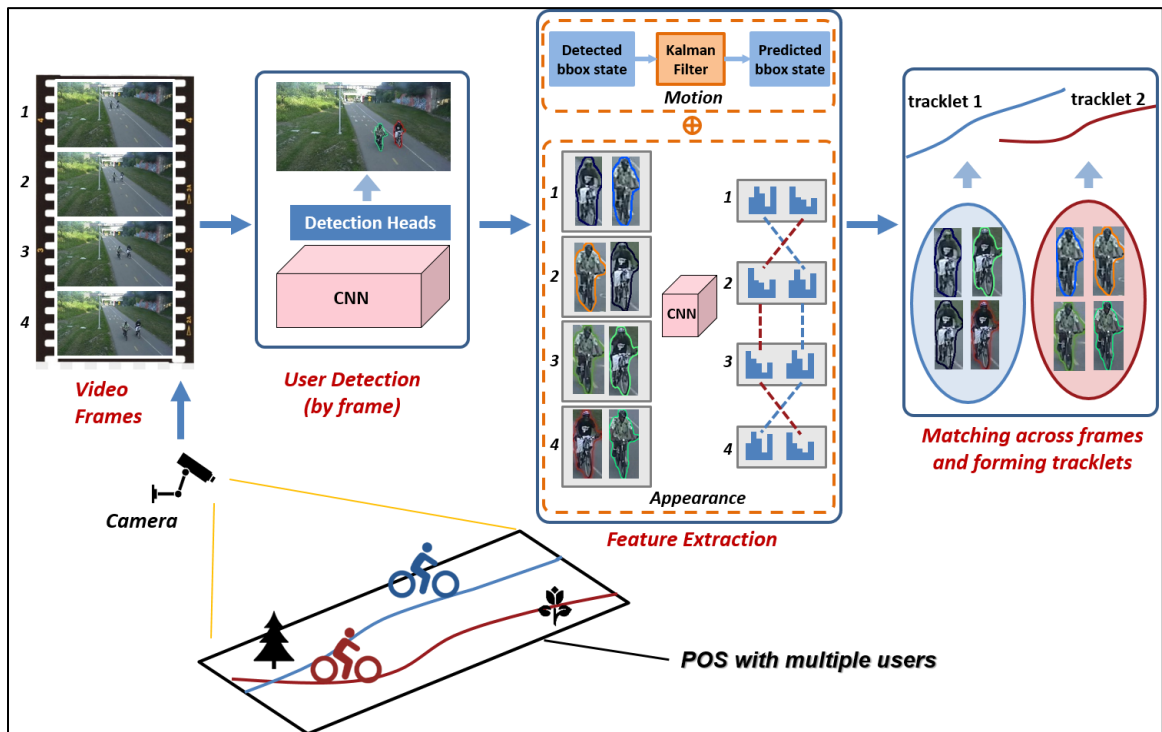


Figure 2.3 High level overview of the DeepSORT tracking-by-detection algorithm.

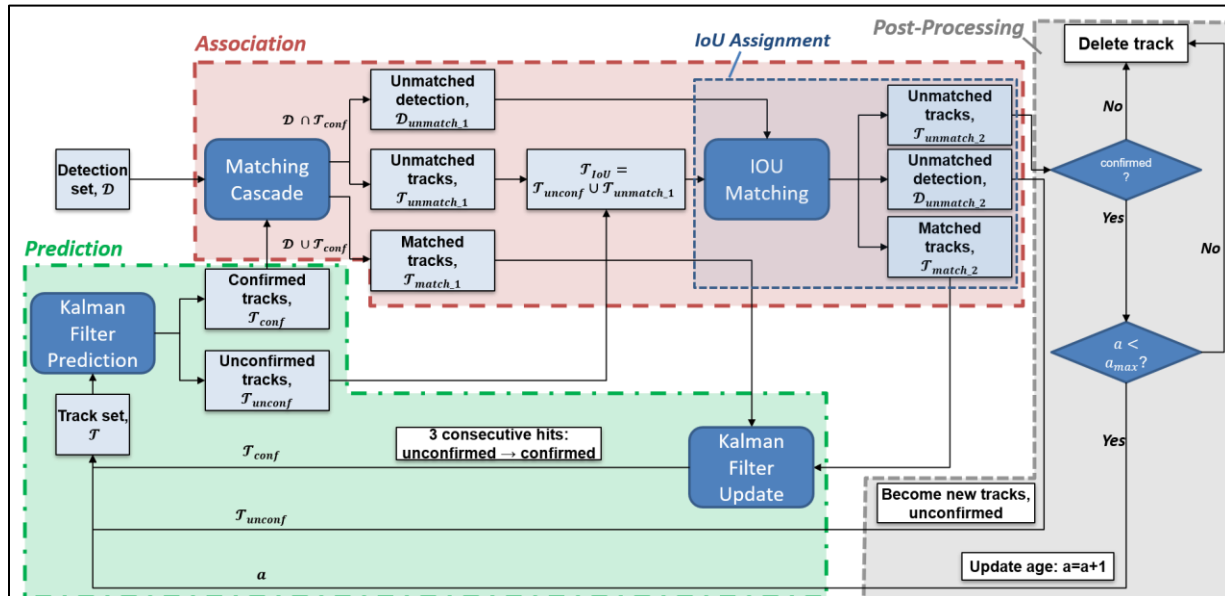


Figure 2.4 DeepSORT tracking algorithm.

An overview of the tracking-by-detection process is provided in Figure 2.3, using four sequential camera frames and two detected bikers on separate trajectories. The first step is people detection by the Mask R-CNN detector and their spatial mappings using the pinhole camera model. As shown in Figure 2.3, the two bikers are detected in each camera frame.

Then, given the bounding box of the detected people, a set of association metrics are extracted from each to enable tracking utilizing both motion and appearance information. Motion association is based on the location of object j detected in frame k to the location of the same object predicted by a Kalman filter using the trajectory of that object collected from prior ($<k$) frames. This stage is referred to as the prediction stage in Figure 2.4. The association metric based on appearance is based on a feature vector extracted from the segmented person object. Together, the motion and appearance association metrics are used to assign the detected object to an existing trajectory (called a tracklet) or used to create a new trajectory (e.g., if it is the first time the object is detected); this stage is referred to as the association stage in Figure 2.4. The

last stage of the three-stage algorithm is the postprocessing stage where existing and new trajectories are managed before advancing the to the $K + 1$ frame.

The prediction stage is primarily based on the use of the Kalman filter (Kalman 1960) to model and predict the motion of the bounding box of a detected person. An observation of a detected person derived from the Mask R-CNN (or YOLO) detector is a vector, \mathbf{Z}_k , of the detected bounding box attributes including the location of the feet of the detected person (x_c and y_c), aspect ratio (a), and height (h), in frame k : $\mathbf{Z}_k = \{x_c, y_c, a, h\}^T$. In this stage, the observation vector, \mathbf{Z}_k , is compared to predictions of the observation vector obtained from a Kalman filter associated with the tracklet set.

To model the motion of the bounding box as a dynamic system, a state vector, \mathbf{X}_k , is defined using the bounding box attributes and their first-order rate of change: $\mathbf{X}_k = \{x_c, y_c, a, h, \dot{x}_c, \dot{y}_c, \dot{a}, \dot{h}\}^T$. The evolution of the bounding box state in discrete time is:

$$\mathbf{X}_k = A\mathbf{X}_{k-1} + w_k \quad (2.4)$$

where $w_k \in \mathbb{R}^{8 \times 1}$ is the process noise defined by a zero mean Gaussian model with covariance matrix $Q_k \in \mathbb{R}^{8 \times 8}$, and $A \in \mathbb{R}^{8 \times 8}$ is the system matrix:

$$A = \begin{bmatrix} I & \Delta t I \\ 0 & I \end{bmatrix} \quad (2.5)$$

where $I \in \mathbb{R}^{4 \times 4}$ is an identity matrix; and Δt is the time interval between the process iterations.

The observation $Z_k \in \mathbb{R}^{4 \times 1}$ is related to the state by $\mathbf{Z}_k = \mathbf{C}\mathbf{X}_k = \mathbf{v}_k$, where $\mathbf{C} \in \mathbb{R}^{4 \times 8}$ is the observation matrix $\mathbf{C} = [\mathbf{I} \ \mathbf{0}]$, and \mathbf{v}_k is the measurement noise defined by a zero-mean Gaussian model with covariance matrix $\mathbf{R}_k \in \mathbb{R}^{4 \times 4}$. Finally, the stochastic nature of the state process is defined by the covariance matrix $\mathbf{P}_k \in \mathbb{R}^{8 \times 8}$.

The prediction of \mathbf{X}_k and \mathbf{P}_k based on the process up to time step $k-1$ without considering measurements at k are defined as $\tilde{X}_k \in \mathbb{R}^{8 \times 1}$ and $\tilde{P}_k \in \mathbb{R}^{8 \times 1}$, respectively

$$\begin{aligned}\tilde{X}_k &= AX_{k-1} \\ \tilde{P}_k &= AP_{k-1}A^T + Q_k\end{aligned}\tag{2.6, 2.7}$$

The Kalman filter determines a Kalman gain matrix $K_k \in \mathbb{R}^{8 \times 4}$ that uses the measured observation of the state \mathbf{Z}_k to update the state (\hat{X}_k) and covariance (\hat{P}_k) matrix predictions

$$\begin{aligned}\hat{X}_k &= \tilde{X}_k + K_k(Z_k - C\tilde{X}_k) \\ \hat{P}_k &= (I - K_kC)\tilde{P}_k\end{aligned}\tag{2.8, 2.9}$$

The optimal Kalman gain that provides the best state and covariance predictions given the measurement noise process is

$$K_k = \tilde{P}_k C^T (C \tilde{P}_k C^T + R_k)^{-1}\tag{2.10}$$

Once the state and covariance matrices have been updated, \mathbf{X}_k and \mathbf{P}_k are set as $X_k = \hat{X}_k$ and $P_k = \hat{P}_k$ to be used in the next time step, $k+1$.

The next stage of the tracking algorithm (Figure 2.4) is to associate detected objects with predictions using the existing set of trajectories termed a tracklet set, T . At frame k , the set of Mask R-CNN/YOLO detected bounding boxes constitute the detection set D . For each tracklet in T , the Kalman filter is used to predict the locations of the tracklets' bounding box at frame k . An association metric, d_{IoU} , is defined based on the j th detected person in D and is compared to the predicted person using the i th tracklet in T .

$$d_{IoU}(i, j) = 1 - IoU_{assign}(i, j)\tag{2.11}$$

Where $\text{IoU}_{\text{assign}}(i,j)$ is the intersection over union (IoU) in the assignment stage measuring the overlap extent of a detected bounding box from the detector k and a predicted bounding box from the Kalman filter

$$\text{IoU}_{\text{assign}}(i,j) = \frac{A_{tra}^i \cap A_{det}^j}{A_{tra}^i \cup A_{det}^j} \quad (2.12)$$

where A_{det}^j is the area of the j th detection bounding box; and A_{tra}^i is the area of the bounding box from the i th Kalman filter prediction.

An additional motion-based association metric based on the probabilistic feature of the Kalman filter prediction is also defined to measure the distance between the j th bounding box detection and the i th bounding box prediction. Parameter d_{mot} is defined using the detected person observation \mathbf{Z}_{kj} to that projected into the measurement space by the Kalman filter based on trajectory i , $\hat{\mathbf{Z}}_{ki}$

$$d_{mot}(i,j) = (\mathbf{Z}_{kj} - \hat{\mathbf{Z}}_{ki})^T \hat{\mathbf{P}}_{ki}^{-1} (\mathbf{Z}_{kj} - \hat{\mathbf{Z}}_{ki}) \quad (2.13)$$

Using motion-based association metrics like d_{IoU} and d_{mot} may result in poor tracking performance when detected objects are occluded by other objects (Wojke et al. 2017). Hence, an appearance-based association metric, d_{app} , that is less sensitive to occlusions was introduced. The metric d_{app} is based on the difference in appearance of the j th detection and the appearance features from the i th trajectory. A small convolutional neural network was proposed by [52] and trained using an open-source Re-ID image library [53] using a 128-term feature vector \mathbf{r}_j extracted from the detected bounding box of the j th person. The appearance descriptor, \mathbf{r}_j , is scaled to have a unit norm (i.e., $\|\mathbf{r}_j\| = 1$). A gallery \mathbf{R}_i of past p appearance descriptors from the

set of bounding boxes associated with the i th tracklet is formed: $R_i = \{r_l^{(i)}\}_{l=1}^p$. The appearance association metric then measures the smallest cosine distance between the gallery of appearance descriptors of trajectory i and that of detection j

$$d_{app}(i, j) = \min\{1 - r_j^T r_l^{(i)} \mid r_l^{(i)} \in R_i\} \quad (2.14)$$

The tracking module utilizes the three defined association metrics to create the trajectory set T over time in a frame-by-frame fashion. The tracking set will be divided into two sets of tracks: T_{conf} corresponding to confirmed tracklets and T_{unconf} which are unconfirmed tracklets. Once an unconfirmed tracklet has three or more detected objects assigned, it is converted to a confirmed tracklet. Each tracklet in T also has an associated age, a , that defines the number of past sequential frames over which no detected object has been assigned.

In the prediction stage, a Kalman filter prediction of where each tracklet is in the k frame is identified providing the predicted observation \hat{Z}_{ki} and correlation matrix \hat{P}_{ki} . In the association stage, matching the j th detected person to an i th tracklet is performed in two steps. First, a cascaded matching process is designed to give priority to those tracklets that have most recently had detected objects appended. To provide this priority, tracklets with an age greater than a_{max} are not considered further. This allows the matching stage to be computationally efficient in searching for matches supporting the real-time execution of the tracking-by-detection framework. In the matching cascade, motion association metric d_{mot} is used as a hard gating parameter for matching; if $d_{mot}(i, j) > (d_{mot})_{max}$, then the j th detection will not be considered for pairing with the i th track because they are physically too far apart to justify further evaluation. For the remaining confirmed tracklets, the appearance-based association matrix is used to make the first round of assignment of detected objects tracklets. A cascade depth index q is defined for

the matching cascade, where q will be varied from 1 to a_{\max} . Given q , a subset of T_{conf} consisting of tracklets whose $a = q$ is formed, T_q , consisting of n tracklets. Similarly, a set of detected objects of D not yet assigned to a tracklet is formed, D_q , consisting of m objects. A cost matrix, $\mathbf{C}_q \in \mathbb{R}^{n \times m}$ is assembled using $d_{\text{app}}(i,j)$. The Hungarian method [50] is used to solve the unbalanced assignment problem using the cost matrix, \mathbf{C}_q . If \mathbf{C}_q is not a square matrix (e.g., there are uneven detections and trackers), \mathbf{C}_q is changed into a square matrix by adding zero rows or columns. The detected objects and tracklets matched at q are removed from further consideration to form a new set of unassigned detected objects and tracklets for the next step $q + 1$ (i.e., D_{q+1} and T_{q+1} , respectively). The matching cascade continues until $q = a_{\max}$. At this point, there may be unmatched objects, D_{unmatch_1} , and unmatched tracklets, T_{unmatch_1} .

The second step of the matching process is conducted by IoU matching using d_{IoU} . In order to perform the next round of matching based on d_{IoU} , a new tracklet set, $T_{\text{IoU}} = T_{\text{unmatch}_1} \cup T_{\text{unconf}}$, and new detection set, $D_{\text{IoU}} = D_{\text{unmatch}_1}$, are formed. Prior to performing the next round of assignment based on IoU, a maximum threshold is established on d_{IoU} as $(d_{\text{IoU}})_{\max}$. Detection-tracklet pairs where $d_{\text{IoU}} > (d_{\text{IoU}})_{\max}$ will not be considered further. The Hungarian method is used to make assignments between the n' detections in d_{IoU} and the m' tracklets in T_{IoU} using a cost matrix, \mathbf{C}_{IoU} , based on d_{IoU} with the Hungarian method applied to make matching assignments. Again, some objects and tracklets may not be paired, resulting in unmatched detection set D_{unmatch_2} and unmatched track set T_{unmatch_2} .

The last stage of the assignment problem is to decide how to treat the unassigned detections (D_{unmatch_2}) and tracklets (T_{unmatch_2}) at frame k . First, the unassigned detections (D_{unmatch_2}) are simply added to the unconfirmed tracklet set, T_{unconf} , for use in the next frame. Second, if a tracklet in T_{unmatch_2} is a member of T_{unconf} , then it is deleted from T prior to the next

frame. If the tracklet in $T_{unmatch_2}$ is a member of T_{conf} , then the age of the tracklet is checked. If $a \geq a_{max}$, then the tracklet is removed from T prior to the next frame. The final step of the postprocessing stage is to move tracklets in T_{unconf} with three consecutive object assignments to T_{conf} .

In our application, the detection set, D , is established based on classified objects with a confidence score of 0.75 or greater. The memory of the tracker-detection process is set as 6s (i.e., a_{max} is six times the frame rate); tracklets that have not been matched for $a > a_{max}$ are not considered for the next round of matching. The calculation of the three metrics, d_{mot} , d_{app} , and d_{IoU} , is performed on a CPU (rather than on a GPU) with their computational time negligible. The average computational time of the tracking-by-detection framework was 0.0015s for importing a video frame, 0.1245s for detection inference, and 0.0014s for tracker-detection association and tracker updating, all on a computer with an Intel Core i7-10700F CPU (Intel, Santa Clara, California) and NVIDIA GeForce RTX-2070S GPU. The threshold of the parameters of the tracker (e.g., thresholds $(d_{mot})_{max}$, $(d_{app})_{max}$, and $(d_{IoU})_{max}$) were determined based on the method proposed by [49], [54] with the threshold for the appearance metric set as $(d_{app})_{max} = 0.40$ and the threshold for the motion metric set as $(d_{mot})_{max} = 9.49$. These thresholds were later confirmed to offer a high level of tracking performance. The cost matrix (\mathbf{C}) used for each time of assignment was mainly built based on $\mathbf{C}(i,j) = d_{app}(i,j)$ except for two scenarios: if $d_{mot}(i,j) > (d_{mot})_{max}$, then $\mathbf{C}(i,j)$ was set as 1×10^5 , making a match impossible for the (i,j) pair; if $d_{app} > (d_{app})_{max}$, then $\mathbf{C}(i,j)$ is set to 0.40, making the match slightly more probable even though the appearances distance between tracklet and detected object are relatively large. The threshold for the IoU association metric is set as $(d_{IoU})_{max} = 0.70$.

2.1.4 Secondary Classification

2.1.4.1 Face mask detection

Certain applications require further classification of identified objects. For example, during the Covid-19 pandemic, park managers wished to know the face mask usage rates of its patrons so they could know where to best distribute signs and reminders of the face mask policy. One approach to solve this issue would be to gather and label examples of pedestrians wearing a face mask, add the examples to training data, and retrain the model. However, this approach has multiple challenges. First, this approach requires the curation of large amounts of training examples, and to perform well as a standalone class the model would need thousands of example instances. Additionally, the new class ('pedestrian with mask') would be extremely similar to pedestrians without masks and would be difficult for the model to accurately distinguish the two, especially considering the small number of pixels detailing face masks.

Another solution would be to add face masks as a detectable class, then in post processing assign detected pedestrians as wearing face masks if there is a large overlap between the face mask's bounding box and the pedestrian's bounding box. However, this approach has similar challenges – it would require thousands of training examples and would not perform well given the size of face masks, approximately 100-400 pixels².

However, these challenges can be simplified or avoided by instead running a secondary classification over existing detections. In this approach, Mask R-CNN/YOLO handles person detection and will forward the cropped images of detections to the face mask detection module. Therefore, the face mask detector only needs to solve a binary classification problem: is the detected person wearing a face mask or not? The need for a larger training set is reduced by

simplifying the problem to binary classification – which is a much easier task than object detection.

Recent work [55] & [56], shows CNN-based detectors are able to detect facial occlusions with impressive accuracy when tested on face mask datasets such as MAFA. The datasets used for training and validation contain clear, high resolution images of faces. Unfortunately, CNN-based face mask detectors that have been trained on higher resolution images with close-up views of faces will struggle when applied to lower resolution crops of faces taken from distant cameras. To address this issue, I manually curated a low-resolution face mask dataset from over 50 hours of surveillance camera footage from the Detroit riverfront. I integrated the dataset, titled OPOS-FM, into the OPOS dataset. OPOS-FM contains 6,039 images of cropped faces. The cropped images are on average 3200 px² with the defining feature (face mask) typically between 100-400 px².

The face mask detector is a CNN-based binary classifier which utilizes a wide residual network [52] architecture. The face mask detection module uses the bounding boxes and tracking information from tracked people. If the velocities from a track state vector indicate a tracked person is headed towards the camera, the bounding box coordinates are used to forward a cropped image of the tracked person to the face mask detection tool. The top 6th of the cropped image is further cropped to isolate the head area. The wide residual network extracts a feature vector from the cropped head through a series of convolutional layers and then follows with a final classification layer; this is illustrated in Figure 2.5. Face mask classification results are associated with the unique person IDs generated from the tracker, so the number of *unique* masks and total usage rate can be calculated. If a tracked person never faces the camera, the face mask classification associated with the ID is left as ``NA'', and the ID is not included in face mask

usage calculations.

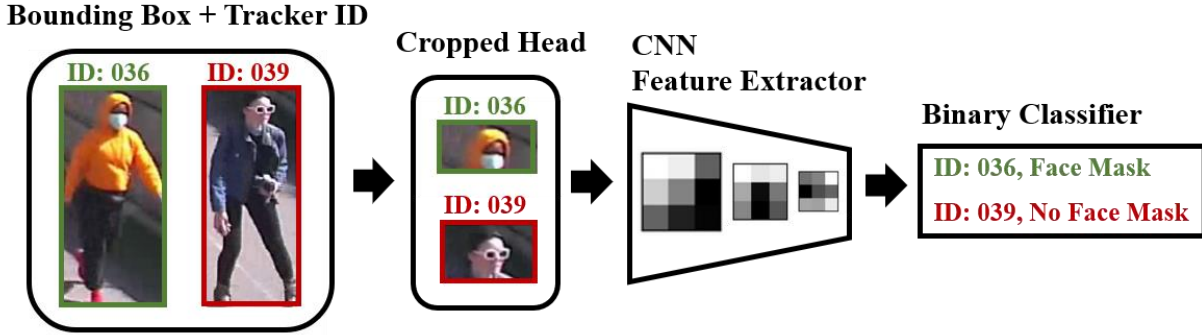


Figure 2.5 Face mask detection using secondary CNN (WideNet) classifier.

2.2 Performance Results

2.2.1 Object Detection

The accuracy of Mask R-CNN and YOLO can be evaluated using precision metrics commonly used for object detectors. The previously defined IoU (Equation 2.12) is often used to assess the performance of the detector, where IoU is defined as the intersection of the ground truth (GT) and the detection result divided by the union of the GT and detection result. Usually, an IoU threshold is set in advance to classify whether a prediction is a true or false positive. For example, 0.50 was set for the PASCAL visual object classes (VOC) challenge metric [57], while 0.75 was used for the COCO challenge metric [43]. The average precision metric at a given IoU threshold for a specific object class, c , is defined as

$$AP_c^{IoU} = \frac{TP_c^{IoU}}{TP_c^{IoU} + FP_c^{IoU}} \quad (2.15)$$

Where TP_c^{IoU} is the total number of true positive results; FP_c^{IoU} is the total number of the false positive results.



Figure 2.6 Performance of Mask R-CNN detector with manually annotated ground truth images (A, C, E) with different weather conditions (sunny, cloudy, rainy) and detector results (B, D, F).

The mean average precision (mAP) is defined as the mean value of the AP_c across different object classes:

$$mAP = \frac{\sum_c AP}{N_c} \tag{2.16}$$

where N_c = total number of classes. However, within various data sets, the number of objects is different. Therefore, the mAP of the most common objects can provide more insight into evaluating the overall performance of a detector. These precision metrics (namely, AP and mAP) were used to define the performance of the detectors using the OPOS class taxonomy [45].

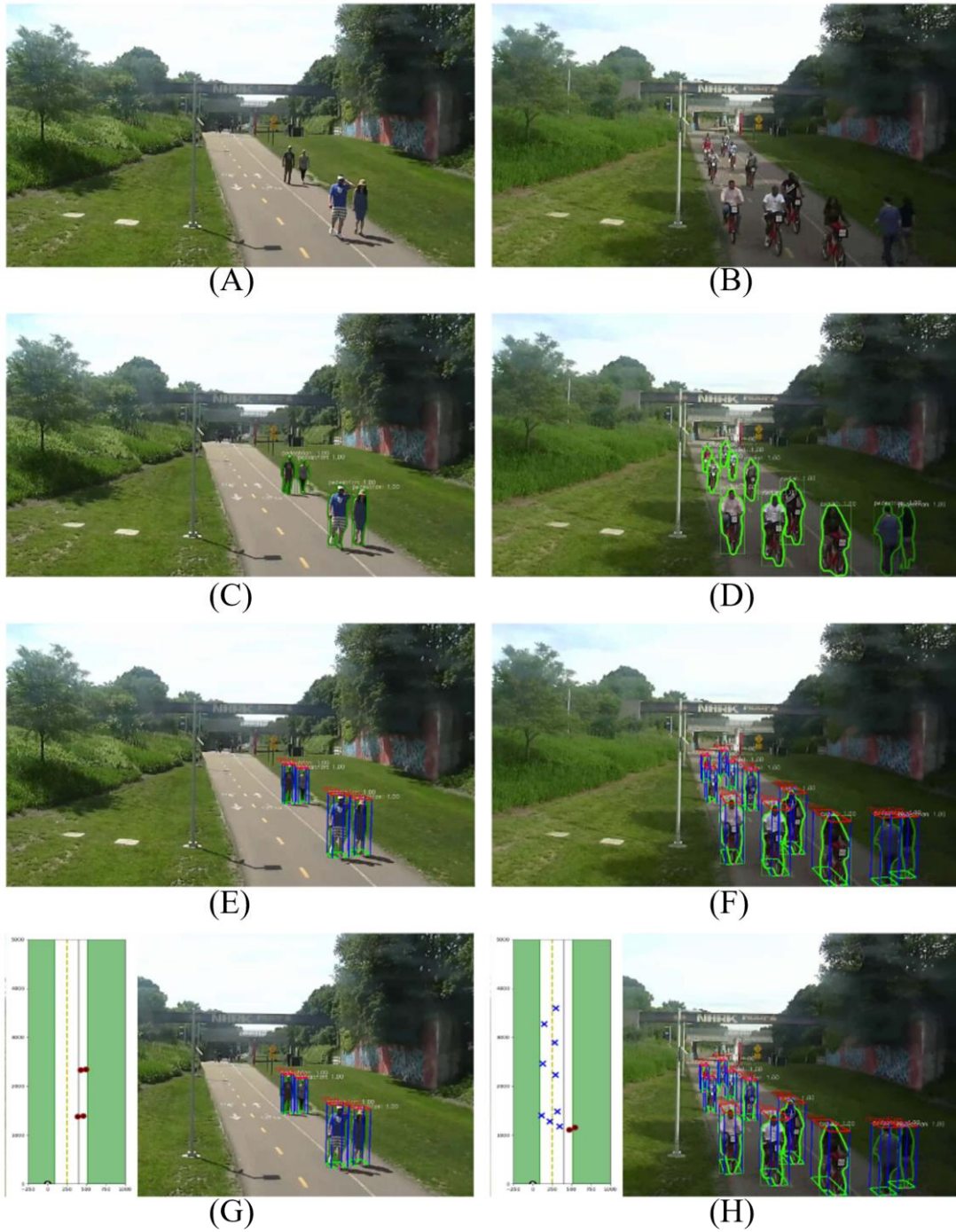


Figure 2.7 Examples at Dequindre Cut (Camera 87) of (A and B) original images with sparse and dense patron distributions; (C and D) processed images with segmentation; (E and F) processed images with 3d bounding boxes; (G and H) processed images with 2D maps.

To evaluate the performance of the Mask R-CNN and YOLO detectors, I withheld a portion of the OPOS data set from training (10% of the images) and placed into the validation set. The validation data set included images taken at various locations and under different weather conditions (e.g., sunny, cloudy, and rainy) to assess the detector performance under varying environmental conditions.

I assessed AP and mAP for the different OPOS classes. The threshold for the confidence score for a positive detection was set as 0.75. The detection results are shown in Figures 2.6 (B, D, and F) for sunny, cloudy, and raining conditions. By comparing the GT [Figures 2.6 (A, C, and E)] and the detection results [Figures 2.6 (B, D, and F)], I found qualitatively that the detection module performs well under various illumination conditions. Patrons (e.g., pedestrian, cyclist, and peopleother) and mobile objects (e.g., stroller, dog, car) were successfully detected and localized. The detected person masks shown in each image can provide detailed information about various parts of a patron or object (e.g., feet, head, tire). The detector also worked well for images with a high density of park uses as shown in Figure 2.7. Images of the Dequindre Cut at Camera 87 with sparse and dense occupation are shown in Figs. 10(A and B), respectively. Figures 2.7 (C and D) are the output of the Mask R-CNN bounding box and segmentation results showing excellent detection capabilities for both sparse and dense occupation of the park space.

Quantitative evaluation of detector performances are summarized in Table 2.1. Using the popular AP_{0.50} metric applied to the Mask R-CNN and YOLO bounding boxes, people detection was excellent with all classes above 89% but with pedestrians and cyclists (by far the most common park user type) above 95%. Some of the objects such as car and stroller were also detected with AP_{0.50} well above 90%. Similar results were obtained for the segmented output of the Mask R-CNN. The mAP was also calculated for the people and object classes. The Mask R-

CNN bounding box results had mAP0.50 of 89.2%, 94.3%, and 91.2% for the people classes, object classes, and overall classes, respectively, while the YOLO bounding box results had mAP0.50 of 86.6%, 92.5%, and 89.0% for the people classes, object classes, and overall classes, respectively. The mask R-CNN segmentation mAP0.50 for the people classes, object classes, and overall classes were 89.1%, 93.5%, and 90.8%. The overall performance of the bounding box head and segmentation head of the Mask R-CNN detector, and YOLO were very similar. With Mask R-CNN scoring slightly higher scores, however, the speed for Mask R-CNN, ~7 fps, was much slower than the ~30 fps speed of the YOLO framework.

Table 2.1 Object detector performance on OPOS validation set

	Mask R-CNN (bbox)	Mask R-CNN (segm)	YOLO (bbox)
APc per ppl class			
Pedestrian	96.4	96.3	96.2
Cyclist	96.5	96.5	95.8
Scooter	89.4	89.4	82.2
Skater	89.5	89.5	80.4
Sitter	89.1	89.5	88.4
Peopleother	74.1	73.1	76.8
APc per obj class			
Car	93.0	93.0	93.0
Stroller	99.7	96.7	96.7
Dog	84.4	84.4	80.3
Umbrella	100.0	100.0	100.0
mAP (ppl)	89.2	89.1	86.6
mAP (obj)	94.3	93.5	92.5
mAP (overall)	91.2	90.8	89.0
Speed (fps)	~6.7	~6.7	~29.9

2.2.2 Object Tracking

2.2.2.1 Performance metrics

I assessed the tracking algorithm using multiple-object tracking (MOT) metrics [58] including multiple-object tracking accuracy (MOTA), multiple-object tracking precision (MOTP), mostly tracked targets (MT), partially tracked targets (PT), mostly lost targets (ML), and the ID switch number (IDs). MOTA and MOTP focus on the performance of the algorithm at each frame. Specifically, MOTA provides an objective comparison of the object assignment at each frame by considering the number of false negative (FN) assignments, false positives (FP) assignments, and tracklet mismatch (IDs) over all frames

$$MOTA = 1 - \frac{\sum_k (FN_k + FP_k + ID_{S_k})}{\sum_k GT_k} \quad (2.12.7)$$

where k = frame index; and GT_k = number of ground-truth objects in the frame. MOTP used here follows the MOT challenge benchmarks definition [59] rather than the original definition in [58]. MOTP, a measure of the precision of the detector localization, is defined as

$$MOTP = \frac{\sum_{k,i} IoU_{k,i}}{\sum_{k,i} M_k} \quad (2.12.8)$$

where M_k = number of matches in frame k ; and $IoU_{k,i}$ = bounding box overlap of target i with its assigned ground-truth object in frame k for all the correctly matched hypotheses and their respective objects.

MT, PT, ML, and IDs parameters focus on the performance of the tracking algorithm at the trajectory level. When comparing the computed trajectories with their GT counterparts, the trajectories are classified into one of the three categories: MT, PT, or ML. An MT trajectory is

defined as one where 80% or more of its life span is tracked (i.e., 80% overlap with the GT), an ML trajectory is one where less than 20% of its life span is tracked, and all other trajectories are classified as PT. Maximization of MT and minimization of ML is preferred. Full tracklets are preferred rather than the GT being divided into smaller tracklet segments with many interruptions; the IDs refers to the number of transitions between smaller tracklets making up the GT.

2.2.2.2 Tracking performance

To evaluate the performance of the proposed DeepSORT tracking-by-detection framework, I built a custom small-size tracking data set using a 30-min-long video (with 53 unique individuals over 9,847 frames at Camera 87) collected along the Dequindre Cut pedestrian path on April 28, 2020, starting at 2:00 p.m. This custom data set includes instances of sparse and dense occupancy of the park space such as that shown in Figure 2.7. The two sets of detection-tracking combinations were evaluated (Table 2.2): (1) Mask R-CNN with DeepSORT and traditional SORT, and (2) YOLO with DeepSORT and traditional SORT. In the table, n_{skip} is used to represent the number of the frames skipped when tracking. For example, $n_{skip} = 0$ represents using detection results from the original video with no frame skipping, while $n_{skip} = 1$ represents skipping every other frame for tracking. The larger n_{skip} is, the quicker the detection and tracking process is due to the reduced computational efforts associated with a slower frame rate. However, skipping too many frames can harm the performance of the tracker, as there are larger gaps of motion information between frames. In practice, tracking patrons on video can often generate a number of false positives in the detection process; however, tracklets associated with false positives from the detector are usually very short over a small number of

frames. Hence, the module has a postprocessing step that filters out these tracklets. Specifically, a threshold of 2x the camera framerate is used to filter tracklets surviving less than 2 seconds.

The results in Table 2.2 reveal excellent performance of the Mask R-CNN and YOLO detection modules with the DeepSORT the tracking framework. The detector and tracker received high scores on the MOTA. Precision metrics are associated with the detector performance, while MOTP is associated with the localization accuracy. When using the original video ($n_{skip} = 0$), DeepSORT was slightly better than using SORT, but when frames were skipped ($n_{skip} = 1$ and $n_{skip} = 2$), DeepSORT outperformed SORT in terms of MOTA, MOTP, and precision. The improved performance of DeepSORT over SORT was more evident in the metrics associated with the number of identified tracklets compared to the GT. For example, MT counts for DeepSORT were dramatically higher than for SORT. The differences in tracking performance for the filtered versus raw tracklet sets were negligible, although the filtering step can slightly reduce the total ID count to the ground truth, making it a useful function in real practice for patron counting.

Table 2.2 Tracking performance on custom video

Method (detect + track)	Skip frame	MOTA (%)	MOTP (%)	Precision (%)	GT	MT	PT	ML	IDs	ID Count
Mask R-CNN + SORT (raw results)	nskip = 0	78.2	85.2	98.3	53	21	32	0	54	80
	nskip = 1	66.1	84.7	99.1	53	9	44	0	48	88
	nskip = 2	-35.3	83.8	33.4	53	0	47	6	34	88
Mask R-CNN + DeepSORT (raw results)	nskip = 0	80.3	87.6	95.1	53	33	20	0	36	79
	nskip = 1	72.4	85.7	95.3	53	18	35	0	33	79
	nskip = 2	56.4	83.8	92.8	53	16	42	5	28	72
Mask R-CNN + DeepSORT (filtered results)	nskip = 0	80.4	87.6	95	53	33	20	0	36	77
	nskip = 1	72.6	85.7	95.5	53	18	35	0	30	76
	nskip = 2	56.1	83.8	95.3	53	16	41	6	27	69
YOLO + SORT (raw results)	nskip = 0	77.1	84.4	97.3	53	20	32	0	55	80
	nskip = 1	65.1	83.9	98.5	53	8	44	0	49	88
	nskip = 2	-36.3	83.1	32.6	53	0	47	6	34	88
YOLO + DeepSORT (raw results)	nskip = 0	78.7	85.6	94.2	53	32	20	0	38	82
	nskip = 1	70.8	83.9	94.3	53	18	34	0	35	80
	nskip = 2	54.4	81.3	91.4	53	16	42	5	30	80
YOLO + DeepSORT (filtered results)	nskip = 0	78.8	85.6	94.2	53	32	20	0	35	79
	nskip = 1	80	83.9	94.6	53	18	34	0	31	77
	nskip = 2	54.2	81.3	92.4	53	16	42	6	29	70

2.3 Experiments and Field Deployment

2.3.1 Detroit Riverfront Camera Network

I performed the experimental work of this study in the park spaces of the DRFC. The DRFC was formed in 2003 with the mission of restoring the riverfront area of downtown Detroit along the Detroit River (which also serves as the US–Canada border). The DRFC manages two major public spaces: Detroit Riverwalk and the Dequindre Cut. The riverwalk consists of a 5.63 km (3.5-mi) long park along the Detroit River starting in the west at West Riverfront Park and ending at Gabriel Richard Park in the east. An additional greenway has been developed running north from the river to Eastern Market called the Dequindre Cut, which is approximately 3.21 km (2 mi) long. The DRFC park areas are an iconic social space for the City of Detroit and attract approximately 3 million visitors annually. I selected the park in this study as an

illustrative example of applying the work to social infrastructure because of the high volume of daily visitors, diverse set of park amenities that support a range of social interactions, and existing camera-based infrastructure. The DRFC was an ideal partner for the study because of their interest in a quantitative approach to assessing utilization of their spaces that could inform their future investments in the management of the park.

The camera network installed in the DRFC parks consists of 100 surveillance cameras with three types: pan-tilt-zoom (PTZ), stationary surveillance, and fisheye. In this study, a total of 15 surveillance stationary cameras were selected from the network to study the usage patterns at the Dequindre Cut and the Riverwalk. I intentionally considered this diverse set of cameras to show the unbiased nature of the people detection and tracking framework trained by OPOS to any specific camera, allowing it to be applied to cameras with different performance characteristics (e.g., resolution, lenses, poses) and scenes.

Along the Dequindre Cut are four cameras located at Gratiot Avenue (Camera 87), Adeliaide Street (Camera 88), Division Street (Camera 89), and Alfred Street (Camera 90); these cameras have a resolution of $1,280 \times 720$ pixel². At the Freight Yard located along the Dequindre Cut are four cameras, Cameras 96A–96D, covering various areas of a concession area with seating. Another camera (Camera 29) at Cullen Plaza was selected, which has a resolution of $1,108 \times 832$ pixel². Selected along the western Riverwalk were 10 cameras, with two near the Detroit Port Authority (Cameras 48 and 49), two near Hart Park (Cameras 50 and 51), two near Cobo Hall (Cameras 52 and 53), and four near Joe Louis Arena (Cameras 54–57); these cameras have a resolution of $1,280 \times 720$ pixel². The cameras record their images at a various rates (5,10,15, and 30 fps) with videos stored in an online database.

The camera images contain representations of different illumination and weather conditions (e.g., sunny, cloudy, and rainy), allowing the robustness of the algorithms to be assessed. The classification of detected people and their trajectories are evaluated on these cameras to show their use in (1) spatially mapping the behavior of park users; (2) automating the counting of people by OPOS classes; and (3) identifying how people interact with park furnishings.

2.3.2 Activity Mapping

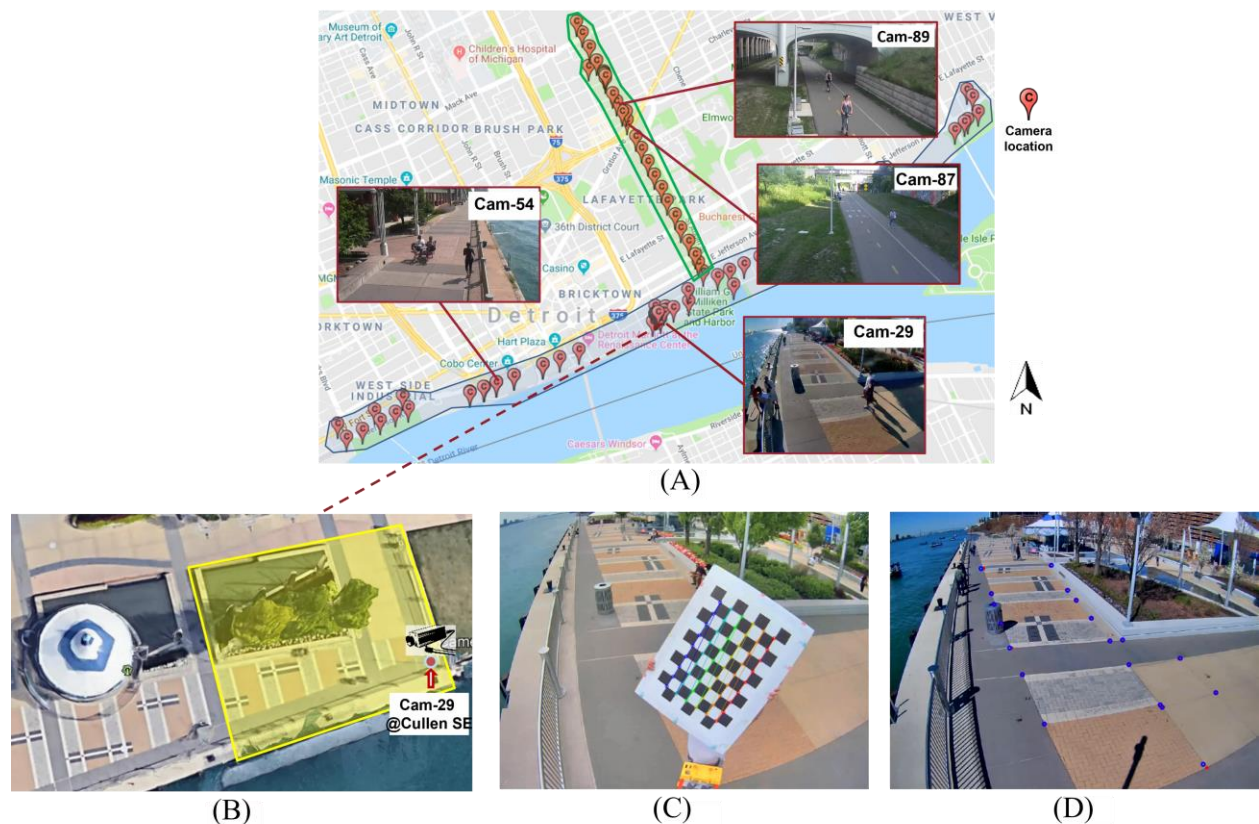


Figure 2.8 (A) Detroit Riverfront parks camera locations; (B) Location of Camera 29 at the southeast corner of Cullen Plaza; (C) Camera calibration with checkerboard; and (D) Calibration reference points.

With this camera network, the research aims to validate the sensing framework's ability to spatially map detected park users on a geographic information system (GIS) map of the monitored space. This spatial mapping provides park managers a rich visual depiction of what spaces people use and how they use them. The map of Detroit riverfront parks and camera locations is shown in Figure 2.8(A), as well as pictures of the four exemplary camera views at Cameras 29, 54, 87, and 89. Cullen Plaza [Figure 2.8 (B)], which is located at the center of the Detroit riverfront area, was adopted for mapping user activity, with all plots and data in this section coming from the summer of 2019. There are three surveillance cameras installed at different locations in the plaza: southeast corner, northeast corner, and southwest corner. The camera on the southeast corner (Camera 29) of the plaza was selected, which has a visual range of roughly 20 m. Camera calibration was performed using a checkerboard [Figure 2.8 (C)] and by defining reference points on the ground plane [Figure 2.8 (B)], resulting in a pinhole camera model capable of mapping detected people in the world coordinate system of the map.

The field of view of the monitored area is about 18×20 m and the dashed line demonstrates the range of the captured view. Fig. 11 provides an example of detected people on Cullen Plaza and the automated mapping of their location and activity on a 2D map layer. Figure 2.9 (A) shows a sparse number of users during afternoons in June 2019 with pedestrian users walking and leaning on the fence, sitter users sitting on the concrete wall around the trees, and scooter users transiting the space. These users are also mapped to the 2D map layer in Figure 2.9 (B). Figures 2.9 (C and D) show a dense number of users; these pedestrians are boarding a boat docked at Cullen Plaza. In both views, the locations of users' feet are marked in each frame.

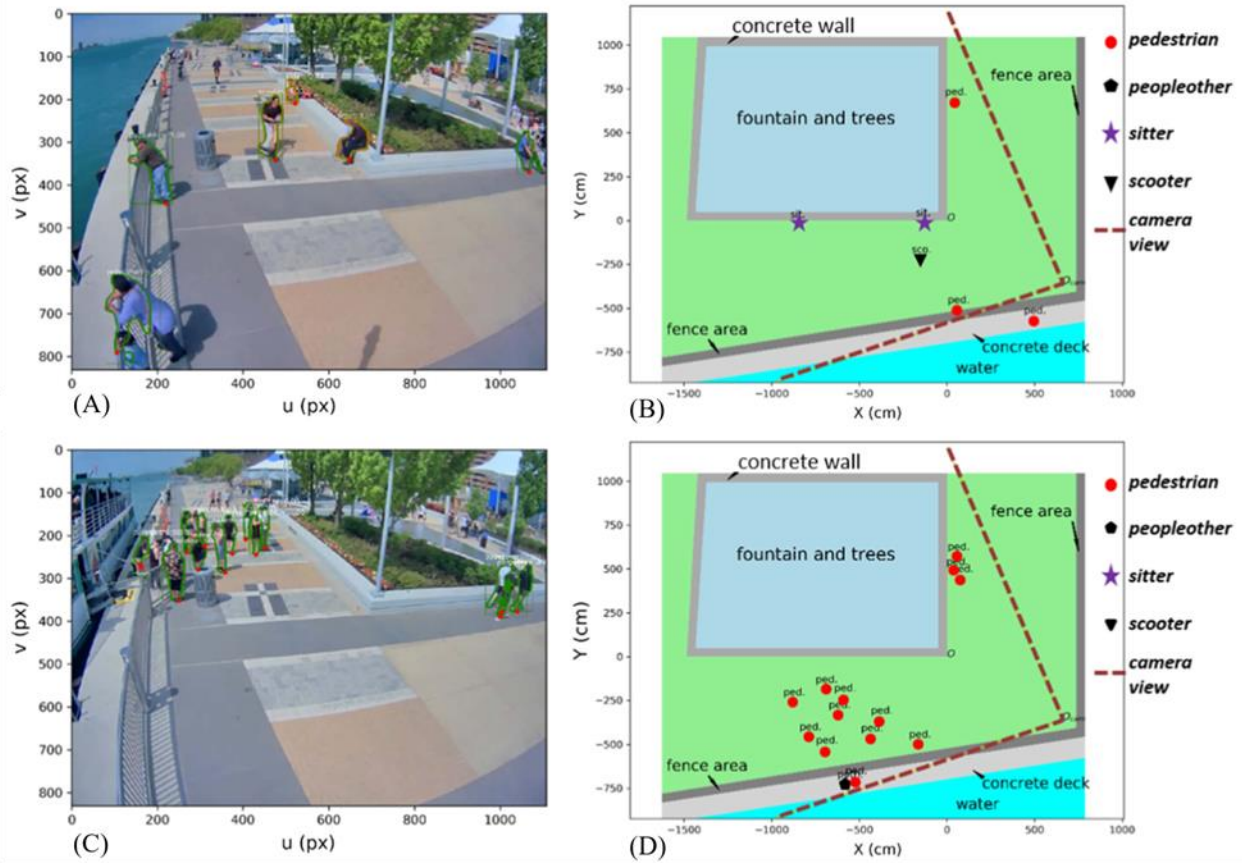


Figure 2.9 Activity classifications in the pixel coordinate system (A and C) and corresponding mapping in the world coordinate system (B and D).

Sometimes, the detector has some difficulty for activity recognition when people appear small.

For example, the patrons sitting on the concrete wall to the far right of the image [Figure 2.9 (A)] are detected as a single pedestrian rather than a pair of sitters. The small number of false detections within the same supercategory (i.e., people) is acceptable and unavoidable for almost all existing CV-based detectors.

To study how people use Cullen Plaza, Figure 2.10 shows a scatter and density map of the plaza for one overcast summer weekday (Wednesday, June 24, 2019), one sunny summer day on the weekend (Saturday, June 29, 2019), and total for the week from June 24 to 30, 2019; the period of observation is from 9:00 a.m. to 5:00 p.m. EST each day. The CNN based detector was applied to the camera at a rate of 1 fps, resulting in 28,800 processed frames each day. The

scatter maps of the accumulated superclass people results over a weekday and a weekend are shown in Figures 2.10 (A and B), respectively. It is apparent from the results that more people enjoy Cullen Plaza on a sunny weekend [Figure 2.10 (B)] than on an overcast weekday [Figure 2.10 (A)]. The hollow rectangular area in the middle part of the plaza ($X \in [-500, 0]$ cm and $Y \in [-300; -500]$ cm) on the scatter plot of Figure 2.10 (B) is due to occlusions from the trash can making detection of people's feet location impossible immediately behind the can. Also, people are detected in the fountain and tree area; this is due to a grounds-keeping crew accessing these spaces as shown in Figure 210 (G).

A more revealing way of presenting the same data is to map space use as a density map [Figures 2.10 (D–F)]. The density maps were generated from the scatter maps by using a kernel density estimation [60] on the detection results. The density maps allow locations of greatest use to be identified. During the weekday, the fence area is of greatest use by park patrons [Figure 2.10 (D)] as they look out along the Detroit River [Figure 2.10 (I)]. On the weekend, especially a sunny one, people make use of the concrete parapet around the fountain and tree area for sitting [Figure 2.10 (E)]; this is due to the shade the trees provide to this spot in the afternoon of a sunny day [Figure 2.10 (H)]. In general, these results align with the edge effect [13] where people feel comfortable and safe when they stay close to the edges of public space (e.g., fountain steps, poles, fence area, statues). Spatiotemporal mapping of user activities offers park managers deeper insight into how to maintain and remake park features to best meet patron needs. For example, on June 30, 2020, between 6:00 and 7:00 p.m., 256 patrons visited Cullen Plaza (Camera 29). Of the 256 patrons, 59 sat on the fountain parapet, indicating a need for more

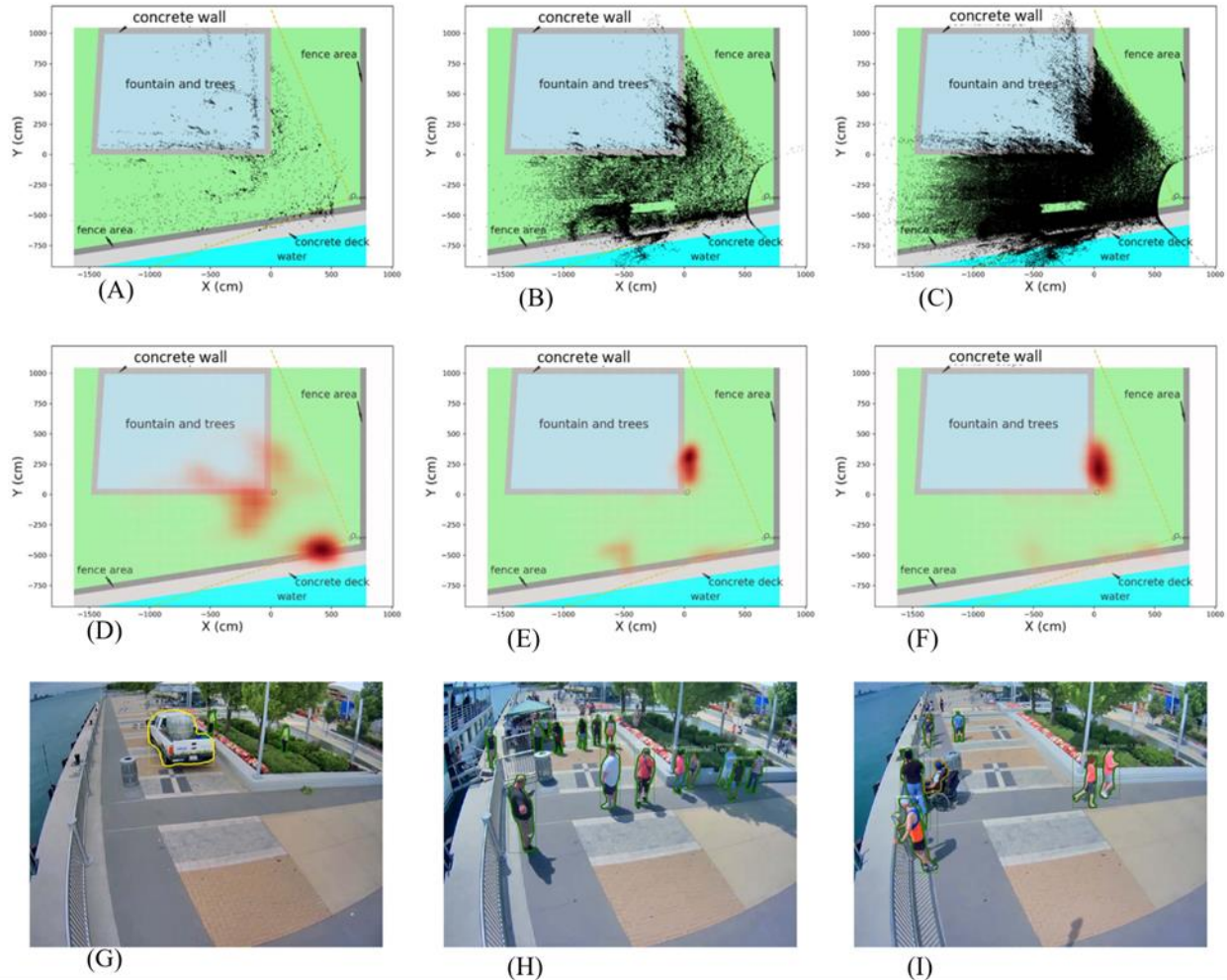


Figure 2.10 Scatter maps and Density plots of detected users on a cloudy weekday (A, D), a sunny weekend (B, E), and total over a week (C, F). Workers trimming trees (G), patrons seeking shade (H), and patrons sightseeing along the fence (I)

comfortable seating in the area. Additionally, 65 patrons stopped to lean against the fence and view the Detroit River. Using these data, park managers could see patrons need more seating options and desire to view the river scenery. With this data, a park manager would know that placing new seating with a good vantage point of the river would be an appreciated investment.

2.3.3 Continuous People Counting with Mobility Direction

People counting is useful to obtain the number of patrons in a public open space.

However, most existing counting methods (e.g., use of passive infrared sensors) can only count

people in a discrete fashion without any indication of their movement or activity. In this study, the tracking-by-detection framework was extended to count park patrons with the direction of their movements quantified to provide a sense of people flow in the park space. The surveillance camera (Camera 87) located on the north-south Dequindre Cut (near the Gratiot Avenue entry) viewing the northbound pedestrian path and the camera (Camera 54) located on the riverwalk looking west were chosen to study the movement of people in these pedestrian pathways. Usually, a camera that is positioned high (e.g., 3–4 m) on a pole can cover a field range of 20–50 m and a horizontal angle range of 60°–90°, which provides a wide viewing field for patron detection and activity recognition. Each camera was calibrated using the aforementioned checkerboard and the reference points assigned to the pedestrian path. The number of unique users of the space was continuously counted from the tracking-by-detection results (i.e., number of tracklets). Furthermore, the spatial direction of people moving was determined based on if the person was moving away or toward the camera based on the height and width of their bounding box: those getting bigger in the field of view correspond to objects moving toward the camera, and those getting smaller are moving away from the camera. Pedestrians were assigned to those moving north or south along the Dequindre Cut path.

Results are shown during a short period (between 11:00 and 11:30 a.m.) on June 8, 2019, along the Dequindre Cut (Camera 87) in Figure 2.11 with the video streaming at 5 fps. Each unique person tracked is denoted in the figures as P# and each number is the unique identification assigned to the detected person during the tracking process. Also shown for each image are the positions of the detected park users on a 2D GIS map layer of the Dequindre Cut. There are two ways for park users to enter the camera view: one is by approaching from far away

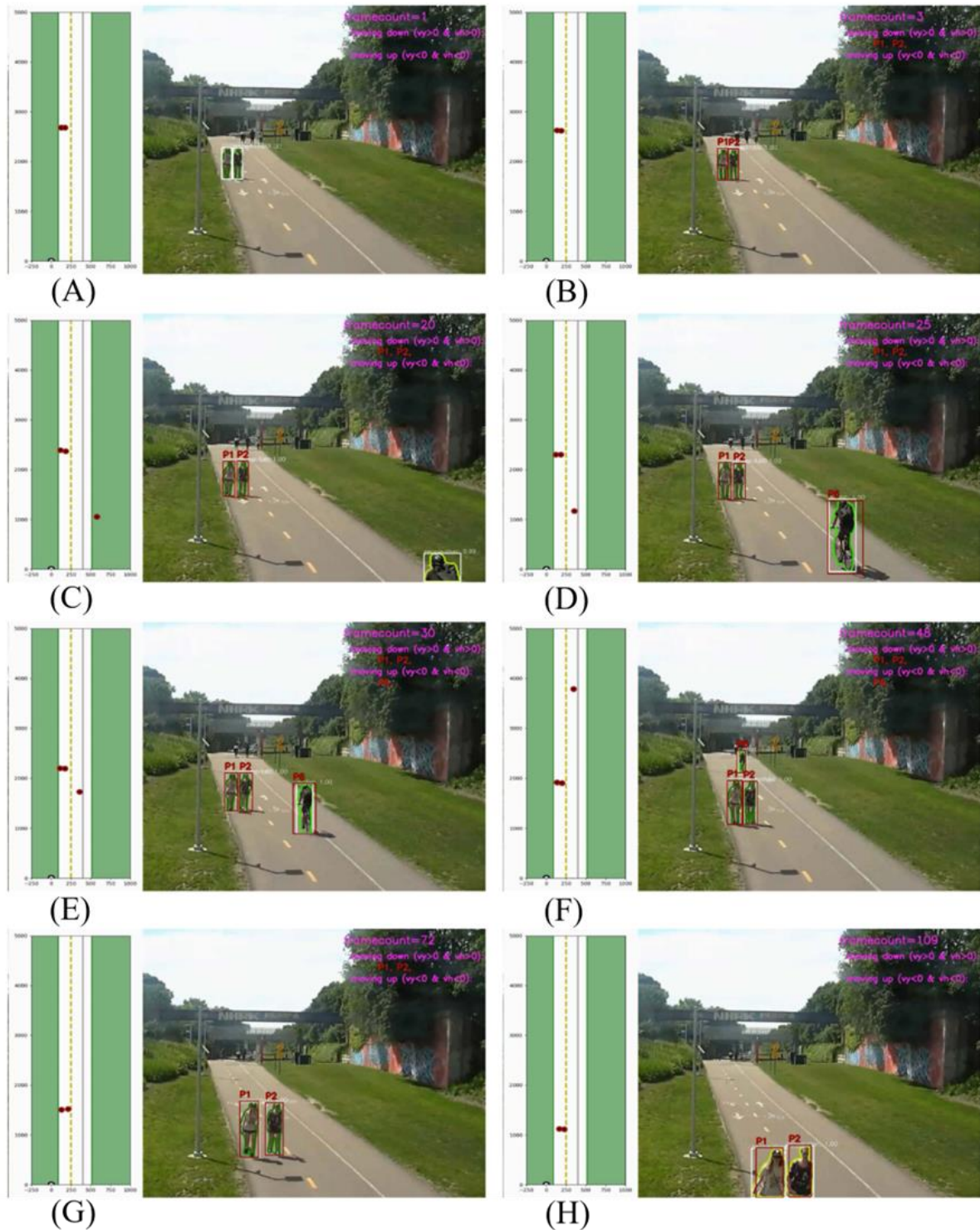


Figure 2.11 Examples at the Dequindre Cut (Camera 87): (A and B) pedestrian tracking and continuous counting with directions; (C and D) cyclist entering the field of view; (E and F) cyclist approaching two pedestrians and about to exit viewing; and (G and H) pedestrian exiting the field of view.

with the person size changing from small to big [e.g., Figure 2.11 (A)]; the other is the person

entering the view from the image boundaries [e.g., Figure 2.11 (C)]. The first frame of the sample video is denoted as frame 1 at $t = 0$ s as shown in Figure 2.11 (A) with two pedestrians initially detected and denoted with a bbox. Because the tracking statuses of P1 and P2 are tentative, they are not plotted on the processed video frame. After being detected for three consecutive frames, P1 and P2 trackers are confirmed by the third frame and shown with bounding boxes at $t = 0.4$ s in Figure 2.11 (B). At the 20th frame ($t = 3.8$ s) shown in Figure 2.11 (C), a cyclist with severe truncation is detected at the bottom of the image and a P3 tracker is created with a tentative status during postprocessing. This is an example of a person entering the camera view from the boundaries of the image. During subsequent frames, the cyclist gets bigger but due to dramatic change in appearance, the DeepSORT tracker does not associate the cyclist with a past frame, so each time algorithm creates a new tracklet identifier (e.g., P4, P5). It is not until the 25th frame [Figure 2.11 (D)] that the DeepSORT tracker confirms the tracklet, at which time the tracklet is assigned P6, suggesting the cyclist was first assigned to tracklet denoted P6 in the 23rd frame. Over defined periods, the number of unique people detected and tracked can be counted to provide direct counts of the space in view. Through the tracking and counting module, the mobility of park users can be obtained on any video feed of interest. For example, from $t = 0.0$ s to $t = 19.0$ s, P1 and P2 are moving downward [Figures 2.11 (G and H)] in the PCS, which corresponds to a northerly direction in the WCS; the average speed of the two pedestrians northward is computed as 0.85 m=s. From $t = 4.4$ s to $t = 9.4$ s, the cyclist labeled P6 is moving upward in the PCS (which corresponds to a southerly direction) and the average speed is computed as 5.43 m=s.

There are two ways for park users to exit the camera view: one is leaving with vanishing size in the far field [e.g., Figure 2.11 (F)], while the other is exiting the view from the image

boundaries with abrupt appearance changes [e.g., Figure 2.11 (H)]. The former is handled well by the detection module because the detector will stop detecting the person if they get too small in the field of view. The exiting of people from the boundaries is more challenging for the tracker because the person goes from a whole person to a truncated person [with the latter likely assigned to the OPOS peopleother class as shown in Figure 2.11 (H)]. In general, the results in Figure 2.11 show very robust performance in dealing with patrons entering and exiting the field of view.

The number of people using a space can be plotted as a time history to understand how space use varies over time. To showcase this capability, Fig. 2.12 plots the number of users per hour along the Dequindre Cut (Camera 89) and riverfront walk (Camera 54) from 11:00 a.m. to 1:00 p.m. each day for the dates shown. Figure 2.12 (A) plots the total number of people using the Dequindre Cut during the summer (from June to August 2020), while Figure 2.12 (B) corresponds to the fall season (from September to November 2020); park users are segregated as pedestrians and cyclists. In general, the use of the Dequindre cut is cyclical, with significantly greater use during the weekends than the weekdays, regardless of season. There are slightly more pedestrians than cyclists at any given day along the Dequindre Cut. The riverfront had similar trends [Figure 2.12 (C)] with higher use on the weekend. However, the riverfront has more pedestrians than cyclists, with a relatively low number of cyclists, averaging 12 cyclists per hour. The Dequindre Cut is used more often for personal transit and features a dedicated cyclist lane, suggesting a greater number of cyclists using it. In contrast, the riverwalk is designed more for pedestrians (i.e., does not have a cyclist lane) strolling along the riverfront. Overcast and rainy days are denoted in Figure 2.12 with numbers greatly depressed during overcast and rainy

weather at both sites. This is especially notable in October as shown in Figure 2.12 (b), which had the lowest traffic and the worst weather.

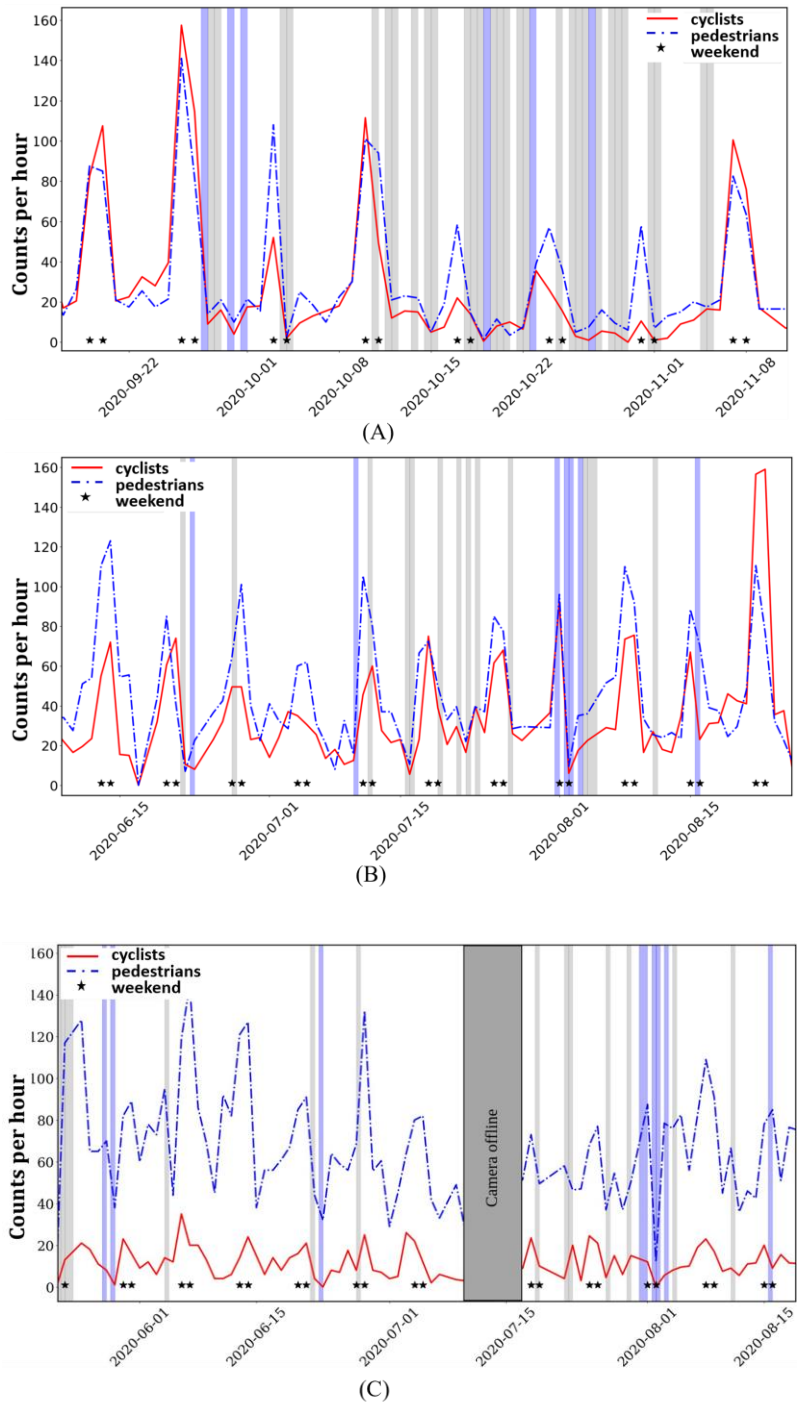


Figure 2.12 Daily pedestrian count and cyclist count from 11:00 to 13:00. (A) at the Dequindre Cut in fall 2020; (B) at the Dequindre cut in summer 2020; (C) at the riverfront in summer 2020. Grey segments represent overcast weather, segments with diagonal stripe pattern represent rainy weather.

2.4 Conclusion

Beyond the engineering work to test, verify and integrate all the individual components into one monitoring framework, this chapter covered a couple of novel intellectual contributions:

- Development of robust training dataset specifically designed to enhance object detection in public spaces using low-resolution cameras.
- A secondary lightweight CNN-based classifier strategy to enable more complex detection (*e.g.*, Face mask, body pose) without the need for exhaustive training data while staying in real-time.

Additionally, this chapter has a number of key findings. First, the work proves the ability of a trained detector to autonomously identify people and their activities reliably in social infrastructure systems. For example, the performance of the detection module using the OPOS data set is impressive with a (bounding box) mAP_{0.50} of 89.17% for people classes and a (bounding box) mAP_{0.50} of 88.19% for the overall classes including mobile objects. This presents a major advancement over current manual observation methods in widespread use. Second, the bounding boxes of the detected persons allow for precise localization and enhances the accuracy of mapping people to a 3D coordinate system. Third, the ability to track unique individuals over a time horizon using consecutive images from a camera stream was also proven, especially during daytime conditions. Specifically, the tracking module modified from DeepSORT was proposed using the detector output. The motion and appearance features were shown to be powerful association metrics for solving the problem of assigning detected people to unique tracklets. Using a Hungarian algorithm for solving the assignment problem, excellent tracking results were achieved with 80.40% in MOTA and 87.60% in MOTP for tracking

performance using the developed tracking data set during the day. Fourth, a simple, low-cost solution for adding additional context to detected persons was proven – face masks of individuals could be tracked without needing to retrain detection models or curate large datasets. Finally, a key finding of the work was the utility of the data collected to managers of social infrastructure spaces. The chapter utilized security cameras at the Detroit riverfront public spaces to showcase how managers of social infrastructure can use the proposed detection and tracking framework to better understand what space the public uses and what activities (e.g., sitting, biking, walking) they do in those spaces.

2.4.1 Mask R-CNN vs YOLO

Mask R-CNN had slightly higher precision and accuracy in the task of person detection (Table 2.1). Additionally, the model provides segmentations of detected objects, providing deeper information than regular bounding boxes. However, the model was significantly slower, running at approximately 7 frames per second versus the 30 frames per second runtime of YOLO. In the early stages of tool development and testing the Mask R-CNN model is suitable. In this chapter, video feeds were processed offline, and time was not an issue. However, the heaviness of the model and the slower speed will cause issues for scaling and real-time solutions. For these reasons, the YOLO framework will be primarily used in future chapters. Not only is YOLO faster, but it is also lighter-weight and can run on edge devices - adding flexibility and power to the sensing framework.



Figure 2.13 Examples of object detection using YOLO framework trained on OPOS

Chapter 3. Extended DeepSORT Framework

This chapter presents novel work in multi-object tracking to track people as part of a comprehensive monitoring framework (Figure 1.2) that can measure the performance of social infrastructure and quantify sociability. Social infrastructure (public open spaces, parks, squares, bus stops, etc.) is an important class of civil infrastructure that facilitates social interaction and healthy lifestyles. This type of tracking data is not only useful to social infrastructure but can also improve city planning and traffic management. Knowing the number of pedestrians and their routes taken through various city streets, plazas, parks, and shopping centers enables data driven decisions regarding city design and management, empowering decisions on public transit operation (schedules, stops, etc.), park design, events, and future investments. Additionally, understanding social behaviors and traffic flow trends can improve evacuation strategies (by increasing knowledge of probable route selections) during emergency events [61]. Person monitoring and activity classification is also especially beneficial on construction sites, vastly improving worker safety, health, and productivity [62]–[64]. However, such data is cumbersome and costly to manually collect.

In an effort to automate person tracking and counting, researchers have developed various solutions. Propelled by the rapid advancement of sensing technology, the emergence of smart cities has fostered a diverse range of autonomous solutions for person tracking and counting. From health tracking wearables to intelligent video surveillance systems, these innovations offer insights into urban dynamics, public safety, and individual well-being, paving

the way for more informed decision-making and targeted interventions. Some industry available pedestrian counting systems utilize discrete sensors (passive infrared, geophones) placed along pathways to give counts of people passing through, and when distributed across a space, counts of pedestrians (and their direction traveled) can be estimated. Other methods utilize smart-phones [25] and WiFi signals [26] to count and localize pedestrians carrying a smart phone. By comparison, video cameras have advantages thanks to the wealth of information they provide, and the ubiquity of cameras already installed in cities for surveillance. For person counting, readily available software has been developed for use on cameras which generates foot traffic and direction estimates through blob detection and centroid tracking, by drawing a line in the frame and counting every time a detected object's centroid passes through the line. However, to be effective the camera angle needs to be aimed directly downwards, limiting the field of view and other uses for the camera. In these methods only counts and direction are estimated.

With recent advances in deep-learning based computer-vision models for object detection, more complex and capable methods for person monitoring have been developed. Tracking-by-detection models such as SORT [51] and DeepSORT [49] leverage outputs from state-of-the-art convolutional neural network (CNN) based object detection models, re-identifying detected objects as they move through the scene. In construction, tracking-by-detection frameworks [51], [65] have been used to track earth-moving equipment [66] and construction workers [67] to increase safety, and productivity and activity assessment. Furthermore, these frameworks can be applied on existing surveillance cameras, using field of views encompassing large segments of a walkway or entire plazas. While such methods are state-of-the-art, and have been successfully implemented in other applications, recording pedestrian traffic and their trajectories in public open spaces [45], they are prone to error when

tracked targets are heavily occluded. Additionally tracking suffers during low illumination or other instances where object detection is challenging.

This chapter seeks to address shortcomings of current multi-object detectors and improve tracking performance, specifically in the application of person monitoring in public open spaces. To better understand how users derive benefits from park spaces and other forms of social infrastructure and meet the goals of the overarching sensing framework (Figure 1.2), it is crucial to be able to reliably delineate individual trajectories. Doing so enables more complex analysis on traffic patterns, feature (fire pits, fountains, accessibility ramps, etc.) use, and patron social behaviors.

This chapter presents novel extensions to the deep-learning tracking-by-detection framework, DeepSORT, seeking to improve tracking performance and resiliency against occlusion and object detection errors. A dynamic tuning strategy is implemented to continuously adjust influence from motion and appearance information based on contexts in the scene, namely the assumed discriminative power of appearance and the uncertainty of motion model. Additionally, changes are made to the calculation of the cost function in DeepSORT for data association. Finally, a person counting “mass-balancing” algorithm is implemented to fix broken trajectories in real-time before duplicate tracklets are made for a single person. In an effort to stay computationally tractable and useful to the wider objective of person monitoring systems, all extensions operate in real-time and the framework is deployable on readily available GPUs and works with low-resolution surveillance cameras.

DeepSORT, along with other computer vision-based tracking algorithms are discussed in the Background section (Section 3.1). In Section 3.2: Methods, extensions to the framework are described in detail as well as the evaluation strategy. Additionally, in Section 3.2, further details

and a review of other relevant tracking and detection algorithms are presented. Performance comparisons between default DeepSORT configurations, other tracking algorithms (OCSORT, ByteTrack), and our extensions are presented in Section 3.3: Results. Finally, concluding thoughts on performance, limitations, and applications are found in Section 3.4: Conclusion.

3.1 Background

3.1.1 *DeepSORT and Multi-object tracking*

This work extends the DeepSORT tracking algorithm [49] to improve its performance for real-time person counting and tracking in public open spaces. At its core, tracking is a complex re-identification (Re-ID) problem, seeking to re-establish the identity of objects within consecutive image frames to track them over time. Multi-object tracking algorithms seek to assign object detections to “tracklets” which contain a unique ID and detection info of each detection assigned to it. Tracking can then be structured as an assignment problem, where objects identified within a specific frame are allocated to existing tracks (using prior tracking outcomes) or designated to initiate a new tracklet.

The DeepSORT algorithm (Fig. 3.1), a tracking-by-detection model, takes a multi-faceted approach by leveraging appearance and motion information for object re-identification. Traditional multi-object tracking approaches have assigned detected objects to tracklets using similarity in appearance features. Past works using appearance features for tracking have used hand crafted visual descriptors such as RGB color histograms, autocorrelograms, and local binary patterns (LBP) to extract color and texture information of detected objects. More recent methods, however, leverage deep learning models instead to extract appearance features for extracting appearance features based on intuitive visual parameters [68].

Other works, such as SORT [51], have focused on motion tracking, modeling a detected object's state via bounding box information and projecting it into the next frame. Locations, in the frame coordinate system, of tracklets in frame $k-1$ are projected into frame k using a Kalman Filter and linear velocity model. Locations of each detection in D , the set of detections in frame k from the object detector, are then compared to each tracklet in T , the set of all existing active tracklets, and used to match detections to tracklets. Performance of SORT is particularly vulnerable to issues caused by occlusion. When an object, such as a person, is occluded from a camera's field of view temporarily, either due to some physical obstruction in the site or another person walking in front of them, they will likely not be detected by the object detector. During these frames, the motion model for the object becomes more and more uncertain as time passes without any updates on the object's location and speed. The location estimate for the objects tracklet becomes poorer as it keeps getting projected into the next frame without any updated observations making it difficult to associate the object back to the original tracklet when it does finally reappear.

Object-Centric SORT (OCSORT) [69], is a recent extension to SORT, and improves performance by alleviating the error accumulation in the Kalman Filter due to lack of observations through a smoothing strategy. OCSORT additionally incorporates object direction when assigning detections to tracklets. ByteTrack [70] is another recent extension to existing tracking-by-detection models. Traditionally, object detectors filter out detections whose confidence scores are below a user-defined threshold. ByteTrack removes the threshold and instead considers every possible detection. Zhang et al.(2022) postulates issues with broken tracks due to partial occlusion can be solved this way, arguing partially occluded targets are usually still detected but are filtered out due to low confidence scores. By including low scoring

detections in the association phase, tracks may match to these detections that would have otherwise been discarded.

DeepSORT is an extension of SORT, projecting tracklet states into future frames and associating detections in those frames to current tracklets. However, DeepSORT utilizes appearance information in addition to motion information to associate detections to tracklets. Assigning detections in frame k to tracklets in T using the DeepSORT algorithm is described in full detail in Section 2.1.3 within Chapter 2.

3.2 Methods

3.2.1 *DeepSORT Extensions*

DeepSORT improved upon its precedents thanks to the combination of motion and appearance information. It remains a popular choice for pedestrian counting and tracking. However, DeepSORT can still struggle with long periods of occlusion, slow frame rates, dark or poor-quality video, and low resolution far-away targets. Other works [69], [70] resolve some of the problems with DeepSORT by improving the Kalman Filter or the quality of information used in the assignment process. However, state-of-the-art models are generally tuned for surveillance scenarios that differ from social infrastructure. They are commonly evaluated on public MOT benchmark challenges like MOT17 [59], which features scenes on public streets and sidewalks, and MOT20 [71], which includes much busier crowded scenes on streets and in transit centers. The nature of public interaction is largely transient, with people entering and exiting the frame as passersby. Social infrastructure, by contrast, is meant to draw community members to a space for enduring social interaction, serving as a destination in itself. Social infrastructure draws individuals and groups to the scene for longer periods of time, giving rise to longer tracklets for

each identified person. Re-identification of identified persons is of paramount importance, as members of a social group commonly occlude one another from the field of view.

This work explores extensions and modifications to the DeepSORT framework aimed at improving stability and robustness in tracking for the social infrastructure scenario. The improvements are inspired by common-sense intuition of the factors important for tracking: 1) Is a person's motion or appearance more informative in this instant? 2) Has a person's appearance changed? 3) Have people that entered the scene exited? Each of the improvements is informed by careful observation of the technical errors made by the DeepSORT algorithm. The sections below describe specific instances of error by the original DeepSORT algorithm as well as three extensions that address the questions posed above.

3.2.1.1 Dynamic Lambda

Equation 3.1 details the cost function DeepSORT uses to assign detections to tracks, with λ weighing the significance of cost associated with appearance versus cost associated with motion. In the original formulation of DeepSORT, λ is assumed to be static ranging from 0 to 1 to offer relative weighting between the two association metrics; it is hand chosen for the application. For cameras that have high periods of significant motion, λ is typically chosen to be close to 0. In applications where lighting is an issue, such as a dark night, a higher λ might be chosen to rely more on motion instead of appearance. However, these are rules of thumb, there is not a rigorous or quantitative approach for choosing λ , additionally once chosen this value is static and used for every frame in the video feed and in computing the cost for every potential association. While a static λ is suitable, a λ which dynamically updates as a function of discriminative power of appearance vectors or uncertainty in Kalman Filter projections would prove more useful for robust tracking. When a track is first created the velocity is unknown and

roughly estimated, the first few projections of the track into subsequent frames will have high uncertainty and will not be as accurate. Additionally, motion information and prediction of future object movement will be less likely when there have been successive frames where the target is occluded and not detected. During these instances the cost analysis should rely more on appearance information for re-identification. Similarly, the discriminative power of appearance vectors may diminish, such as when working with distant objects with limited pixels describing the object. In these instances the cost associated with the motion metric should have the higher impact.

To incorporate a λ able to adapt to changing scenarios the following three equations are introduced:

$$\lambda_1(i, j) = \frac{\sum_{n=1}^8 \min \left\{ \frac{\sigma_n^2}{\sigma_{n,\theta}^2}, 1 \right\}}{8} \quad (3.1)$$

$$\lambda_2(i, j) = \min \left\{ \frac{Area(det_i)}{7500}, 1 \right\} \quad (3.2)$$

$$\lambda_3(i, j) = avg(\lambda_1, \lambda_2) \quad (3.3)$$

Where σ is the set of diagonals of the error covariance matrix of the state estimate for track j at frame instance i , σ_n is the variance of each state variable of X_n , $\sigma_{n,\theta}$ is the initial variance of the state variable, and det_i is the i th detected object in the frame.

λ_1 increases with the error of the track state estimate, decreasing influence from motion estimation in times of high uncertainty in the state projections. Equation 3.4, instead tunes λ_2 with the assumed discriminative power of a detection's appearance vector. An assumption here is

the relationship between discriminative power of an image crop and the number of pixels detailing the object. To test this assumption, an image library of 500 different pedestrians was used to analyze the relationship between size of image and discriminative power of appearance vectors extracted by the MARS-trained CNN. The image library contains image crops of pedestrians taken from a variety of public surveillance cameras at the Detroit Riverfront. The library contains 2000 image crops of unique pedestrians of varying sizes. Appearance feature vectors were extracted for each image and appearance vectors of images with similar size (areas within 90-110% of each other) were scored by cosine difference. About 99,000 pairs were compared. Figure 3.1 details the distribution of cosine dissimilarities between the different IDs over average area of the compared image crops. Image pairs with more pixels were more distinguishable from each other (same vs different). At sizes around 7250 *pixels*, the average cosine difference between image pairs of different IDs was 0.54 and 0.13 for pairs of the same ID, which is an improvement from 0.48 and 0.28 at lower pixel sizes (~1500 *pixels*). While the empirical relationship between pixel size and discriminative power is not decisive, it still has utility. Equation 3.4 scales lambda with size of det_i's image crop, and stops at 1 when an image crop contains more than 7,500 pixels. Equation 3.5, is an average of λ_1 and λ_2 set by both

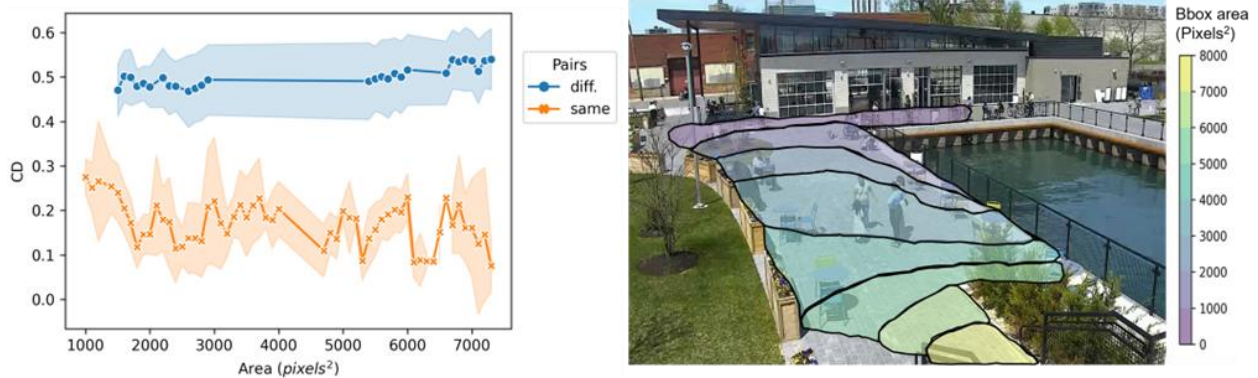


Figure 3.1 Cosine difference between crops of the same person (n=16,000) and crops of different people (n=83,000).

discriminative power of appearance and variance in Kalman Filter, and is the equation used in our dynamic lambda configuration.

3.2.1.2 Appearance feature vector gallery

When comparing a detection's appearance vector to a track, the cosine difference between the appearance vector is taken between each appearance vector in the track's gallery, which includes the 100 most recent detections assigned to the track. The smallest cosine difference is used for the appearance cost. While this method has its advantages, it opens opportunities for erroneous id switches due to extreme data points. For example, a man with a white shirt might briefly walk under a heavily shaded area and have a few images in the gallery with a dark appearing shirt, later a person with a dark shirt might match with the track due to a low cosine difference with the appearance vectors of the shaded white shirt in the other track's gallery. In this instance a new tracker ID is not created for the man in the black shirt, instead he is assigned to an existing track and would be missed by the counter. Additionally, when one pedestrian, A, crosses another, B, and partially occludes the view of B, there will be a couple frames where part of A is visible in the image crop of B. By taking the minimum cosine difference across the entire gallery, this single data point with a mixed view of both can cause A and B to switch tracker ids. A different approach is to compare the appearance vector of the detection versus the mean of appearance vectors in the track's gallery. A single cosine difference will be calculated and used for the appearance cost. This method will average out extreme data points and aims to build robustness in the tracker.

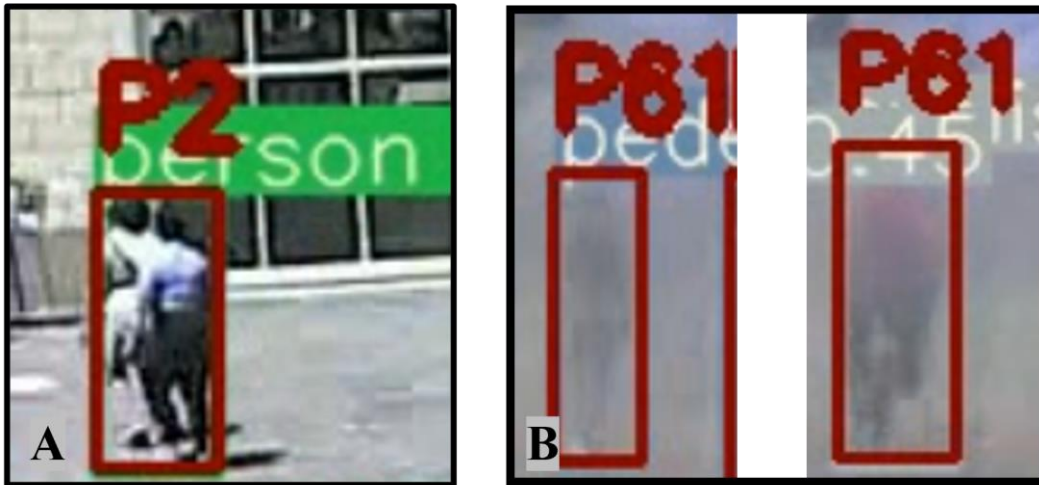


Figure 3.2 (A) Detection error – detecting merged persons as a single person. (B) Precision tracking error due to poor illumination and distance from camera – matching two different persons to same ID.

3.2.2 Mass Balancing

To further improve person counting and ensure individuals are only assigned to one tracklet ID, this work proposes an additional algorithm to remedy lingering tracking mistakes. Specifically, the "mass-balancing" algorithm aims to correct instances of a singular target generating multiple IDs. These types of mistakes are typically caused by longer periods of occlusion, where a target is hidden in the scene for multiple frames and is then assigned a new ID once again visible. The dynamic lambda extension attempts to resolve this problem by shifting the cost function to ignore the untrustworthy location from the Kalman Filter and instead rely mostly on appearance, however this approach is not perfect. The appearance of a reemerged individual may differ for a variety of reasons. For example, the target could still be partially occluded, or the detection box may feature elements of another individual if they are close together (Figure 3.2 (A)). The proposed algorithm follows a mass balance approach and operates on the assumption that all objects which enter the scene must eventually exit. Additionally, the

algorithm assumes objects do not disappear or appear in the middle of a scene. The boundaries of a scene (the entrance and exit points) dictate whether a tracklet is complete or not (Fig. 3.3). The mass balance algorithm is a heuristic that keeps a running tally of all the number of individuals that have entered and exited the scene as well as how many are still in the scene itself. If a tracklet goes unmatched before it's flagged by the algorithm as leaving the scene it enters the “lost” gallery which is a collection of all the final states of tracklets that ‘disappeared’ (tracklets that terminate while still on site). Whenever a new tracklet is generated, the algorithm checks the location, flagging the tracklet as ‘appeared’ if it originated outside of the entrance/exit boundaries. Before the new tracklet is finalized the algorithm checks if the tracklet matches to any currently in the ‘lost’ gallery based on time of origin/termination and location of origin/termination.

As with the DeepSORT tracker, the assignments are solved via a cost function and the Hungarian Sorting algorithm. The cost for a match is a function of distance and time

$$C(j_l, j_a) = \sqrt{(p_{l,f}(x) - p_{a,i}(x))^2 + (p_{l,f}(y) - p_{a,i}(y))^2} + (t_{l,f} - t_{a,i}) \quad (3.6)$$

where $p_{l,f}(x,y)$ and $p_{a,i}(x,y)$ are the center (x, y) pixel coordinates of the final position of the ‘lost’ tracklet j_l and the initial position of the ‘appeared’ tracklet, respectively. Additionally, the amount of frames between the final point in j_l and initial point in j_a is computed and added to the cost. A cost threshold is used to inhibit infeasible matches, ensuring the cost between a match is lower than the threshold before being confirmed.

Lastly, before matches are confirmed, the mass balance algorithm performs a continuity check. If tracklets in a proposed match co-existed at some point, the match is discarded.

Confirmed matches are stitched together, becoming one tracklet (Fig. 3.4). After a certain user-

defined time, tracklets in the 'lost' gallery will expire and be removed from the gallery if they are never matched. The algorithm will remove incomplete trajectories that disrupt the mass balance of the scene (accounting for the exceptions of targets already existing in the site at the start of the scene and targets currently still on site).

3.2.3 Tracker Challenges

To test the tracking framework's performance in person counting and tracking applications, such as monitoring traffic in public open spaces through surveillance cameras, three custom tracker challenges were created. Challenges, otherwise known as benchmarks, are common in the multi-object tracking (MOT) community which include fully labelled video sequences and provide

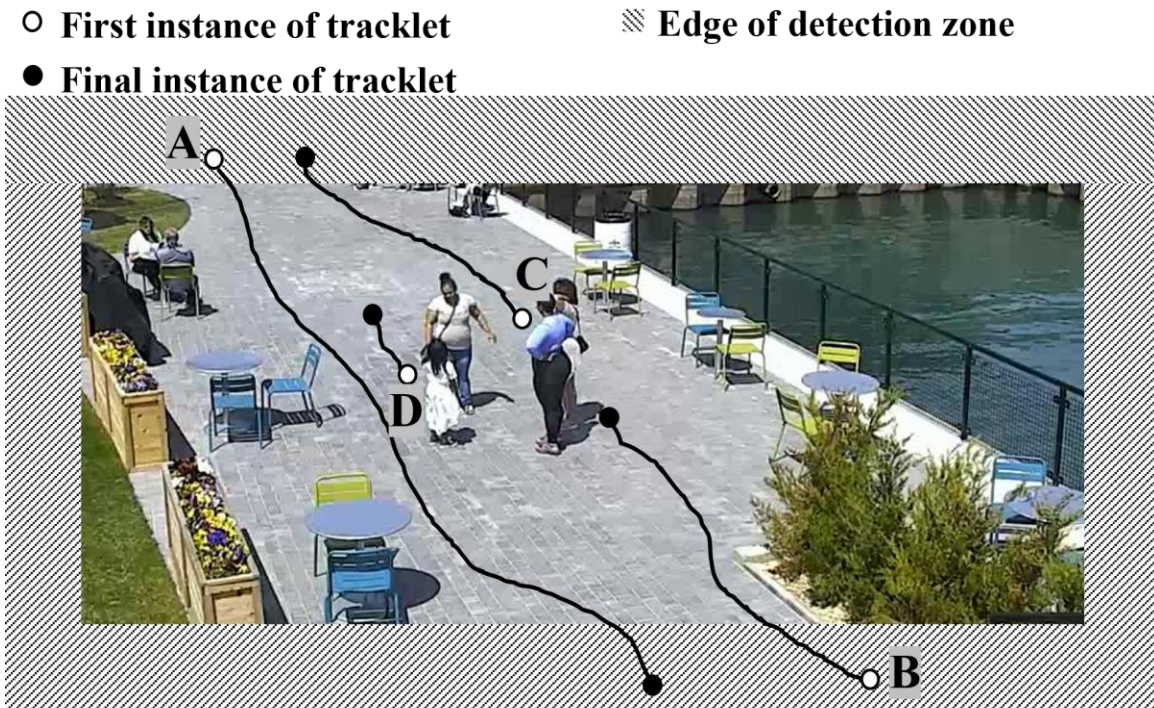


Figure 3.3 Examples of tracklets that entered and exited (A), entered but was lost (B), appeared then exited (C), and appeared then lost (D).

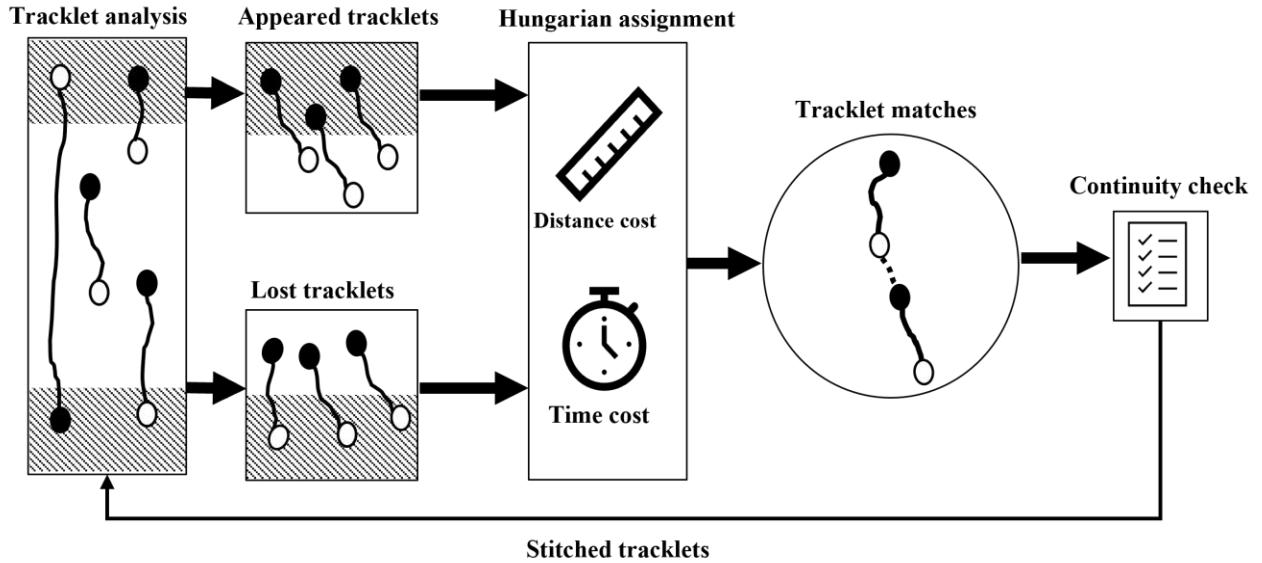


Figure 3.4 Overview of tracklet balancing and stitching.

standardized evaluation protocols for comparing and analyzing the performance of multi-object tracking algorithms. Various benchmarks are tailored for evaluating performance in specific applications, such as the KITTI tracking challenge [72], which utilizes footage from moving vehicles to evaluate tracking algorithms for autonomous vehicle applications. Other common open source challenges which include various scenes and footage of crowded public spaces, are the MOT benchmark [58], [59], [71] and VOT benchmark [73].

Three challenges were created specifically for this research, to evaluate the tracking algorithm's performance in person tracking on distant surveillance cameras. Each challenge is thirty minutes long and contains fully annotated footage from three different surveillance cameras in Detroit public open spaces (Fig. 3.5). The first two challenges contain footage from designated pedestrian walking routes: one along the Detroit river (Fig. 3.5 (C-D)), and another along a greenway running through the city (Fig. 3.5 (B)). The first challenge contains footage during a bright summer day, tracking pedestrians and cyclists as they move along a paved pathway. The second challenge also documents a similar scene of cyclists and pedestrians



Figure 3.5 Scenes from the three tracker challenges. A) Mixed-use plaza. B) Pedestrian greenway. C) Riverfront walking path -sunset. D) Riverfront walking path – night.

moving along a path but occurs during a summer night with limited visibility provided by streetlights. The lighting in the challenge gets darker throughout, with dim sunlight at the start (Fig. 3.5 (C)) that gets darker as the evening progresses (Fig. 3.5 (D)). Finally, the third challenge is the most difficult and covers a busy mixed-use plaza in a public park (Fig. 3.5 (A)). The footage comes from a sunny spring day but includes multiple occasions of long periods of occlusion (from groups socializing closely together) and multiple pose transitions (from sitting at dining table to standing) which can be challenging due to quickly changing appearance and bounding box geometry. The challenges represent practical scenarios and come from distant low resolution (720p) and low frame rate cameras (5 fps) representative of common surveillance

cameras in public open spaces. Each frame was manually reviewed and annotated with ground truth detection and tracking data. Higher Order Tracking Association (HOTA) metrics [74], which are further defined in the results section, and ground truth labels are were to analyze the performance of the output of the tracking framework and its various configurations.

3.3 Results

Four tracking algorithms were evaluated on each of the challenges. Default DeepSORT, OCSORT, ByteTrack, and DeepSORT with our extensions (dynamic λ scaling (Eq. 3.5), assignment by mean of appearance, and the use of mass-balancing). The outputs of each tracking algorithm were compared against the ground truth using the Higher Order Tracking Association (HOTA) metrics [74] for analysis: association precision AS_{Pr} , association recall AS_{Re} , and association accuracy AS_{Ac} .

$$AS_{Pr} = \frac{1}{TP} \sum_{c \in TP} \frac{|TPA(c)|}{|TPA(c)| + |FPA(c)|} \quad (3.4)$$

$$AS_{Re} = \frac{1}{TP} \sum_{c \in TP} \frac{|TPA(c)|}{|TPA(c)| + |FNA(c)|} \quad (3.5)$$

$$AS_{Ac} = \frac{AS_{Re} \times AS_{Pr}}{AS_{Re} + AS_{Pr} - AS_{Re} \times AS_{Pr}} \quad (3.6)$$

While there are various other multiple object tracking metrics [59], these HOTA metrics were chosen because they solely measure the performance of the tracking, while other metrics include measures of detection accuracy in the calculation. The HOTA metrics quantify performance by analyzing two sources of error from the tracker: false positive associations (FPA) and false

negative associations (FNA). For every true positive (TP) detection of a person, there exists c , the ground truth ID ($gtID$) and the ID assigned by the tracker ($trID$). For each c , the set of true positive associations (TPAs) is the set of TPs which have both the same $gtID$ and the same $trID$ as c . The set of FNAs includes TPs that have different $trIDs$ but the same $gtID$, while FPAs includes TPs that have the same $trID$ but belong to a different $gtIDs$. In short, the association precision tracks how well tracklets stay on a single target, low precision scores result when a $trID$ is assigned to multiple $gtIDs$ (one tracker ID being assigned to multiple people), while low association recall scores result when a $gtID$ is assigned various $trIDs$ throughout its trajectory (One person being assigned multiple tracker IDs).

Additionally, to further test the performance for pedestrian counting applications the "count" - number of unique ID's tracked - is compared against the ground truth. To generate 'counts' for each challenge, the number of unique tracklets was used (while filtering out short lived tracklets that lasted less than 15 frames) for default DeepSORT, OCSORT, and ByteTrack. The running tally of complete trajectories from the mass-balancing algorithm was used to generate the count for the extended DeepSORT proposed in this chapter. The performance for the various tracking algorithms on each of the challenges are tabulated in Table 3.1.

Table 3.1 Evaluation of the tracking performance of various tracking-by-detection algorithms applied to three custom tracking challenges using HOTA-challenge Metrics and ID count.

Method (tracking algorithm)	Challenge	Tracking Accuracy	Tracking Recall	Tracking Precision	Count (GT)	Counting Accuracy
Default DeepSORT	Busy Plaza (Day)	56%	61%	88%	82 (51)	39%
	Greenway (Day)	76%	84%	89%	50 (53)	94%
	Riverwalk (Night)	61%	76%	76%	65 (60)	92%
Extended DeepSORT	Busy Plaza (Day)	69%	80%	83%	55 (51)	92%
	Greenway (Day)	80%	84%	95%	53 (53)	100%
	Riverwalk (Night)	67%	77%	84%	62 (60)	97%
BYTETRACK	Busy Plaza (Day)	59%	64%	88%	74 (51)	55%
	Greenway (Day)	76%	79%	95%	30 (53)	57%
	Riverwalk (Night)	64%	74%	83%	63 (60)	95%
OCSORT	Busy Plaza (Day)	66%	73%	87%	86 (51)	31%
	Greenway (Day)	72%	80%	88%	53 (53)	100%
	Riverwalk (Night)	67%	84%	77%	68 (60)	87%

3.3.1 Association Recall Improvement

The most dramatic improvement from the extensions observed was the reduction of recall error. This improvement was especially notable in the tracker's performance in a busy plaza scene (challenge 3), where there were numerous occlusions caused by crossing groups. When one person is assigned a *trID* then is occluded from the camera view by a passing group and is assigned a different *trID* upon reemergence a recall error is generated. For example, during frame 3585 of tracker challenge 3, two recall errors were generated using default DeepSORT when two persons (*gtID* 27 and *gtID* 28) walking together are occluded by a passing group and then switch *trIDs* once back into view. An analysis of this example showed that while the appearance cost of matching *gtID* 27 back to itself was 0.13 as compared to 0.26 for matching to *gtID* 28 the distance costs (marred with uncertainty due to multiple frames of missing data) were 0.63 and 0.10 respectively, leaving the final costs (Eq. 3.1) as 0.38 and 0.18 - driving the

• Recall Error

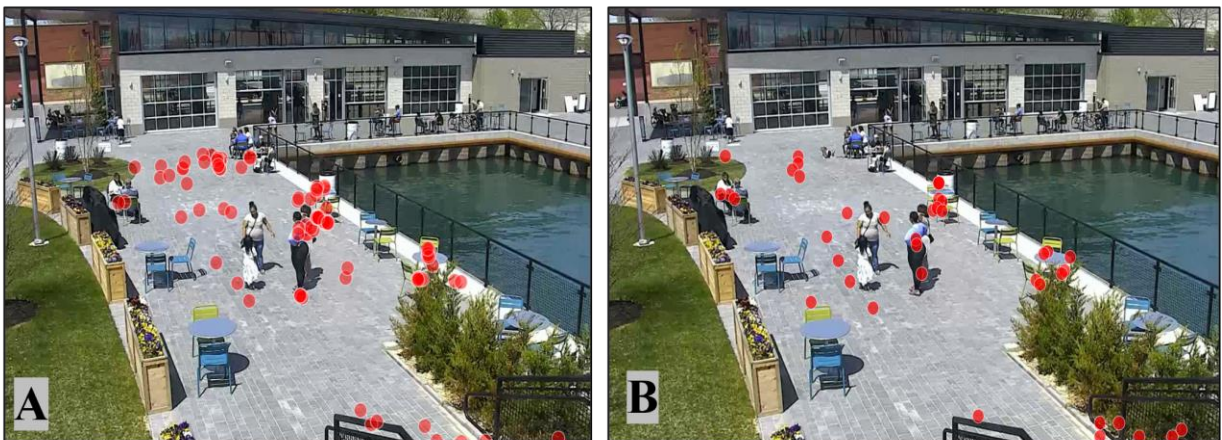


Figure 3.6 Recall errors in the mixed-use plaza challenge. Using Default DeepSORT (A) and a dynamic lambda strategy with mean of the appearance feature gallery (B).

mismatch. However, when using a dynamic lambda strategy, at this instance λ was equal to 0.90, leaving the final cost to match *gtID 27* back to itself to be 0.18 and 0.25 to match to *gtID 28* thus resolving the error. Earlier in challenge 3, in frame 602, another ID switch occurred using default DeepSORT when *gtID 2* moved in front of *gtID 1* partially blocking the person from the field of view. In this moment the detector made an error, detecting the two persons as a single person (Figure 3.2), creating an erroneous appearance feature vector including elements of both persons. The merged detection was matched to *gtID 2* (the matching in this frame is not of any importance as both were in the botched detection), adding the feature vector (with elements of *gtID 1*) in *gtID 2*'s feature gallery. In the subsequent frame, when the two were again detected separately, the appearance cost between the detection of *gtID 1* and the tracklet for *gtID 2* used the cosine difference score between the current detection of *gtID 1* and the erroneous detection of *gtID 2* which included elements of *gtID 1* because it was the minimum, ignoring the other evidence - the much higher difference scores of the other 99 vectors in the gallery - causing a mismatch. However, this specific instance and others like it are resolved when the mean is used instead of the minimum. The mass balancing further improved association recall scores by stitching together broken tracklets. Together a strategy utilizing a dynamic lambda, the mean of the appearance feature gallery, and mass-balancing improved the association recall score by 19 points over the default. Figure 3.6 plots the recall errors for challenge 3, showing the improvements from a dynamic strategy were not specific to location but were general and resolved various sources of recall error throughout the scene.

3.3.2 Association Precision Improvement

Precision errors occur when one *trID* extends over to different *gtIDs*. In general the default tracker had less precision errors, leaving less room for improvement, however, the extensions improved performance during low light conditions, improving precision by 8 points in the riverwalk night challenge (challenge 2). Figure 3.2 shows the two person crops from challenge 2 that resulted in a precision error when using default DeepSORT. When persons are dark and distant, such as in Figure 3.2 (B), the ReID CNN is unable to extract truly discriminative appearance vectors. In this specific case the person on the right entered the scene a few frames after the person on the left exited. The tracker matched the two detections as the same ID due to strong similarity between the extracted appearance feature vectors. However, when a dynamic lambda strategy was tested the different persons were not assigned to the same *trID*, as the influence from the appearance cost of the distant targets was tuned down.

3.3.3 Evaluation of specific extensions

The busy mixed-use plaza (Challenge 3) was the toughest challenge, thus it had the highest capacity for improvement. For this reason this challenge was chosen to further evaluate the impact each DeepSORT extension had on the increased performance observed in Table 3.1. For challenge three, extended DeepSORT was run three times, with each step adding one of the proposed extensions. Results are listed in Table 3.2.

Table 3.2 Evaluation of the tracking performance of DeepSORT (with various static λ s) and extended DeepSORT (with various extensions used) on a busy mixed-use plaza tracking challenge.

Method (tracking algorithm)	Challenge	Tracking Accuracy	Tracking Recall	Tracking Precision
DeepSORT ($\lambda = 0.0$)	Busy Plaza (Day)	32%	46%	57%
DeepSORT ($\lambda = 0.1$)	Busy Plaza (Day)	59%	64%	88%
DeepSORT ($\lambda = 0.2$)	Busy Plaza (Day)	54%	58%	89%
DeepSORT ($\lambda = 0.3$)	Busy Plaza (Day)	54%	58%	89%
DeepSORT ($\lambda = 0.4$)	Busy Plaza (Day)	52%	56%	88%
DeepSORT ($\lambda = 0.5$)	Busy Plaza (Day)	54%	58%	89%
DeepSORT ($\lambda = 0.6$)	Busy Plaza (Day)	56%	61%	89%
DeepSORT ($\lambda = 0.7$)	Busy Plaza (Day)	48%	53%	87%
DeepSORT ($\lambda = 0.8$)	Busy Plaza (Day)	54%	59%	86%
DeepSORT ($\lambda = 0.9$)	Busy Plaza (Day)	54%	59%	86%
DeepSORT ($\lambda = 1.0$)	Busy Plaza (Day)	52%	57%	87%
<hr/>				
Extended DeepSORT				
Dynamic λ	Busy Plaza (Day)	63%	68%	89%
Dynamic λ + mean of appearance	Busy Plaza (Day)	64%	69%	89%
Dynamic λ + mean of appearance + mass-balancing	Busy Plaza (Day)	69%	80%	83%

First, extended DeepSORT only added the dynamic lambda strategy and improved the default score by 7 points. Next, the strategy of taking the mean of the appearance feature vector gallery was implemented and increased the score by another point. Lastly, the real-time mass-balancing algorithm was implemented and increased the accuracy further by 5 points. It was expected the strategy of using the mean of appearance gallery would have the smallest impact. The extension is added to fix tracking errors due to erroneous detections, such as two people detected as one (Fig. 3.3 (A)), or other instances of brief outlier appearance changes (heavy shade). These instances are infrequent and thus have lower impact on accuracy scores.

Additionally, the original implementation of DeepSORT was evaluated using 11 different lambdas stepping from 0.0 to 1.0 at intervals of 0.1. While some static lambdas performed better than others, the highest performing static lambda (0.1) still performed worse than extended DeepSORT using a dynamic lambda approach. Even if an optimum lambda is correctly chosen for the application (which would be difficult to predict) it would be outperformed by a lambda able to adapt frame to frame. Expectedly, a static lambda of 0.0 performed much worse than all

other options - as it completely removes influence of appearance information from the cost function effectively reducing the algorithm to just SORT.

3.4 Conclusion

This chapter presented the technical details as well as analysis on testing and validation of novel extensions to DeepSORT tracking algorithm. The intellectual contributions of the chapter are:

- Enhanced tracking performance of DeepSORT via a novel approach to dynamically tune the association cost matrix based on real-time information in the image.
- Further enhanced tracking performance and person counting through real-time algorithms adopting conservation of mass principles.
- Developed diverse pedestrian traffic focused tracking challenges to robustly measure tracking and counting performance of person tracking algorithms.

In general, the proposed extensions improved DeepSORT's tracking performance by increasing stability in periods of occlusion and reducing ID switches. The diversity of the tracker challenges showcased the different strengths of the extensions. It needs to be stated the accuracy, recall, and precision scores are deflated. The metrics chosen for analysis include tracking errors only in the calculation, however, detection errors still influence the scores. The ground truth was generated via manual annotation and thus has perfect detection. Whenever there is a detection error and merged persons are detected as one, the tracker performance will be penalized as there is only one detection available to work with. Undetected persons in low-light and merged

detections of partially occluded people (Figure 3.2 A) will be counted as FNAs and FPAs, respectively. Therefore, without perfect detection, even perfect tracking would have flawed HOTA metrics.

While the greenway challenge was relatively easy for tracking, leaving little to no room for improvement, the low-light conditions of challenge 2 and the business of challenge 3 highlighted the problem areas for computer vision-based trackers and showed where the proposed extensions can improve performance. A higher lambda was needed during times of occlusion to suppress distance costs of uncertain location estimates, while a lower lambda was needed to suppress appearance costs of distant, dark, and blurry persons. It is often uneasy to predict if a scene would need a higher or lower lambda, moreover, as demonstrated this would change case by case, and why a dynamic lambda strategy is a necessary extension. Switching the appearance cost function to take a mean cosine difference over the gallery instead of the minimum also increased stability and reduced ID switches, while the mass-balancing algorithm made further stability improvements and drastically increased the counting accuracy. The mass-balance addition is especially useful in person traffic monitoring, in all three challenges the mass-balancing brought the counting accuracy to within 92%. The challenges were intentionally chosen to be representative of practical conditions (low resolution and low frame-rate), as such the framework is capable of being deployed on existing surveillance camera infrastructure. With the demonstrated performance of the extensions, the tracker can provide a lot of utility to park managers or city officials hoping to tap into existing infrastructure to monitor and analyze person traffic and trajectories throughout their spaces and will serve as the tracking module in our sensing framework (Figure 1.2).

Chapter 4. Exploring Sociability in Public Open Spaces via in-depth Interviews with Park Managers and Operators

4.1 Goals and Objectives

The sensing framework so far can reliably detect and track pedestrians, cyclists, scooters, skaters, dogs, and people sitting. Furthermore, with the mapping module, the framework is able trace how patrons move through the site and which areas they engage with. At this point, the framework is an outstanding urban sensing tool, more practical and capable than the plethora of other sensing options covered in Chapter 1.2.2, such as passive infrared sensors, Wi-Fi sniffers, and other simplified camera-based person counting systems. However, as far as complexity in activity labels and social classifications, the framework is unable to compete with human observation. Researchers are still utilizing the human eye, along with “pen and paper” to monitor and record the social health of the space and the social activities it is supporting. In Chapter 1 we review the impacts the social health of community can have on the social sustainability and resilience of a community and how social infrastructure can play a critical role in building said social health and capital.

Going back to the overarching goal of this work, we want to move away from time intensive approaches (manual observation) and instead build a framework capable of autonomous quantification of sociability in public open spaces. By doing so we hope to drastically improve the monitoring, managing, and planning of social infrastructure.

However, sociability can be a vague term, what does this mean in the context of public open spaces? Drawing upon previous research, a sociable public space has been defined as an environment that enables individuals to comfortably and securely pursue their activities while engaging in social interactions, participating in events, ceremonies, and spectacles, or merely sitting and waiting [75]–[80]. These spaces foster social and recreational endeavors, whether pursued individually or collectively, and provide individuals with a venue for bonding and gathering. In the context of this study, the sociability of a space refers to its present capacity to facilitate the formation of a social environment.

But functionally, how can we delineate activities and behaviors observed in public spaces in a manner that is useful when analyzing social performance of a space or evaluating if it's meeting the community's needs? In Chapter 1.2.1 “Sociability frameworks and taxonomies” we pull from other research to learn how historically researchers have classified and labeled social activities and behaviors in the public realm. The existing frameworks [11], [18], [19] offer a starting point and are a consistent reference and frame of view when building out the social sensing capabilities of our framework.

However, to go a step further and ensure our sensing framework's utility to those managing, operating, and investing in social infrastructure, we explored the concept of sociability in public open spaces through a series of structured in-depth interviews with those in the space. We connected with various park managers and operators across the US to understand

their thoughts on a park's role in sociability, what behaviors and activities are believed to contribute to sociability, and how might we capture and measure these activities. Additionally, subjects were questioned about data visualizations and analytics, to get a sense of the most impactful and actionable delivery forms of data collected by our sensing toolset. Specifically, our objectives for the interviews are as follows:

- 1. Identify desired outputs (activities and behaviors) from social programming.*
- 2. Identify desired outputs from park features (benches, fountains, fire pits, etc.).*
- 3. Identify common activities in public open spaces and how they are perceived to contribute to sociability.*
- 4. Identify public social behaviors perceived to contribute to sociability*
- 5. Identify physical features and structures perceived to be relevant to or help contribute to sociability*
- 6. Identify desired web-dashboard features and visualizations*
- 7. Identify key sociability related analytics*
- 8. Ascertain importance of sociability and user experience to park managers and investors*

By meeting these objectives, we will be able identify key activities and behaviors related to sociability, ascertain which are measurable or inferable through our toolset, and reconcile meaningful classifications and delineations from our respondents with existing frameworks to generate a complex fully autonomous sociability and activity labelling schema. Furthermore, from these interviews we aim to build performance indices' capable of abstracting detailed social activity reports into meaningful performance scores so stakeholders can quickly and confidently assess the impacts of social programs as well as design and management decisions.

4.2 Interview structure

The eight objectives of the interviews were used to generate seventeen questions across four sections. A majority of the questions are open ended, allowing respondents to freely answer and discuss their views, opinions and thoughts on sociability in public open spaces. In an attempt to quantify certain attitudes or perceptions of specific activities a few questions do require a numerical response. Aside from questions on a respondent's background, all questions are directly linked to a specific objective - intentionally written to ensure discussions relevant to the goals and objectives are thoroughly discussed.

The interview protocol is broken into four sections or themes. First, respondents are asked about their professional background and job duties, as well as general thoughts about sociability and its importance in the context of public open spaces. Second respondents are asked a series of questions on sociability in park spaces – what activities are considered more sociable? What are target goals of social programming? What drives sociability? Third, respondents are then asked about data in public spaces on park usage and patron behaviors – what is important or meaningful enough to measure? And fourth, respondents are asked about the presentation of data and how could a web-dashboard best be designed to deliver visualizations and analytics in a meaningful and impactful way. The interview protocol can be found in Appendix A, along with full results in Appendix B.

4.3 Results

4.3.1 Background and importance of sociability

In total there were fifteen interview subjects. The participants represented various park systems, and coalitions. Ten of the respondents were in management positions – directors of social programming, program managers, and community engagement leaders, while the other five were executive – CEOs, presidents, and executive directors. As shown in Figure 4.1, the fifteen respondents came from different areas in the US – Detroit, Bell Isle, Memphis, Saint Louis, and Oregon City.

To explore the respondents' general thoughts on the concept of sociability in the context of public open spaces – they were asked to define sociability in their own terms, read a working definition, provide comments on the definition, rank the importance of sociability on a scale, and then discuss if they believed maintaining sociability was a top priority for them maintain as part of their job description. When discussing what sociability in public open spaces means to them three common themes were discussed:


- Interacting with others and actively building community
- Being around others and cohabitating (without direct interaction)
- Inclusive spaces and welcoming atmosphere

These themes are all elements of our working definition, which can be found in Appendix A, underneath the bulleted interview objectives. In short, respondents believed sociable spaces supported community building in a safe and inclusive way. Thirteen of the fifteen respondents agreed to our definition without adding or removing any components, the other two suggested a sentence in the definition that specifically addresses inclusivity of a space. One respondent

believed friendly interactions between strangers are closely connected to the concept of sociability and should be highlighted in the definition.

When discussing the importance of sociability in park spaces and job priorities respondents were in overwhelming agreement that sociability is of utmost importance with all respondents ranking sociability as a 5 on a scale (1-5) of importance to public spaces. Additionally, ten respondents stated it could be considered their number one priority to maintain while the other five said it was in their top three priorities. Five respondents also used the words “critical” or “foundational” when describing the role of sociability in park spaces.

Affiliated Parks/Organization:

Detroit Riverfront	:	     (5)
St. Louis Arch	:	    (4)
Detroit Parks Coalition	:	  (2)
Belle Isle	:	 (1)
Huron Metro Parks	:	 (1)
Willamette Falls, Oregon City	:	 (1)
Memphis Parks Riverfront	:	 (1)

Mentioned duties/responsibilities:






















- Develop social programs :      (5)
- Secure funding :      (5)
- Develop Vision :     (4)
- Find Partners :     (4)
- Attract visitors into space :   (2)
- Assure public access :   (2)
- Hire staff :   (2)
- Data collection on users :   (2)
- Training :  (1)












Figure 4.1 Affiliated organizations/parks and listed job duties of the fifteen respondents.

4.3.2 Exploring sociability in Public open spaces

The following section of the interview focused on specific activities and behaviors perceived to contribute to sociability. Respondents were asked a variety of questions regarding “sociable” activities, desired outputs of social programming, and impacts on sociability from park assets. When discussing what impacts social programming or park assets would need to have in order to be considered successful respondents talked mostly about building community. Specifically, respondents hoped their social programming and park assets would help build

community by facilitating conversation and connectivity. For social programming, respondents wanted to create friendly and welcoming environments – with a focus on making the event and space approachable. They want the events to increase patron time on site, as longer stay times suggest the patron is happy and enjoying the space and leads to more opportunities to connect and engage with others. They want a welcoming and inviting space, so people feel safe and will seek new connections and spontaneous interactions. Similarly, respondents wanted park assets and furniture to encourage new connections and spontaneous interactions – respondents want larger tables that multiple families and groups use, they want their amenities to be shared and enjoyed with other members of the community.

Goals of park programs:

- Build community :  (8)
- Make the place approachable :  (8)
- Increase patron time on site :  (7)
- Encourage mental wellness activities :  (5)
- Drive traffic :  (3)
- Expose people to new things :  (2)
- Create spontaneous interactions :  (2)
- Outdoor education :  (2)
- Build relationship with space :  (2)
- Bring in/support local business :  (1)
- Highlight diverse cultures and minorities :  (1)

Goals of park assets:










- Facilitate conversation and connectivity :  (11)
- Increase comfort/sense of belonging :  (5)
- increase accessibility :  (5)
- Does it fit a need?
-sit/respice/shade :  (5)
- People are using it :  (4)
- Facilitate connection to nature :  (2)
- Encourage people to spend more time in space :  (2)
- Self sustaining :  (2)
- Facilitate interactions between strangers :  (1)

Figure 4.2 Goals and desired outputs of social programming and park assets.

Additionally, they want their assets to also increase time on site by ensuring needs are met – are there restrooms, shaded areas and places to rest? Is there food, and water? When people are comfortable and have their needs met, they will stay longer, build relationships with the space, and have more opportunities to connect with and meet others.

Outside of building community and providing opportunities to connect, respondents want their programming and assets to encourage physical and mental wellness. Spaces should encourage and support exercise, meditation, reflective thought, and connection to nature. The quote below, taken directly from one of the respondents during the interview, highlights common discussion points and views held by multiple respondents.

*“Let's give more people easy opportunities to be **fit to live healthy lives**. Let's give easy opportunities for people to protect **their mental health**. Let's give people, you know, easy opportunities to have fun. I think all of those are really important... I think to the extent that those activities also **encourage people to meet strangers** - Then I think that's even better.” - interview respondent.*

When asked to describe in detail a park setting with a high degree of sociability, respondents discussed most the importance of diversity in both activity and demographics. Ten out of fifteen respondents specifically mentioned “diverse activities” when describing their scene – people walking, talking, biking, playing, eating, dancing, etc. Specific actions mattered less as long as the site was able to support and cater to a range of activities. Additionally, respondents desired diversity in demographics, six respondents talked about “generational gatherings” –

parents, grandparents, and children alike all out enjoying the space together. Furthermore, diversity in race, gender, and socioeconomic status were often mentioned. Along with the diversity, respondents specifically mentioned strangers conversing, new connections, and an overall sense of joy and laughter as being evident in their scenes.

Respondents were also given a list of specific activities and social behaviors and asked to rate them on a scale of one to five, with one being not sociable at all, and five being extremely sociable. The list of rated activities can be found in Figure 4.3. Figure 4.4 shows the distribution of answers (means and standard deviations) for each activity.

1. Leisurely stroll through park space alone:
2. Leisurely stroll through park space with others:
3. Running/Jogging through park space alone:
4. Running/Jogging through park space with others:
5. Sitting on a bench alone:
6. Sitting on a bench with others:
7. Dining alone on park tables:
8. Dining with others on park tables:
9. Hanging out alone:
10. Conversing with friends:
11. Conversing with strangers:
12. Organized group play(volleyball nets, yoga in green space, etc.):
13. Biking/skating/scootering through park space alone:
14. Biking/skating/scootering through park space with others:
15. Engaging/viewing park scenery (art piece, gardens, fountain, river view, etc.) alone:
16. Engaging/viewing park scenery (art piece, gardens, fountain, river view, etc.) with others:
17. Walking pet alone:
18. Walking pet with others:
19. Engaging with park performer (listening to musician, watching magician, etc.):

Figure 4.3 List of activities respondents were given and asked to rank on a scale of 1-5 on how “sociable” they believe them to be.

In general, average scores for activities of people with others (average score of 4.0 with standard deviation of 0.7) were ranked 1.2 points higher than those same activities but conducted alone (average score of 2.8 with a standard deviation of 1.3). Additionally, scores of activities of

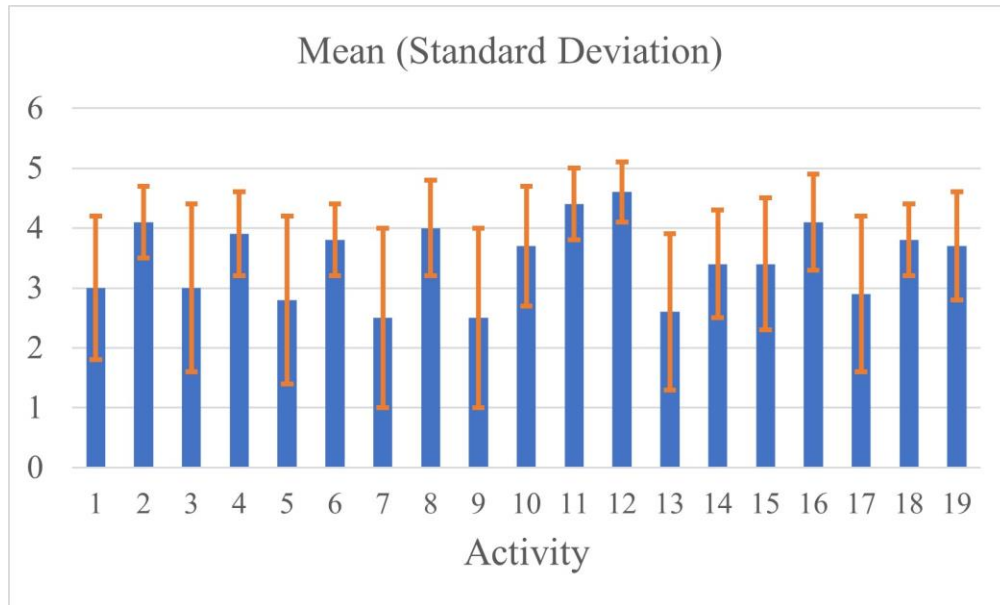


Figure 4.4 Compiled responses for each item in figure 4.3 showing the mean and standard deviation.

people with others had lower standard deviation, suggesting a greater agreement on high scores for groups. When delineating activities by movement: are people moving through space or staying? Activities keeping patrons to a specific area scored slightly higher (average score of 4.1 with 0.7 standard deviation) than those moving through the space (average score of 3.8 with 0.7 standard deviation), such as bikers or joggers.

Interestingly, while a majority of respondents generally gave higher scores to those engaging in activities with others versus alone a couple of respondents had differing opinions and fresh insights – stating they thought when people come alone to a space it’s an indicator of the space being welcoming. The idea is if someone is comfortable enough to sit in the space alone it is reflective of the space feeling safe.
















“But I think that you know the solo suggests some amount of comfort with what’s going on....it’s a reflection that the space makes them feel comfortable and safe to be able to take a walk by themselves....Is that person clearly a member of a vulnerable group in some way or other? And that that would even, to me, expand the sociability of the place” -interview respondent

These comments connect back to the idea of a space needing to feel inclusive, safe, and welcoming to help it thrive socially. While groups conversing in and engaging with a space are perceived to be more sociable than those that are alone and playing a more passive role, it is important to not dismiss single individuals when evaluating the sociability of a scene. Someone just passing through the scene, to move from point A to point B, may not contribute much to the social scene, but someone sitting in the space, reading, meditating, or simply people watching does contribute. Furthermore, in such cases it can be seen as supporting evidence that the space is welcoming and inviting.

4.3.3 Data on park use

Next, respondents were asked to think about the types of data they could gather about their park spaces and how people are using it. Discussions focused on what type of data would be most helpful or insightful. Respondents were asked what they would like to know most regarding their users and their social behaviors. Outside of patron demographics and where they were visiting from, respondents wanted to know most how long people were staying on site, where specifically they were spending their time, and if they were engaging with others.

Most desired data on park users:

- time on site/duration of use :  (11)
- origins :  (10)
- heatmap/most used locations :  (7)
- demographics/age :  (7)
- interactions with furniture :  (6)
- traffic/volume of use :  (5)
- return visits :  (4)
- are they connecting :  (3)
- public perceptions on safety of space :  (2)
- purpose of visit :  (2)
- air and sound quality :  (1)
- health of fauna :  (1)
- entrance points :  (1)
- time of day people visit :  (1)
- mode of transportation to park :  (1)

Social behaviors to track:









- spontaneous interactions with strangers :  (8)
- Are people with others? :  (7)
- sharing (sharing amenities/assets) :  (2)
- different demographics connecting :  (2)
- are people happy? :  (2)
- indicators of dissatisfaction :  (1)
- sleeping in park :  (1)
- do people help others :  (1)

Figure 4.5 Types of data on park users and usage patterns respondents would like to have.

Respondents also wanted to know the specifics of spontaneous interactions - where are these occurring? How frequently are strangers connecting? Are patrons sharing amenities? In this vein, respondents want to know if specifics of the space are facilitating new connections or pushing people together.

“Does the space and the way it's managed encourage people to interact with strangers. I'll put it that way. And particularly again with strangers who don't

look like them, whether it's age, or whether it's, you know, income.” – interview respondent.

4.3.4 Web dashboard

Lastly, respondents were asked about how they would like to engage with data about their park spaces. They were asked to describe a web-dashboard of which they could log into and interact with data – what would be the most desired features? What visualizations and analytics would be most impactful? Overwhelmingly respondents agreed that wanted something simple and straightforward, with immediately digestible charts, figures, and trends. Respondents want heatmaps of the most trafficked areas, with filters on social interactions, weather, days of week and time of day. Respondents would also like lists of most common activities, busiest time of day, and busiest areas. Respondents would like to be able to quickly learn which sites are performing well and when. One respondent mentioned a color sliding bar at the top which gives an idea of how socially active a site currently is. A majority of respondents stressed they would not want to be overwhelmed with spreadsheets, or overly complex and detailed charts. One respondent in particular compared the difference between academia and industry, suggesting academic level visualizations and charts would not be received well.

“something that's like really easy to comprehend – on a digestible level for a park manager. Right? I think that it's just a general difference between like Academia, and you know practical life is that like we need to be able to quickly understand. So infographics, or something that illustrates data in a real kind of fast and simple way” – interview respondent.

In addition to simple and interactive maps and charts, respondents desired the ability to export visualizations – to place into presentations and brochures. The overall take away from this section of questions was simplicity and practicality.

4.4 Conclusion

Overall, the series of interviews were insightful. There were a multitude of common themes discussed showing respondents agreed on various ideas surrounding sociability in public spaces, how a space might contribute to sociability, and desired social behaviors of park patrons. Respondents were also in agreement on data about their park spaces – what would be most impactful to track and how it should be delivered. Consistent ideas and themes found throughout the interviews are compelling and inform how we might extend the sensing framework to measure and quantify the sociability of a scene. The eight objectives are revisited below with brief commentary on the insights for each gleaned from the interviews. Specifically, each objective is revisited with the sensing framework and toolset in mind – how can we functionally and practically measure and address the themes brought up in the interviews?

4.4.1 Revisiting interview objectives

Objective 1: Identify desired outputs (activities and behaviors) from social programming.

Increasing sociability (via community building and increasing visitor sense of belonging) was the most commonly desired output. Functionally, respondents thought evidence of these goals being met would be:

- Increased time on site
- Groups conversing
- Spontaneous interactions with strangers

- Increased traffic

Each of these things is measurable or at least inferable and can be built into the capabilities of the toolset. Additionally, most respondents included increased wellness and participation in wellness activities as target goals. Functionally, this is more difficult for a vision-based sensing system to track. However, with site specific information this can be somewhat tracked by noting specific spaces (or specific times) with wellness activities – such as meditation zones or organized group yoga, and tracking activity in this space.

Objective 2: Identify desired outputs from park features.

With the overarching goal of building community and increasing sociability, respondents want their park assets and physical features to facilitate connection. Driving social connections was by far the most common desire respondents discussed. Respondents believed assets, such as fire pits, would increase a person's time on site, make them feel more comfortable and thus led to more social engagement with others. Additionally, respondents desired assets that helped visitors build a relationship and attachment to the physical space. Even if no new connections are formed, assets such as fire pits can provide a space for people to further develop existing relationships and grow attachment to the place. Other examples would be features that made something like fishing, bird watching, yoga, or meditation more accessible. Assets that support such activities build a sense of belonging and attachment to the area, which respondents thought were critical to increasing wellbeing of their community and park patrons.

Functionally, this is something our toolset could capture – by tracking patron time on site, interactions with specific furniture, and interactions with others, we could evaluate which assets drive connection and increase a patron's time on site.

Objective 3: Identify common activities in public open spaces and how they are perceived to relate to sociability.

When it came to specific activities, respondents overwhelmingly agreed that diversity in activities is key. When asked to describe a scene that has a current high degree of sociability, almost all participants focused on the diversity and range of activities being supported by the space. Beyond diversity in activities respondents often described group activities such as volleyball, picnic, singing and dancing. Functionally, diversity of activities and group activities is something we could track with our toolset and can be highlighted in indices', and visualizations. Activities that kept participants on the site longer were also mentioned frequently. If patrons spend more time on the site it increases the likelihood of connections and is also a reflection of the comfort of the patron and how welcome they feel.

Objective 4: Identify public social behaviors perceived to contribute to sociability.

Do people feel welcome? Are they comfortable - with each other and with the space? Any and all behaviors that provide evidence patrons feel welcome and safe were desired by respondents and perceived to contribute to sociability. Particularly interactions with strangers were repeatedly discussed in most questions on this topic. In a similar vein, sharing was a frequent discussion topic – are patrons sharing amenities? If so, this helps build community and is evident people feel welcome and are comfortable in the space. Interestingly a couple of respondents also focused heavily on the passive social category – stating they liked to see people alone reading books, working on laptops, or people watching as those type of behaviors are indicative of the comfort of the space and public perceptions of safety. Functionally, interactions

with strangers is something we can track and build into our toolset. However, careful thought must be given into how we delineate and categorize social interactions.

Special consideration must be given to the person alone, as pointed out by the respondents, being alone does not always mean the activity is non-social. In *Life Between Buildings*, Gehl used the three categories of optional, necessary, and social to categorize activities in public spaces. Gehl did incorporate examples of solitary activities like people watching as being social, suggesting there should be nuance to social classifications of solitary people. A person moving through the space alone is quite different to someone sitting and passively participating in the space's social scene.

Objective 5: Identify physical features and structures perceived to be relevant to sociability or help contribute to sociability.

Like objective 2, respondents believed anything that facilitated connection, kept people on site, or enabled patrons to build a deeper connection the space helped contribute to increasing sociability. Nature was the biggest recurring theme during these discussions. Respondents believed natural features and assets (gardens, river views, fishing spots, active bird spots) encouraged patrons to stay, actively engage with nature, and deepen bonds to the physical space. Nature that supported activities such as fishing, or bird watching were specifically mentioned often – as they bring families and friends together to enjoy activities and time spent with each other.

Additionally, physical structures which increase patron comfort and safety impact sociability – as it increases patron time on site and willingness to engage and interact with others.

Objective 6: Identify desired web dashboard features and visualizations.

Overwhelmingly patrons pushed for simplicity – charts, graphs, and maps that are immediately digestible. Functionally, the toolset should deliver figures such as heatmaps that quickly visualize where activities or certain socializations take place. The dashboard should also allow users to filter data and visualizations by time of day, weather, and date. Additionally, users need abstractions of the data to quickly show them which sites are under or over performing according to their standards and desires. The dashboard should also allow users to export visualizations so they can add them into their reports and presentations.

Objective 7: Identify key sociability-related analytics

In short, respondents want to know:

- Are people connecting?
- Where are people connecting?
- How long are people staying?
- When people stay long, where are they staying?
- What programs/assets drive traffic?
- Is the space supporting diverse activities?

Outputs from the sensing framework need to be able to answer these questions. Functionally, this means the toolset needs to be able to track social interactions, classify them on if they are new connections or not, and track patron's movements, locations, and time on site.

Objective 8: Ascertain importance of sociability and user experience to park managers and investors.

All respondents ranked sociability as a five on a scale of importance with five being the highest score. Additionally, two-thirds of respondents mentioned sociability was their number one priority with the others stating it was in their top three. Park managers and investors care deeply about the experience of their patrons and feel they have a responsibility to help usher new connections and provide a safe and welcome space. To paraphrase one of the respondents – church is no longer the go to place to meet new people and connect with others, less and less families are a part of these communities. So where do adults go to make new friends, get connected with locals, and feel a part of something bigger than their house?

The overarching goal of the interviews was to help refine our toolset and inform the design of the sociability module (Chapter 5) and the activity and social indices’ (Chapter 6). By addressing the interview objectives, we are able to think critically of the sensing framework and ensure it: captures pertinent social behaviors and activities, abstracts key sociability data through effective indices’, accurately measures space, asset, and program performance, and provides visualizations and insights in simple yet meaningful ways. The intellectual contributions of the chapter can be summarized as follows:

- The development of a structured interview template on topics surrounding sociability in the context of public open spaces and how it relates to outputs of park programming and design choices.

- An analysis of fifteen interviews, representing a subset of critical park stakeholders, on measurable social behaviors and park outcomes considered to contribute to sociability in public open spaces.

Chapter 5. Real-Time Tracking and Visualizing of Sociability in Public Open Spaces

In this chapter we introduce the sociability module (SM) to the sensing framework (Figure 1.2). An extension of the framework aimed to capture and measure activities and behaviours relevant to the social health of the scene. The SM is designed to capture evidence, as seen in research and the interviews (Chapter 4), that the space is conducive and supporting a thriving social scene.

To recap, the sensing framework is designed to ingest image feeds from cameras installed in a public space. The framework was developed using pre-installed surveillance cameras and is flexible enough to adapt to a variety of image streams and angles, often able to use already existing infrastructure. A custom trained CNN performs object detection, identifying relevant objects in the image (Chapter 2). A mapping algorithm transforms the detections from a 2D pixel coordinate space (PCS) to a 3D world coordinate system (WCS) tied to the physical dimensions of the camera's FoV (Chapter 2). A deep-learning based object tracking algorithm identifies unique objects as they move they move through the camera's field of view (FoV) and produces trajectories of each object (Chapter 3).

Using outputs from the tracker and mapping module, the SM builds upon the sensing framework analysing each unique trajectory and annotating social interactions, infrastructure interactions, movement speeds, and total time on site. The SM generates a report for each tracked person (denoted with a numerical identifier), detailing their activity (*e.g.*, stroll, exercise, Cycling, Dog-walking), social interactions (*e.g.*, enduring relationship with ID 36, fleeting interaction with ID 43), and engaged areas (*e.g.*, Interacted with fountain, sat next to fire pit). The SM is described in detail in Section 5.1: Methods. Additionally, in this section, the live web-dashboard, powered through cloud services, which delivers real-time analytics to park managers and operators is presented.

In Section 5.2: Performance the SM is evaluated on manually annotated videos (~ 2 hours of footage from 2 different cameras) of two different summer scenes along the Detroit Riverfront. Additionally, the SM is evaluated on a verification dataset – a manually annotated video of various choreographed social interactions and activities.

In Section 5.3: Results, the SM is experimentally deployed at two park sites during the summer and fall of 2022 and is used to generate sociability analytics and assess site performance. Lastly, a discussion on the SM - its capabilities and limitations, can be found in the Section 5.4: Conclusion.

Frame	T_ID	Class	Bounding box (PCS)	Position (WCS)
1	12	person	[489, 350, 35, 110]	(1200.0, 630.0)
1	13	cyclist	[685, 750, 50, 80]	(1050.0, 1230.0)
2	12	person	[483, 351, 34, 110]	(1201.0, 630.0)

Figure 5.1 Example of data output by the tracking module (with information on locations in the WCS from the mapping module).

5.1 Methods

5.1.1 Sociability module

The taxonomy used to train the object detector on the OPOS and COCO datasets largely focuses on simple human activities (*e.g.* pedestrians, cyclists, scooters, sitters, etc.)(Chapter 2). However, the SM enables the framework to detect and track more complex activities and social interactions. The tracking module (Chapter 3) outputs the track ID, base class, and bounding box coordinates (top left corner coordinates, width, and height) of each tracked individual for each frame. Additionally, the locations of each ID in the world coordinate system produced by the mapping module (Chapter 2) are incorporated in the output (Figure 5.1). The SM analyses trajectories of each unique track ID produced from the tracking algorithm to generate detailed activity reports and classify social interactions. To do so the SM analyses each track ID at three different layers: movement, social, and location (Figure 5.2).

5.1.1.1 Movement Layer

In the movement layer, track IDs are annotated at each frame with their estimated speed. The speed (in meters per second) of an individual (track ID), can be estimated at each frame by taking the Euclidean distance between the individual's current location (in world coordinates) and previous location and multiplying by the camera's frame rate. Track IDs in the pedestrian class will be marked as either jogging, walking, or stationary based on their speed, using metrics of average walking and jogging speeds [81], [82]. Cyclist and scooter class track IDs will be marked as either stationary or moving. Sitter class track IDs will only be marked as stationary.

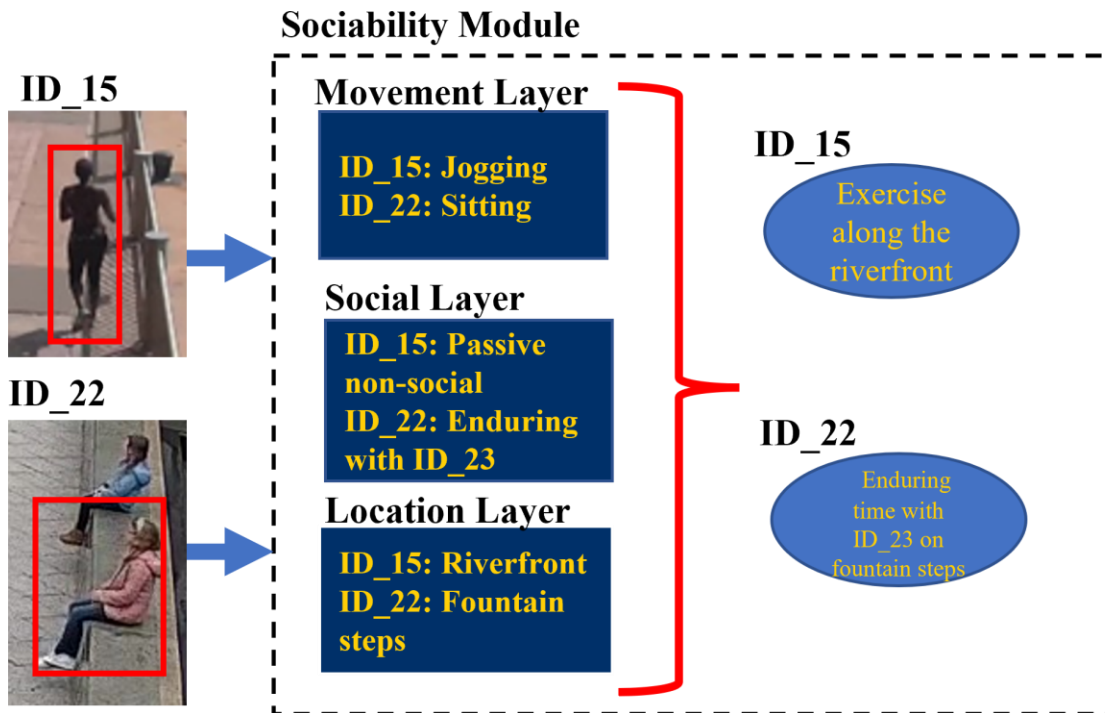


Figure 5.2 Diagram of the Sociability Module and its three layers: movement, social, and location.

A new activity will be generated for a track ID for each sustained change in movement category. For example, a pedestrian walking through a scene will be noted as a stroll through the space, but if the pedestrian stops for greater than 10 seconds the stationary interaction will be noted as an activity as well.

Additionally, in this layer, tracks are further annotated if they are walking a dog or pushing a stroller (Figure 5.3). Dogs and strollers are also tracked, just as pedestrians and cyclists. To assign a dog or stroller to an individual, the movement layer analyses each frame containing the tracked object and generates a dataframe (CPDF) of the closest pedestrian at each frame. There is a max distance, and persons outside of that distance are not considered. Once the CPDF is generated, the most frequent ID in the CPDF is assigned to the dog or stroller and their track ID is annotated with as walking a dog or pushing a stroller. If the CPDF is empty then the tracked dog or stroller or removed and it is assumed it was only a stray or a misclassification.

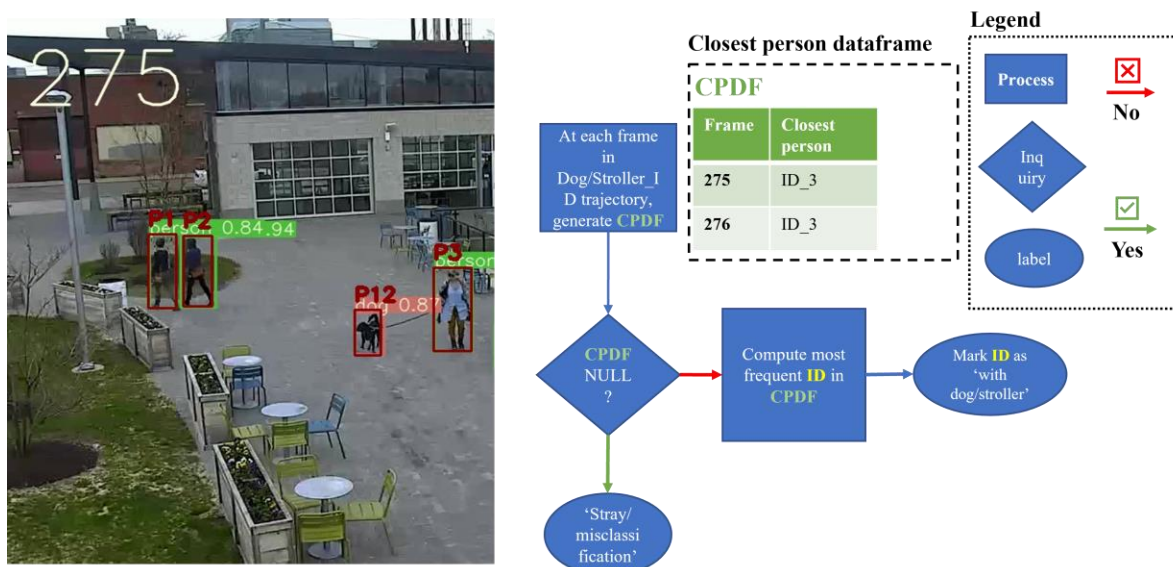


Figure 5.3 Movement layer decision tree for assigning dogs and strollers to their owners.

5.1.1.2 Social Layer

The social layer of the SM operationalizes and builds on the frameworks put forth by Mehta, Gehl, and Anderson [11], [18], [19] and discussed in Chapter 1.2.1 – Sociability frameworks and taxonomies. The SM starts with the classification schema put forth by Mehta in [19], where social interactions are categorized as either fleeting, enduring or passive. This framework aligns well with the insights pulled from the sociability interviews, ensuring ‘new connections’ are highlighted and uniquely tracked. During the interviews, interactions with strangers was repeatedly discussed and perceived by stakeholders to be evidence of welcoming space and healthy social scene. This is in agreement with other research linking the number of spontaneous interactions between community residents to sense of community, attachment to place, and perceptions of safety [6], [83], [84]. Therefore, the ‘fleeting’ (short spontaneous interactions) category is necessary and tracks number of ‘new connections’ and sharing interactions which were both, according to interviewed stakeholders, heavily desired outputs of social programming and contribute to increasing a site’s sociability.

The other two socialization categories put forth by Mehta [19]: ‘enduring’ (intimate interaction between friends and family) and ‘passive’ (people by themselves) delineate those that are alone from established social groups (friends, families). The enduring category tracks the number of families and friends that use the space to deepen bonds and spend time with others. This category is important as these interactions build pride and attachment to place as well as impact physical and social well-being and therefore is also included in the taxonomy.

If the aim of the SM is to properly categorize and track pro-social behaviours, which have been identified by past research or the sociability interviews as impacting the sociability of a site, then the ‘passive’ socialization category requires more nuance. While fleeting and enduring

interactions are inherently social activities and capture desired outputs identified in Chapter 4, passive behavior can be both social and non-social. This is apparent in Gehl's activity classifications in *Life Between Buildings* (necessary, optional, and social) [11], where certain activities that would be labelled as 'passive' in Mehta's schema are classified under 'social' while others are classified as either 'optional' or 'necessary'. A person actively choosing to be in a space around others, either to read or simply people watch is socially different than a person just moving through the space. This is also in agreement with views held by the sociability interview respondents, which believed persons spending time alone in the site is evidence of the sense of security around the place and is an indicator of the social quality of the space. Therefore the 'passive' category is subdivided into 'passive social' and 'passive non-social'. Adhering to the schema in [11], activities that would be labeled as necessary – such as walking through the space to get to and from work, are labeled as 'passive non-social'. Activities that would be labeled as social – such as people watching or sunbathing, are labeled 'passive social'.

Using this schema, the social layer of the SM analyzes each activity generated from the movement layer and infers any social interactions categorizing them into one of four categories. The social categories reflect the three socialization categories for public behavior from the Mehta framework [19], but with the passive subdivision: fleeting, enduring, passive-social, and passive non-social.

These categories (enduring, fleeting, passive social, and passive non-social) each have their own impact on community wellbeing and cohesion. For sociability classification each unique track ID goes through the process documented in Figure 5.4. First, proximity buffers are generated around the location of the unique ID, (let's call this specific ID 'T_ID') at each frame. A buffer interaction dataframe (BIDF) keeps track of other ID's that are within T_ID's proximity

buffer at each frame. Next, the social interaction dataframe (SIDF) generates a compiled list of every ID that was within the T_ID's proximity buffer along with which frames they were together. Additionally, the SIDF compares the entrance and exit trajectories of the IDs to the T_ID, and an entrance and exit score (Ent. R and Ext. R) – the ratio of entrance and exit frames of the T_ID trajectory that are spent with the IDs. To filter out interactions of strangers just passing by closely to each other, IDs in the SIDF that spent less than 15 seconds with T_ID are removed from the SIDF. If the SIDF is non-empty, the trajectory of each ID in the SIDF is compared against the trajectory of T_ID.

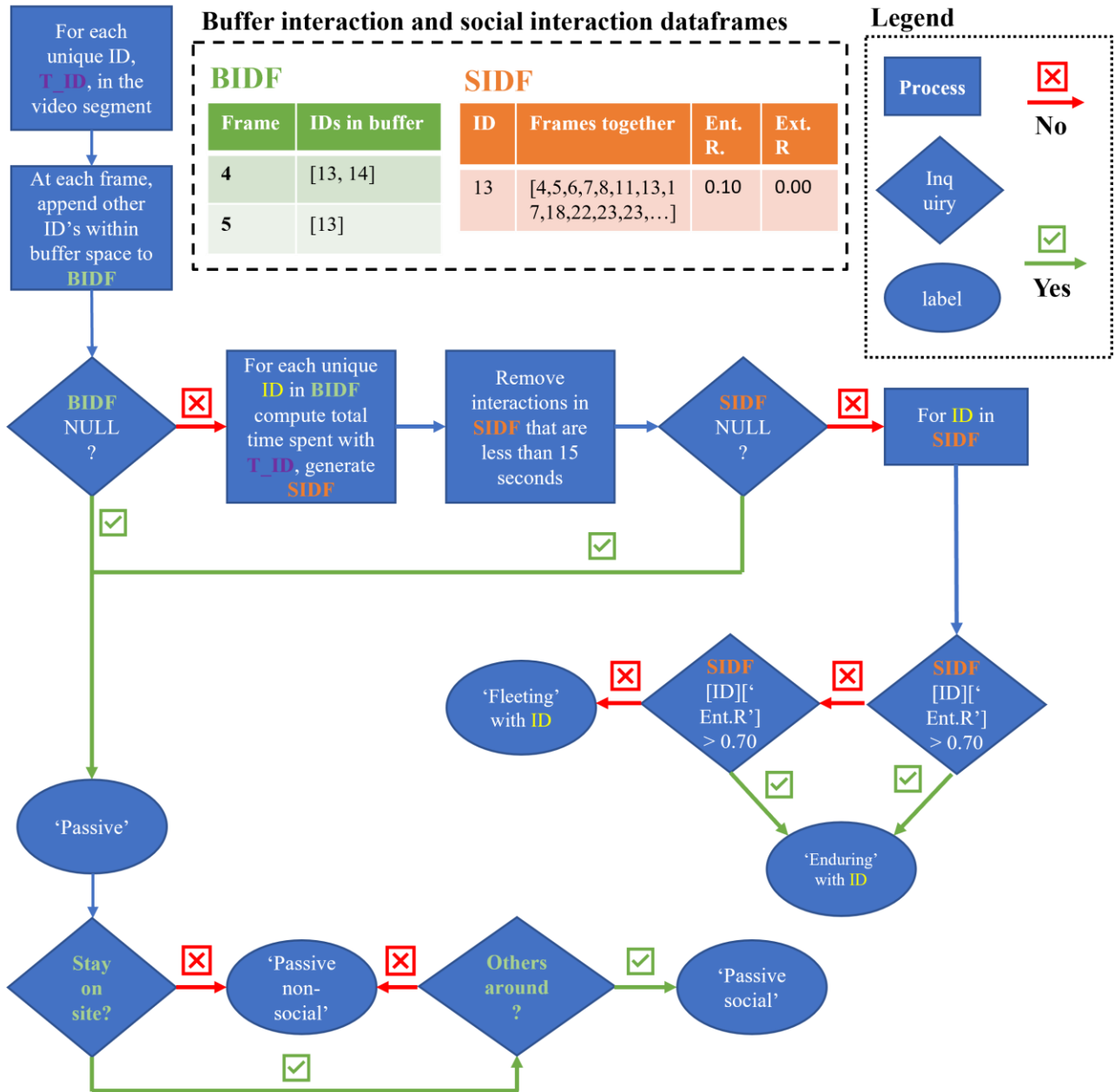


Figure 5.4 Social layer decision tree of the SM

If an ID in the SIDF either left or entered the site with T_ID (An Ent. R or Ext. R score being higher than 0.7) the social interaction is labeled as “Enduring”, otherwise the interaction is labeled as ‘fleeting’. If the SIDF is empty for a given T_ID, then a different set of inquiries is asked. The location layer of the T_ID is used to analyze the patrons time on site and asset

interactions. If a patron only walked through space or did not spend any significant time engaging in the area or with an asset, then they are labeled as “passive non-social”. However, if they did stay on site, and interact with an asset (sit on fountain steps) while there were others in the site, then they are labeled as “passive social”. Taking for example the classical bench (Figure 5.5), an example of “passive non-social” would be a pedestrian walking past a bench and through the park space without interacting with any assets or persons (Figure 5.5 A). If the pedestrian decides to sit on the bench and observes the people nearby, this may be classified as “passive social – people watching” (Figure 5.5 B). An example of “fleeting” socialization would be two people sitting on the bench at the same time for a short period but with uncorrelated arrival and departure (Figure 5.5 C). Finally, an example of “enduring” socialization (Figure 5.5 D) would be a case when two pedestrians become sitters on a bench with correlated arrival and/or departure.

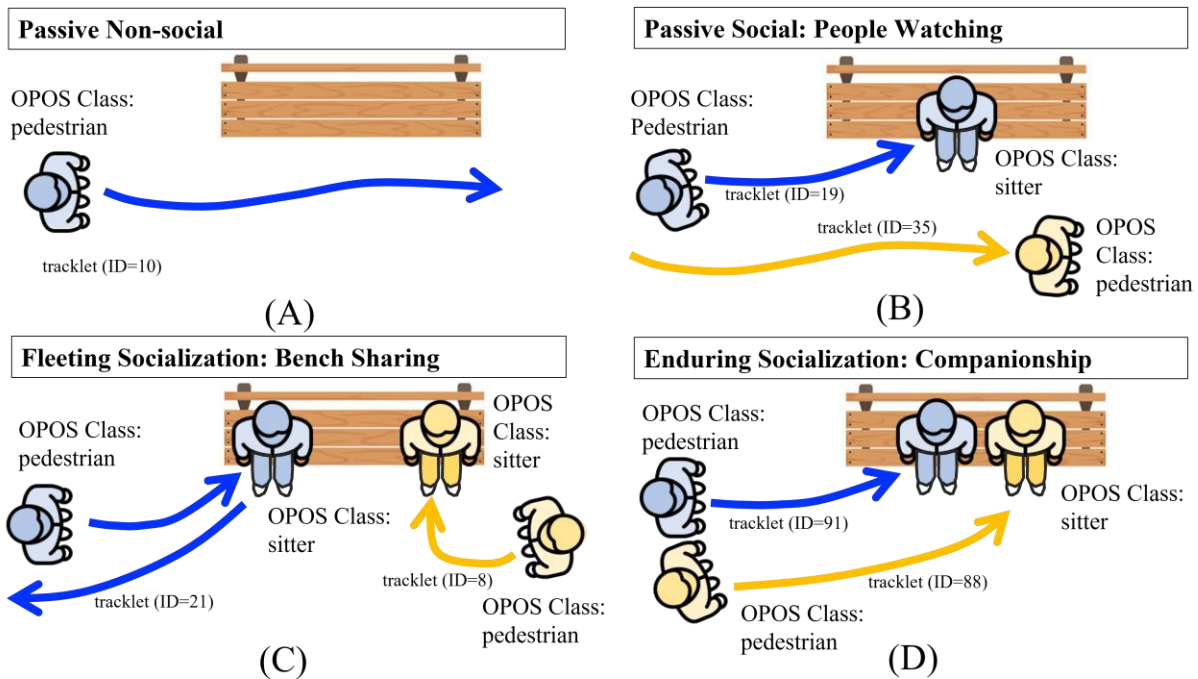


Figure 5.5 Examples of the SM socialization categories around a park bench

5.1.1.3 Location Layer

To add further context to user activities and behaviours, track IDs are also classified by their interactions and engagement with physical assets and features in the scene. For each scene, pertinent locations and infrastructure are demarcated in the image (Figure 5.6). If a camera's FoV has been calibrated (Chapter 2), demarcations are in the WCS and trajectories of tracked individuals in the WCS from the mapping module are used. However, the location layer of the SM can also use the coordinates of tracked individuals in the PCS. If a camera's FoV has not been calibrated or if the mapping is complicated or tricky due to complex camera angles and lens distortions, demarcations are made in the PCS. The location layer analyses the full trajectory of each unique tracked individual. At each frame the location layer annotates the location (*e.g.*, plaza, riverfront, street intersection) of the ID and any inferred interactions with assets (*e.g.*, fountain steps, bench, table). An activity is generated for each sustained change (longer than 15 seconds) in location. For example, an individual may walk along the riverfront and then choose to sit and rest on a bench. The location layer of the SM will annotate said individual as having been in both the riverfront path and on the bench. If the time in a certain location or an interaction with an asset does not last more than 15 seconds, the location layer will not record the activity. This 15 second requirement filters out brief encounters, where a patron may quickly closely pass a bench or cut through the dining areas without stopping or engaging in the location. In doing so, the SM can annotate activities with their engaged areas and furniture. Utilizing all three layers the SM can generate detailed activity reports for each track ID. Figure 5.6 shows the location layer annotating three tracked individuals (Ids 1, 2, and 8) by their location (plaza) and if they are interacting with the physical asset (fountain). Additionally, their WCS locations are visualized in a map of the area.

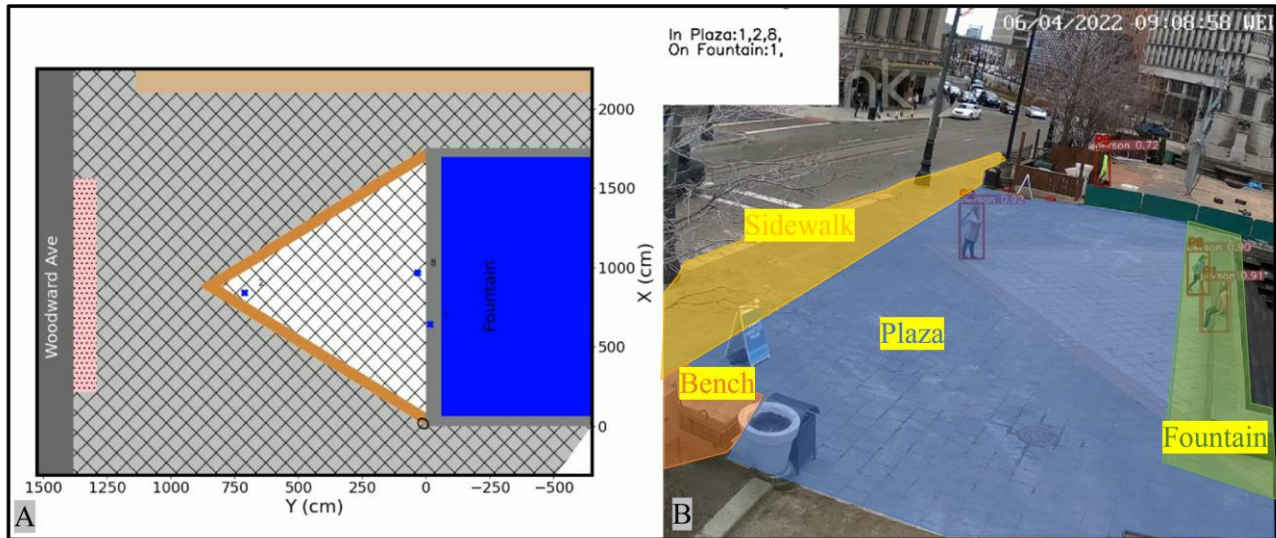


Figure 5.6 Location demarcations of a scene (B) with mapping visualization showing individual locations in the WCS of a calibrated FoV (A).

5.1.1.4 Activity Report

Utilizing all three layers, the SM can generate detailed activity reports for each tracked individual. For example, three tracked pedestrians can now be identified as a jogger exercising along the riverfront and two close people (enduring socialization) enjoying the fire pit in the plaza. An individual may have multiple activities listed in its report. The SM generates a new activity for each change in location (*e.g.*, from plaza to fountain), movement (*e.g.*, from walking to sitting), and social status (*e.g.*, from passive to fleeting). For each generated activity, the SM pulls context from all three layers to produce a label for said activity. Activities for each individual are compiled and JSON [85] formatted along with data regarding, time of visit, and time on each activity (in seconds). A typical activity report for an individual that spent time in the space might look like the example shown in figure 5.7.

5.1.2 Real-time implementation and web-dashboard

Pulling from the final section of the interviews, where respondents were questioned on how they would best like to engage with sociability data, a live web-dashboard was created.

Cloud infrastructure, powered through amazon web services (AWS), is implemented to enable real-time data visualizations and analytics which can be made accessible online to community members and stakeholders. Outputs from the tracking module (figure 5.1) and the SM (Figure 5.7) are pushed to and stored on AWS' simple storage service (S3) [86] using the AWS API. The data on AWS S3 is then visualized on an interactive web application which is hosted through AWS Elastic Beanstalk [87] - an easy-to-use service for deploying and scaling web services and applications. The dashboard is coded in Python with Dash [88], an open-source productive Python framework for building web applications. Dash is an attractive solution for developing interactive dashboards as it allows for easy integration of powerful Python data visualization packages. Two dashboards, one for DRFC park spaces and another for Campus Mauritius are shown in figure 5.8. The development of the dashboard is an on-going process, with iterative designs incorporating feedback and requests from stakeholders using the platform.


```
"Valade Park (Cam 97)": {
  "Date": {
    "8/1/22": {
      "Time": {
        "18:00": {
          "People": {
            "1": {
              "Class": "Person",
              "Activities": {
                "Stroll ": {
                  "location": "Dining area",
                  "Time on Activity": 45,
                  "Enduring": "ID 2"
                },
                "Sitting": {
                  "location": "Dining tables",
                  "Time on Activity": 500,
                  "Enduring": "ID 2"
                },
                "Stroll": {
                  "location": "Plaza",
                  "Time on Activity": 18,
                  "Enduring": "ID 2"
                },
                "Stationary": {
                  "location": "Riverfront fence",
                  "Time on Activity": 107,
                  "Enduring": "ID 2"
                },
                "Stroll ": {
                  "location": "Plaza",
                  "Time on Activity": 16,
                  "Enduring": "ID 2"
                }
              }
            }
          }
        }
      }
    }
  }
}
```

Figure 5.7 Example JSON encoded activity report for ID_1 at Valade Park on 8/1/22.

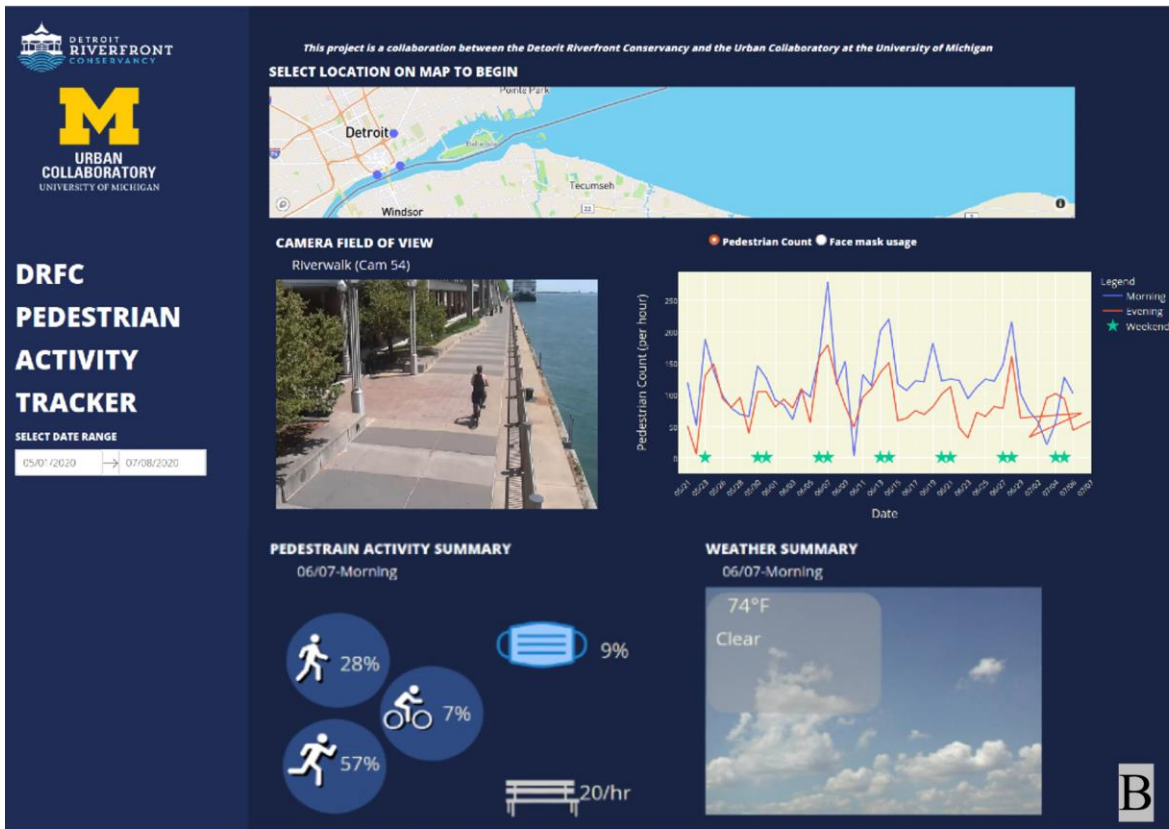
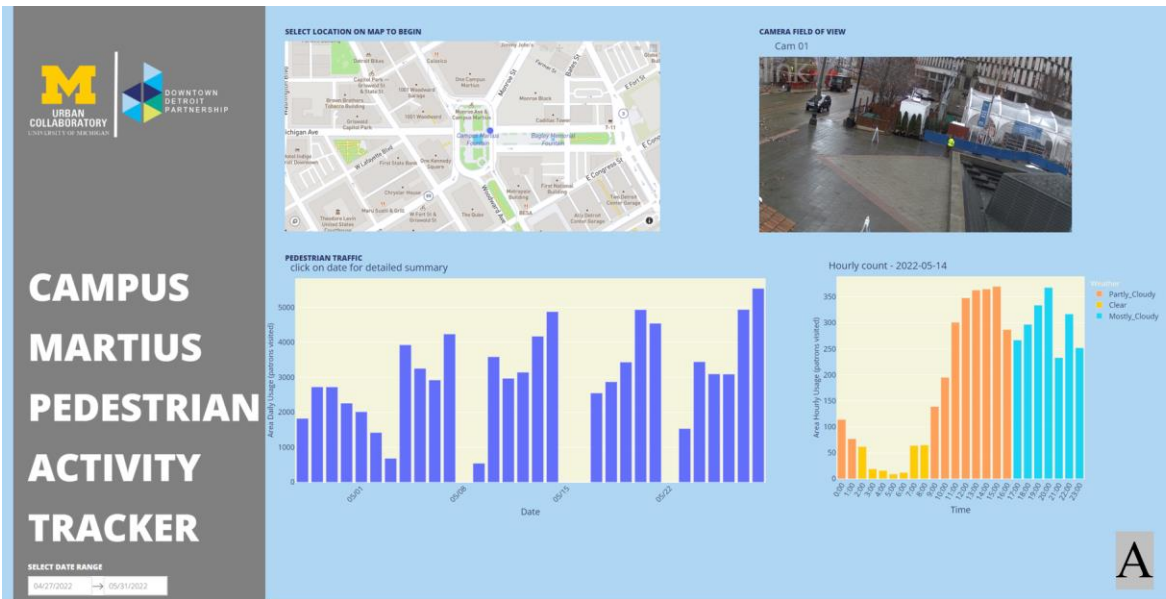


Figure 5.8 Web dashboards for Campus Martius (A) and the DRFC park spaces (B)

5.2 Performance

The performance of the sociability module was evaluated on video feeds from the DRFC parks. Additionally, due to the nature of the low occurrence of fleeting interactions, a series of controlled social interactions was recorded in the park space and used for additional performance testing.

5.2.1 Evaluation on live data

The social interactions of two hours of footage from two different cameras (one hour of footage from each) was manually annotated and classified. Cullen Plaza and Valade Park (Figure 4.9) were chosen as the two sites for testing as they have more social assets (tables, benches, riverfront views) and higher potential for interesting social interactions than some of the other views that focus on jogging paths or walkways. The results from the sociability report generated from the SM were compared against the manual classifications. The precision, recall, and accuracy of each social classification was computed using the following formulas:

$$Precision(class) = \frac{TP(class)}{TP(class) + FP(class)} \quad (5.1)$$

$$Recall(class) = \frac{TP(class)}{TP(class) + FN(class)} \quad (5.2)$$

$$Accuracy(class) = \frac{Precision(class) \times Recall(class)}{Precision(class) + Recall(class) - Precision(class) \times Recall(class)} \quad (5.3)$$

Where TP, FP, and FN are the total number of true positives, false positives, and false negatives, respectively, for a given social classification, *class*. A true positive for an enduring social

classification would be correctly identifying one ID for having an enduring interaction with another ID, with both IDs listed being the correct participants in the interaction. A false positive would be assigning an ID as having an enduring interaction with an ID it did not have an enduring interaction with. Lastly, a false negative would be classifying an ID as only having passive or fleeting interactions when it had an enduring interaction with another patron. A summary of the SM’s performance results on the two hours of footage are presented in Table 5.1 below.

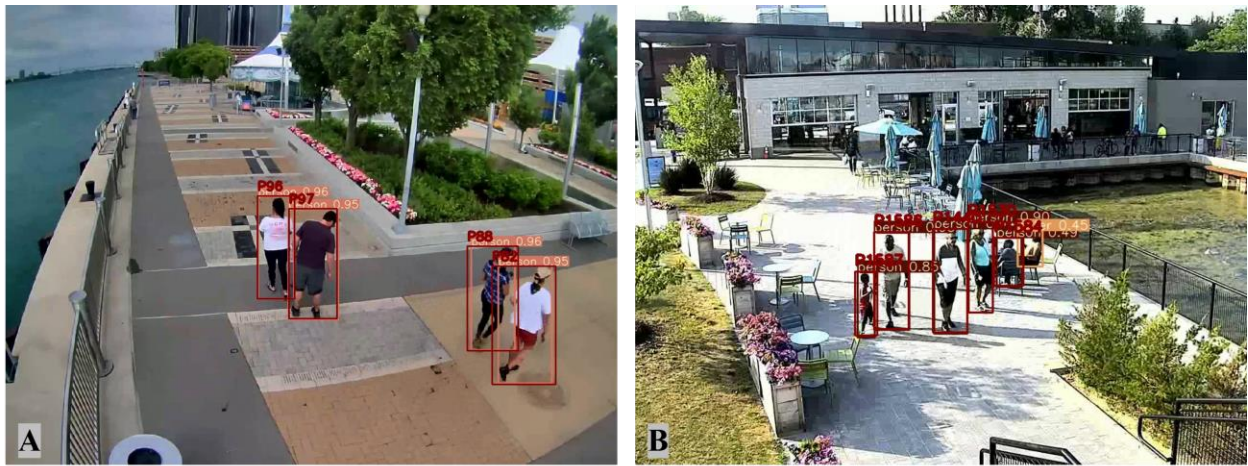


Figure 5.9 Example footage from Cullen Plaza (A), and Valade Park (B) used for evaluating SM performance.

Table 5.1 SM performance results on footage from Cullen Plaza and Valade Park

	Class		
	Enduring	Passive	Fleeting
TP	93.0	51.0	2
FP	1.0	10.0	0
FN	8.0	2.0	1
Precision	0.99	0.84	NA
Recall	0.92	0.96	NA
Accuracy	0.91	0.81	NA

For the purpose of this test passive social and passive non-social classes were merged into just a “passive” class. A total of 157 social interactions were observed, a majority of which were either enduring or passive. The fleeting count is too few to evaluate performance, and so for these reasons accuracy metrics were withheld for this class (a deeper evaluation of the fleeting class is accomplished in 5.2.2). With accuracy and recall above 90% and a precision score of 99% the performance of the SM on categorizing enduring relationships was strong. The SM also performed well classifying passive interactions – with precision, recall, and accuracy scores of 84%, 96%, and 81%, respectively. In general, most of the mistakes came from assigning IDs in enduring relationships as passive – and is why the false positives for the “passive” class are higher while the false negatives for the “enduring” class are higher. The SM algorithm is selective when deciding to assign ID’s as having a social relationship and is why the precision score for this class is quite high (99%) – when it assigns ID’s as being together it is quite certain. When there is not enough evidence of a social interaction ID’s are defaulted to the “passive” categorization, leading to the high recall score (96%) of the passive class – making it hard for passive ID’s to be misclassified as another social class. The most common mistake came from two friends or partners walking through the space together but farther apart, or from a child running too far ahead of the family. The human eye can tell the two came into the space together through body language, or that the child belongs to the family lagging behind, however, their positions are too far apart from each other for the algorithm to assume they are together. Other mistakes are generated from detection or tracking errors, on a couple of occasions a parent was carrying a child or blocking the child from the camera’s FoV. In these occasions the child was either undetected or merged in the detection of the parent, without any information on the child the SM is unable to classify the parent child relationship. There were a total of three fleeting



Figure 5.10 Examples of two observed fleeting interactions at Valade Park.

interactions observed, two were happening at the same time (Figure 4.10). In Figure 4.10, the bottom interaction (A), occurred between a security staff and a patron at a table who stopped him to ask a question – the interaction was correctly labelled as a fleeting conversation. Just above you see a couple stopping and talking to another patron at a table (B). However, the positions of the patrons are just outside the radius (by a couple of inches) the algorithm uses to infer if a social interaction took place. Instead, the SM labels the interaction as follows: the couple is an enduring conversation in the plaza while the other patron at the table is labeled passive-social and people watching. The buffer radius around individuals is tunable and can be changed so interactions like the one shown in Figure 4.10 (B) are correctly classified. However, through trial and error the current radius has proven to be most effective, increasing the radius incurs the SM to produce more false positives on enduring and fleeting interactions and hurts overall accuracy.

5.2.2 Evaluation on choreographed sequences

On a typical day, fleeting interactions occur much less frequently than enduring and passive. Curating natural fleeting interactions through manual observation, to evaluate performance of the SM is too inefficient. Additionally, more complex social maneuvers, such as an individual having two fleeting interactions with two separate groups, or patrons moving between dining tables, can be hard to find. For these reasons a set of controlled choreographies between volunteers was performed and recorded on the camera at Valade Park (Figure 4.9 (B)). The choreographies are a series of basic and more complex social interactions, with finely tuned changes (do the patrons meet on site? Leave together? Leave separately? Stop at a table? Change meeting location?) to assess the SM on specific interactions and pin point the causes of any failure. The choreographies were designed to test each aspect of the SM and ranged from simple situations like two friends walking through the space to a group of friends meeting at a table and then engaging together in spontaneous interactions with others. The interactions take place at various points within the camera's scene. Additionally various physical assets are utilized. The choreographies were conducted in view of the camera installed at Valde Park and took approximately 40 minutes. A brief description of each choreography, the involved social interactions, and notes on the SM's performance of the specific choreography are shown in Table 5.2. All social interactions during the controlled and choreographed sequences were correctly classified except two. Both misclassifications were due to detection error, where significant parts of the interaction were occluded from view of the camera. Figure 5.11, showcases one the errors. In the figure two choreographies are shown, A-C and D-F. Both choreographies are the same interaction, an enduring couple having a fleeting exchange with a passing patron, with the only change being the location of the scene. In the top choreography (A-C), ID 97 is blocked from

Table 5.2 SM results on the controlled verification dataset

Choreography	Social Categories	Results
2 persons passing by	<i>Passive</i>	Correct classification(s)
2 persons spontaneous conversation	<i>Fleeting (during conversation)</i> <i>Passive (outside conversation)</i>	Correct classification(s)
2 persons walking together	<i>Enduring</i>	Correct classification(s)
2 persons meet on site and leave together	<i>Enduring</i>	Correct classification(s)
2 persons arrive on site together and leave separately	<i>Enduring</i>	Correct classification(s)
Groups passes by individual	<i>Enduring (group)</i> <i>Passive (individual)</i>	Correct classification(s)
Group spontaneous conversation with individual	<i>Enduring (group)</i> <i>Passive (individual)</i> <i>Fleeting (group + individual conversation)</i>	Correct classifications of enduring group members Correct classification of fleeting conversation when all members are in view of camera (see figure 4.11)
Group meets up with individual	<i>Enduring</i>	Correct classification(s)
Group meets individual on site and leaves together	<i>Enduring</i>	Correct classification(s)
Group arrives together, an individual leaves separately	<i>Enduring</i>	Correct classification(s)
Group all meets on site then leaves together	<i>Enduring</i>	Correct classification(s)
Group arrives together and then all leave separately	<i>Enduring</i>	Correct classification(s)
2 persons dine together at table	<i>Enduring</i>	Correct classification(s)
2 persons meet at table and leave separately	<i>Fleeting (during table conversation)</i> <i>Passive (outside conversation)</i>	Correct classification(s)
2 persons meet at table and leave together	<i>Enduring</i>	Correct classification(s)
Individual briefly joins group at table	<i>Enduring (Group)</i> <i>Passive (Individual)</i> <i>Fleeting (Group + individual conversation)</i>	Correct classification(s)
Individual meets group at table and all leave together	<i>Enduring</i>	Correct classification of group members that arrived together Individual is incorrectly classified as fleeting relationship with group - occluded from view when group exits
Group, all meet at table and leave together	<i>Enduring</i>	Correct classification(s)
Individual A, has two spontaneous conversations with individual B and individual C	<i>Passive (individuals)</i> <i>Fleeting (during each conversation)</i>	Correct classification(s)
2 persons jogging together	<i>Enduring</i>	Correct classification(s)
Individual walking dog	<i>Passive</i>	Correct classification(s)
Individual riding bike	<i>Passive</i>	Correct classification(s)

view of the camera during the entire fleeting interaction with the couple and is undetected for the duration of their conversation. Because of this the SM misses the interaction. However, in the bottom choreography (D-F), the patrons are only partially occluded, allowing all patrons to be detected and tracked during the interaction and the fleeting interaction is correctly classified by the SM. The other error of the SM was also due to occlusion, this time a member of an enduring group was misclassified as only having a fleeting interaction. The ID in question was completely occluded from the camera by the group of friends when entering and leaving the scene. The individual is only tracked while the group sits at the table, for these reasons this SM believes the individual did not leave or enter with the group and thus only had a fleeting interaction with them. Everyone else in the group was correctly classified. Overall, the performance of the SM is

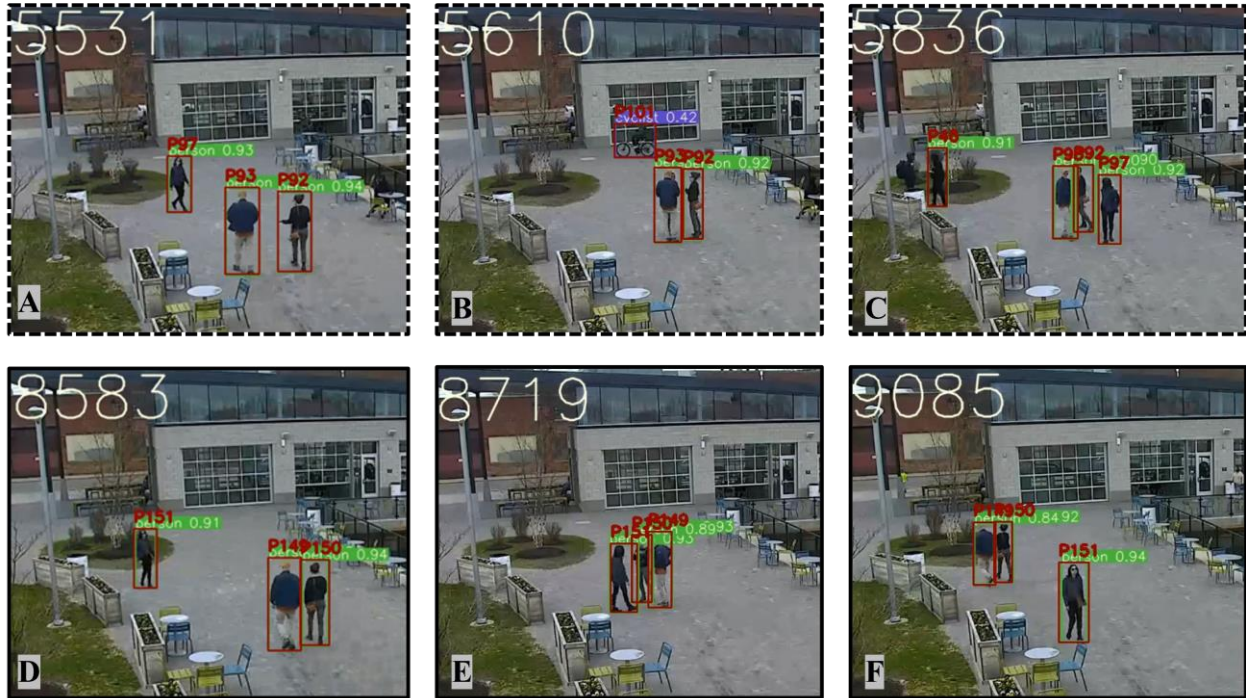


Figure 5.11 Two example choreographies from the controlled SM verification dataset. A-C are different frames from the same sequence, as well as D-F.

quite strong with few errors coming from the detection and tracking. While the tracking algorithm is robust and able to track most individuals through prolonged periods of occlusion, the SM is unable to work with the missing frames where the location of the ID is uncertain and thus will have issues when this occurs.

5.3 Experiments

The updated sensing framework with the SM was deployed and tested on various park scenes in the DRFC park spaces during summer and fall of 2022. The framework captured multiple social programming events planned by DRFC staff as well as some independent social gatherings from the community. A total of four park cameras, each at a different park space and location, were used to capture a diverse range of activities and programs. Additionally, each

space offers a unique design and purpose, allowing the tool to gain insights into the differences in behavior when different assets are available. An exercise focused beltway for joggers, walkers, and cyclists was observed at a point where it intersects with a dining space that offers food, drink, tables, and games. A separate beltway that intersects with pickleball courts and features an open field was also observed with the toolset along with a riverfront plaza featuring gardens and pleasant views of the water and city across the bridge. Lastly, a hub near the waterfront featuring plentiful dining tables with playground equipment, sand volleyball courts, and dining carts nearby was also studied.

5.3.1 Cullen Plaza

Cullen Plaza sits on the east riverfront, not only does it feature views of the water, but it also contains a carousel, garden beds, a fountain, and dining options via the Cullen Cafe and a Tiki bar. The site also hosts bike rentals and is a boarding point for Diamond Jack's river tours. The sensing framework was deployed on a camera feed coming from the southeast corner (Figure 2.8) that captures the fence along the riverfront that many rest on as they take in the view, flower beds whose trees and parapets provide shade and a place to rest, as well as the walking path to riverboat boarding point.

During the weekday, on Monday and Thursday morning, DRFC hosts the "river walkers" program – where patrons meet to walk through park spaces and along the riverfront. Additionally, many of Diamond Jack's tours depart during the weekday mornings in the summer. The framework analyzed weekend afternoons and evenings to get a sense of the social behaviors during its busiest times but also analyzed footage from weekday mornings to capture the impacts of the social programming and Diamond Jack's tours on typical weekday morning behavior.

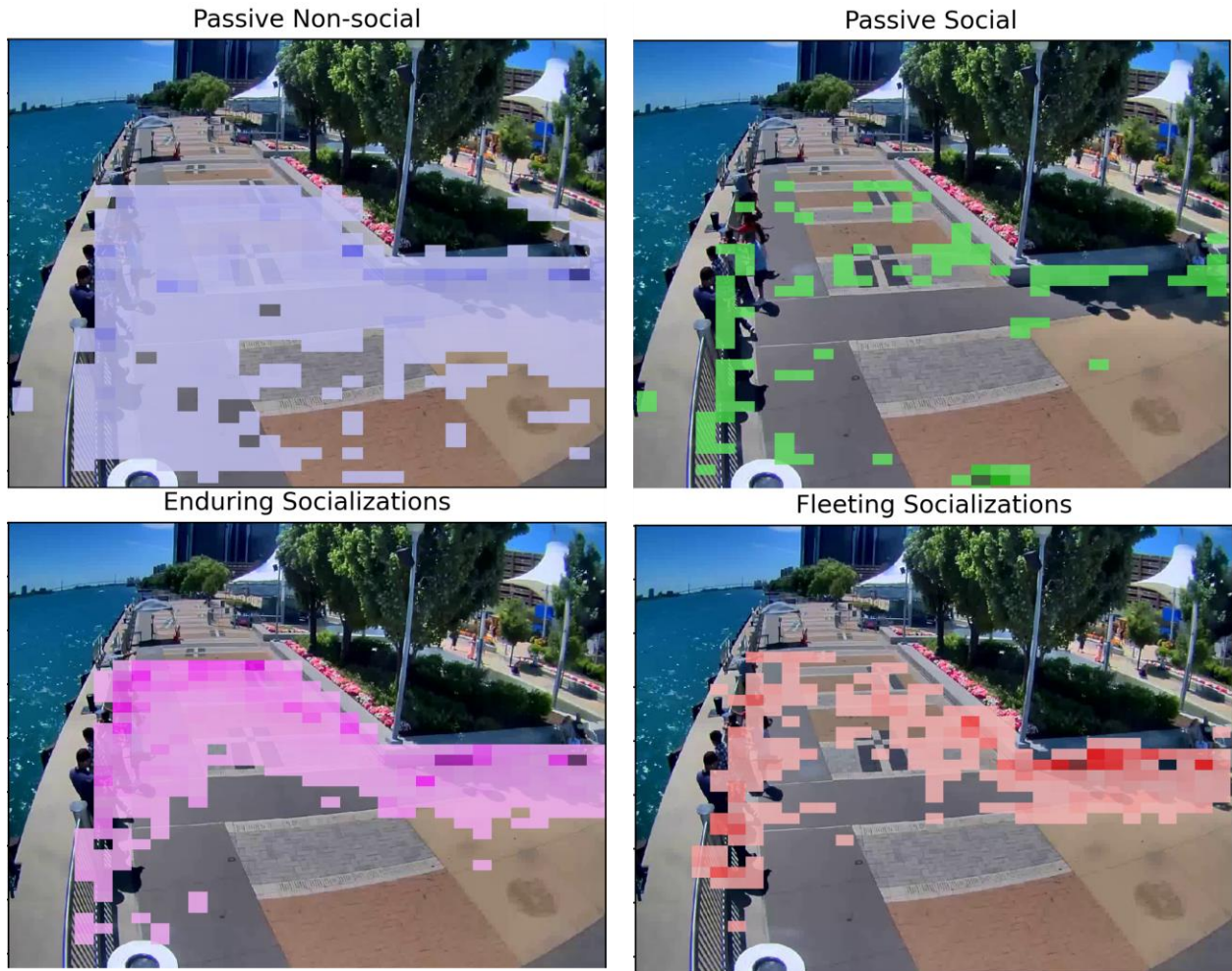


Figure 5.12 Heatmaps of social behavior at Cullen Plaza during July and August of 2022.

Figure 5.12 shows a location heatmap of the four different socialization categories. A majority of the fleeting interactions take place along the flower beds or at the riverview fence. The data shows people are more willing to have spontaneous interactions and engage with others while leaning on the fence, taking pictures of the view or while sitting on the flower bed parapets. Enduring socializations, people coming to the park with partners or friends also congregate along the parapets and riverview fence. Additionally, there is a hot spot in the plaza at the top of the detection zone – this is where the line for Diamond Jacks often extends and is likely a hotspot for enduring socializations as friends wait in line to board the river cruise. Interestingly, there is a slight difference between the fleeting and enduring interactions around

the parapet – the enduring hotspots are along the entire parapet while the fleeting only has a hot spot on the side and corner that is most often shaded. Perhaps, respite from the sun drives people to be physically closer and drives new connections and spontaneous interactions. The passive non-social category is fairly evenly distributed with some light hotspots around the parapets. The passive social category is similar to fleeting, with a majority of these interactions taking place at either the riverview fence or flower bed parapets. When people come into the place alone and choose to stay and either people watch or reflect, they choose to do so at the edges of the space, where they can sit or lean on the fence and enjoy the view. Figure 5.13 shows a breakdown of each social category by location, reinforcing the trends we see in the heatmaps.

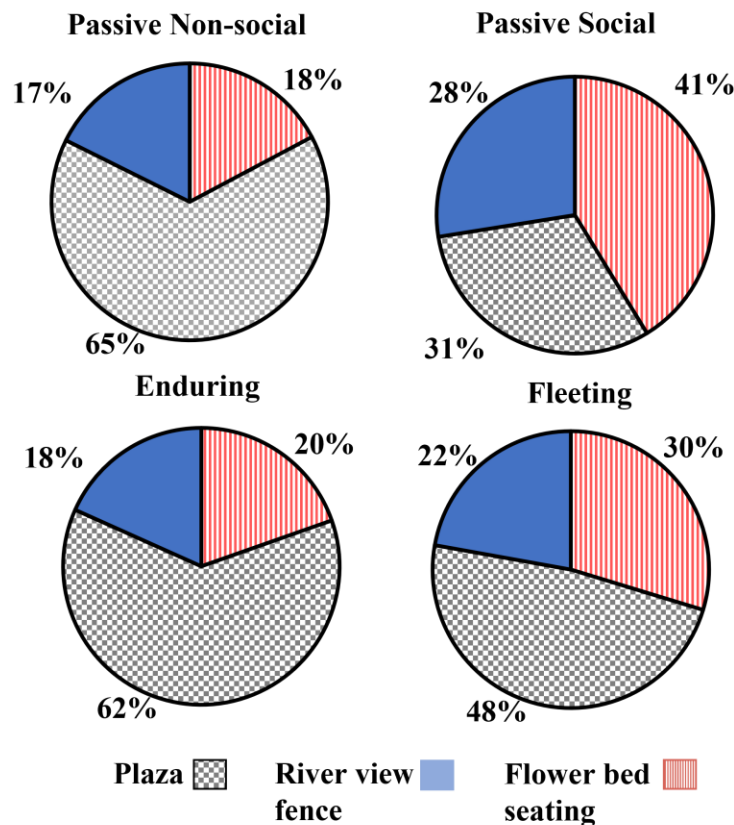


Figure 5.13 Cullen Plaza social interactions by location during July and August 2022.

More than two-thirds of “passive social” interactions take place along the edges, similarly this is also where over half of fleeting interactions take place. “Passive Non-social” and “Enduring” happen much more often in the plaza as couples and friends walk through the space or wait in line for Diamond Jacks. Data collected during weekday mornings (from 8:00am-11:00am) is divided into three categories: mornings with organized riverwalkers, morning with a Diamond Jacks tour, and normal weekday mornings. Figure 5.14 show social and traffic trends for each category. Hourly morning visitation saw a 21% increase from normal weekday mornings during ‘riverwalker’ days. Additionally, during the mornings when this program took place, patrons stayed on site for 40% longer. The breakdown of social categories between normal weekdays and ‘riverwalker’ days were mostly similar except for an 11% increase in the percentage of passive social behavior. However, there is a drastic difference on the mornings with a Diamond

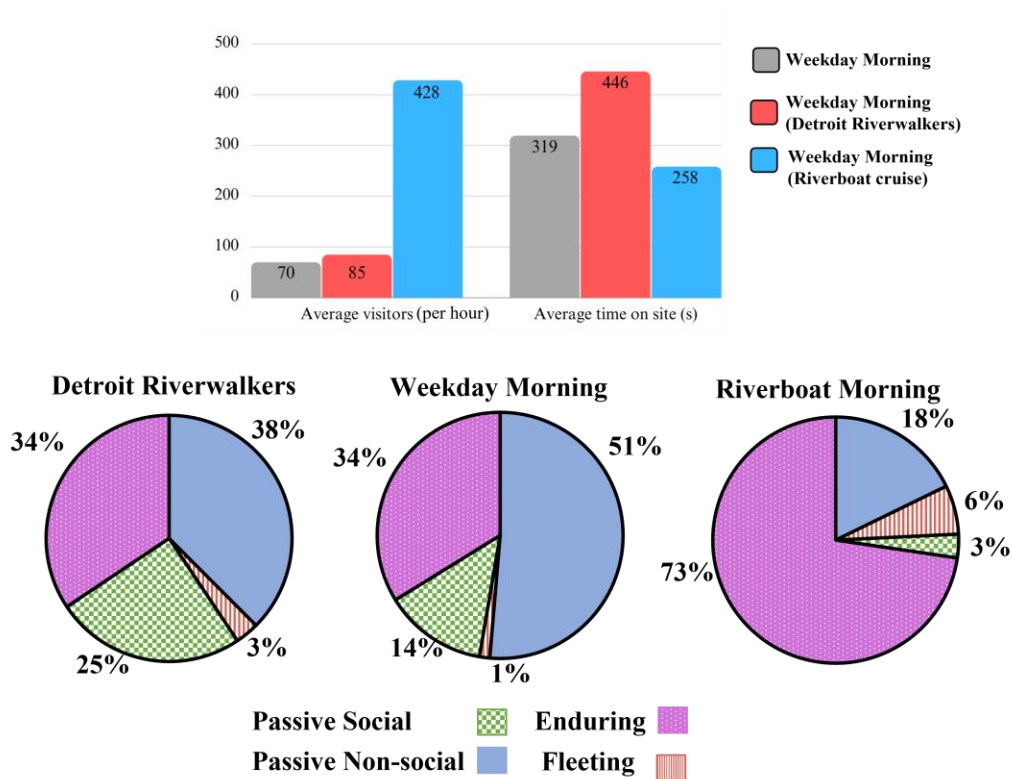


Figure 5.14 Traffic and sociability statistics of weekday mornings at Cullen Plaza in August and July 2022.

Jacks tour. During these mornings the space saw an increase of over 500% in foot traffic, however the average time spent in the space by these patrons was considerable shorter – suggesting a majority of the traffic was just passing through to board the ship. Additionally, on these mornings nearly three fourths of all guests were classified as enduring – showing the large number of partners and friends passing through to go on the river cruise together.

5.3.2 *Freight Yard*

The Freight Yard is an outdoor beer and wine garden that sits along the Dequindre Cut near the Eastern Market in Detroit. The Dequindre cut is a two-mile greenway linking the East Riverfront, Eastern Market, and several residential neighborhoods in between via paved walking and cyclist paths. The Freight Yard is composed of multiple shipping containers boasting a DJ booth along with various drink, food, and retail booths. The Freight Yard is open on weekends during the summer and fall season.

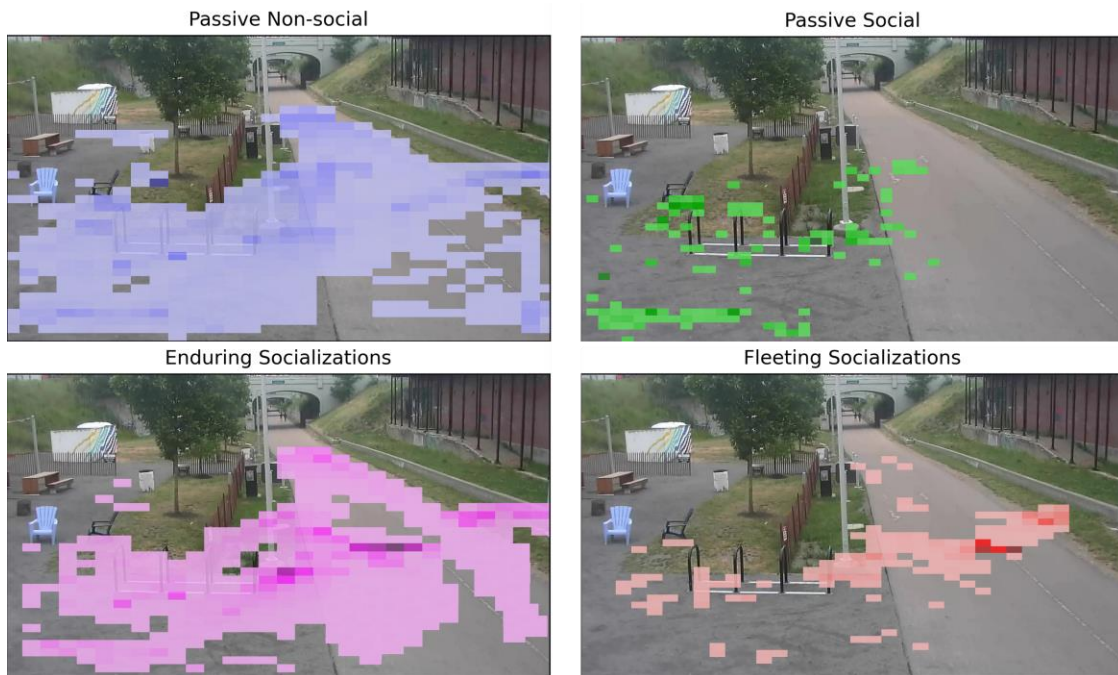


Figure 5.15 Heatmaps of social behavior at Freight Yard entrances during July and August of 2022.

The camera at the Freight Yard captures the two entrance points from the Dequindre Cut, as well as features a small section of chairs and tables inside the Freight Yard. The sensing framework analyzed weekend footage during July and August of 2022, focusing on Sunday hours around the “Summer Sundays” programming. “Summer Sundays” features live music as well as games and activities on specific advertised Sundays through the summer.

Figure 4.15 shows a location heatmap of the four different socialization categories during weekend hours at the Freight Yard. Most interestingly, you can see a majority of the fleeting interactions take place at the entrance and exit points, or along the walking path just as it passes the entrance to the Freight Yard. Once again, activities labeled as “passive social” occur along the edges and in the Adirondack chairs. The passive non-social and enduring activities happen throughout the scene with clear patterns in the enduring heatmap showing couples/groups enter the site or stay to the right of the path. Further insights into the social behaviors are gained by delineating the “Summer Sundays” from normal Sundays.

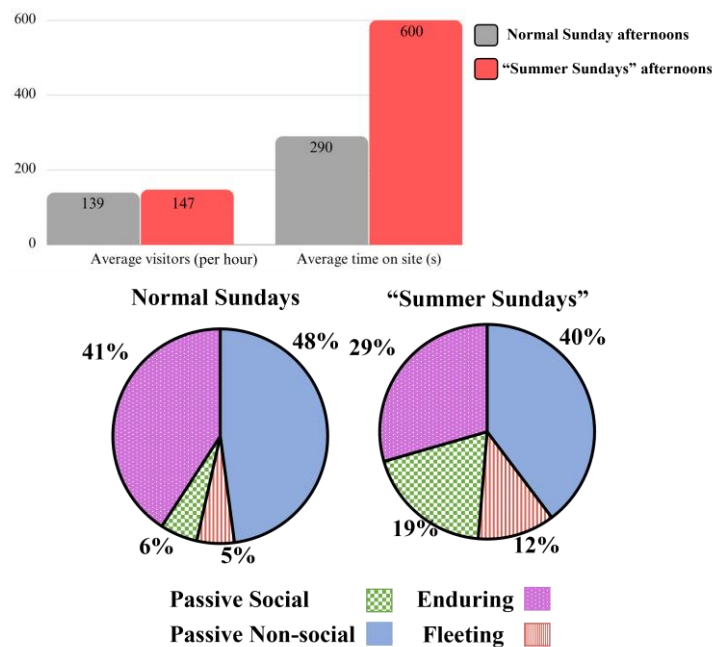


Figure 5.16 Traffic and sociability statistics of Sunday afternoons at the Freight Yard in August and July 2022.

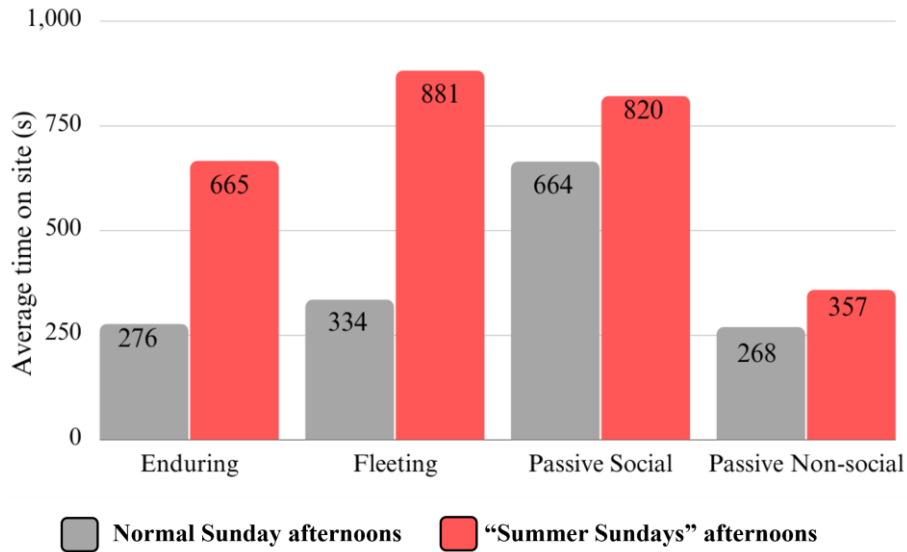


Figure 5.17 Average time on site at the Freight Yard by socialization category.

Figure 5.16 shows the programming did not increase the foot traffic however it did significantly increase the time visitors spent on site (~ 106% increase). Additionally, during the programming, those passing through the site alone were more likely to engage in social behaviors. The number of fleeting interactions was also doubled. When looking at the average time on site for the different socialization categories (Figure 5.17), it is apparent those that engaged with strangers (had a fleeting interaction) spent a considerably larger amount of time on site than others. From figure 5.17 we can also see the programming had the biggest impact on the enduring and fleeting categories with respect to time on site. Predictably, the programming had little impact with respect to time spent on site to those choosing just to pass through, however the decrease in passive non-social and increase in passive social, does suggest the programming may encourage patrons originally planning on passing through to stop, stay, and passively observe or engage with the area.

5.3.3 Archer greenway

The Dennis Archer Greenway is 1.2-mile greenway providing east siderers with safe and easy access to the Detroit Riverfront. Pickleball courts, playgrounds, open fields, benches, and tables are featured along the path. The camera chosen for the study captures a segment of the greenway with entrances to pickleball courts on the left and an open field on the right. Additionally, a few tables can be seen farther along the path. To capture the busiest hours and gain a sense of the social behaviors the evening hours of 4:00-7:00pm were recorded and analyzed each day of the week during August-November of 2022. These months captured an organized ice cream event along with multiple independent gatherings – for things like cheer practice, yoga sessions, and barbecues. Figure 5.18 shows a location heatmap for the enduring, fleeting, and passive socialization categories (passive social and non-social were merged for this camera location). The largest hotspots for enduring and fleeting interactions take place in the field just to the right of the path. This was expected as the field is often used by school cheerleading teams for practice, however, the sensing framework picked up increased activity in the field even on days without practice. Upon seeing this, the recorded footage was manually



Figure 5.18 Location heatmaps of the socialization categories at Archer greenway and examples of independent field gatherings (A-C).

reviewed to verify the accuracy of the toolchain. Investigation revealed the field was often being used by various groups for stretching, yoga, picnics, barbecues, and general play. It should be noted the amount of fleeting interactions are likely overestimated in this case, as large distant groups pose a difficult tracking problem, and many enduring relationships will be falsely labeled as enduring, as it's too challenging to infer the individual relationships of members of groups as large as 100, the accuracy is not expected to be as strong during these events as it was on the tracking challenges – which were representative of typical day to day traffic. Nevertheless, the estimates are still insightful and do inform that large gatherings often took place in the field.

In addition to highlighting the importance of the field to the social landscape of the scene the heatmaps show that only those participating in either fleeting or enduring interactions utilized the tables at the end of the path. It is also important to note that almost all patrons labeled as passive stuck to the walking path and did not cross into the field. The fleeting interaction heatmap also tells a story on the importance of the field, as the only interactions on the walkway are directly in front of the most used part of the field, suggesting patrons are stopping as they pass by and engaging with those in the field. At least two instances of this was manually verified when quickly reviewing the footage to investigate the spikes in attendance on non-cheer days.

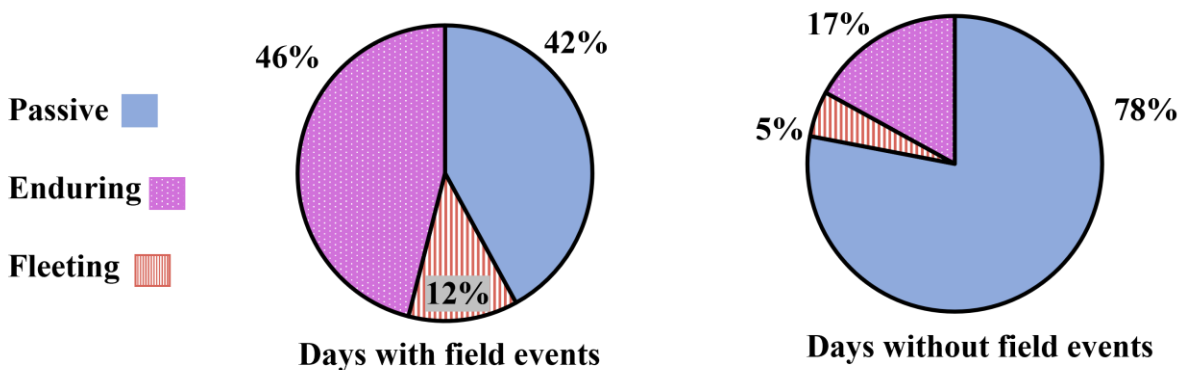


Figure 5.19 Sociability stats at teh Archer greenway on days with and without field events.

The data was parsed between days that had a large field gathering, organized by the DRFC or organic, and the days that did not. Figure 4.19 gives a breakdown of the number by social category. We see a dramatic increase in the number of enduring and fleeting interactions on the days that feature some event in the field. It is clear through these figures and heatmaps the role the field has in enabling the social scene of the space.

In addition to the independent cheer practices, yoga groups, and barbecues, the DRFC organized an ice cream event in early August in attempt to generate more traffic to the area. The DRFC also sent out mailers to the surrounding neighborhoods advertising the space in the first week of September. Figure 5.20 shows a daily count (from 4:00-7:00 PM) of patrons during the late summer and fall months, with a breakdown of the locations visited by the patrons. After isolating foot traffic from the field events, by considering traffic only passing through the walkway, we learn the ice cream event generated 52% more traffic in the 14 days after the event when compared to the 14 days prior. However, the mailers did not have the same impact, increasing traffic by only 10% in the weeks after.

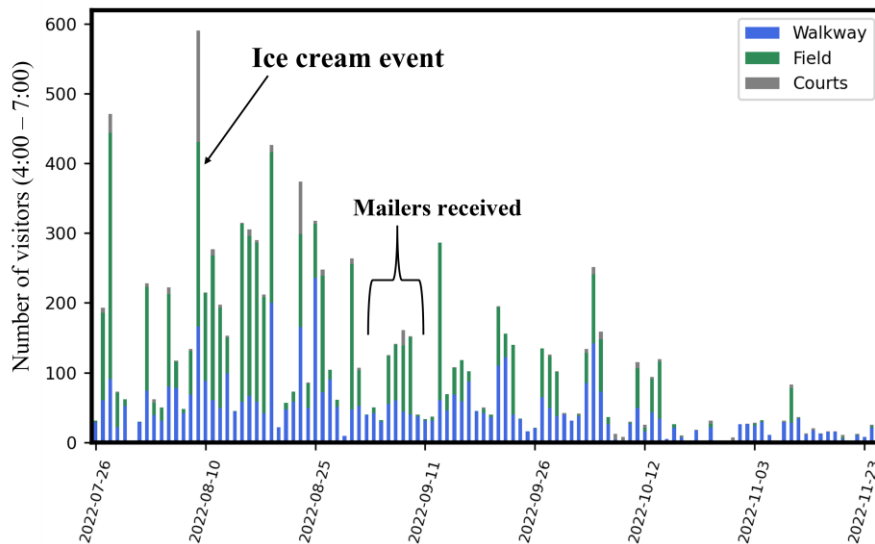


Figure 5.20 Daily visitation by location at the Archer greenway during August-November of 2022.

5.3.4 Valade Park

Valade Park sits along the riverfront and features a multitude of assets for patrons to enjoy. The site features a café, playground, children’s musical garden, beach volleyball court with chairs, as well as Bob’s barge – a floating bar open in the summers. The camera chosen for the study captures a plaza or the “central hub” of the area. In view are multiple dining tables and chairs as well as the path that runs between them connecting the café and playground to the beach volleyball court and riverfront. To capture the busiest times as well other park programs, image feeds from the camera were processed every evening from 6:00pm-9:00pm as well as Wednesdays from 9:00am-3:00pm during July and August of 2022. Wednesdays were chosen for additional analysis due to the “Work from the park” programming that ran a few Wednesdays during the summer. The program incentivized people to get outside and work remotely from the park, by offering free WiFi and free vouchers for a stationed coffee and tea food truck.

Figure 5.21 shows location heatmaps of the four socialization categories during a weekday afternoon (A-D) and a weekday evening (E-H). The chosen weekday was a normal Wednesday without park programming. The social traffic varied greatly between the two time frames: 1:00pm-3:00pm and 7:00om – 9:00pm. In the afternoon, there were zero fleeting or passive social interactions, most were passive non-social. However, the social scene did awaken in the evening, with the dining tables driving both the fleeting and passive social interactions. With patrons choosing sit and rest at the tables and watch others pass by and some choosing to interact and converse with strangers. Again, the passive social interactions occur on the edges of the scene, mainly at the dining tables or fence along the water. Enduring couples, friends, and groups, largely pass through the site - likely headed to the beach area, bar, or riverfront view or they stop to use the tables.

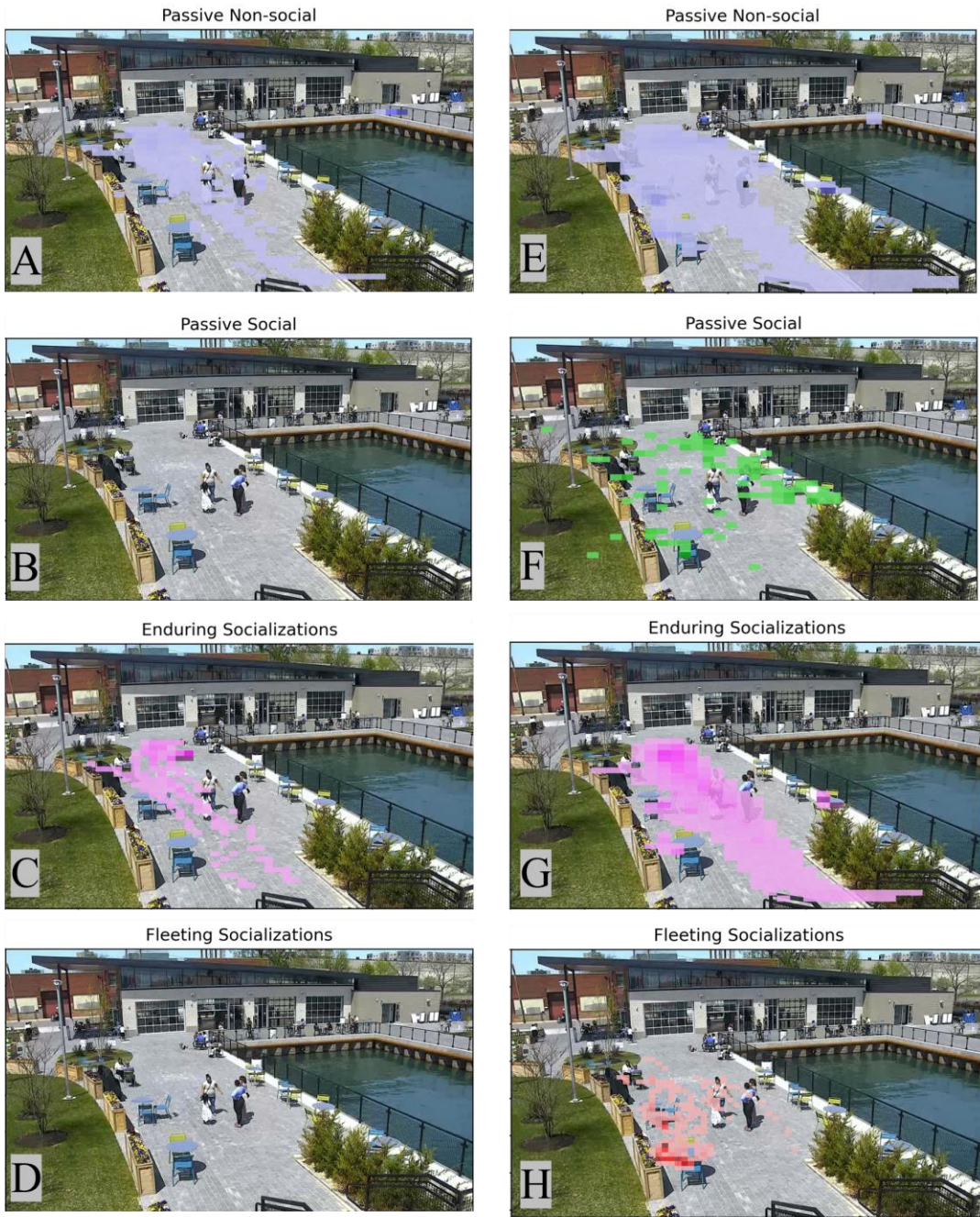


Figure 5.21 Sociability heatmaps at Valde Park during a summer Wednesday afternoon (A-D), and a summer Wednesday evening (E-H).

When comparing the evenings of weekends to weekdays the breakdown of social activity stayed extremely similar even though the visitation numbers are much higher on the weekend – an average of 566 patrons from 6:00pm-9:00pm versus an average of 161 patrons. Enduring,

fleeting, passive non-social, and passive social are 28%, 4%, 55%, and 13% of the interactions of a weekday evening respectively, while weekend evening interactions are composed of 29%, 4%, 12%, and 55% respectively.

The “work from park” programming did not have an impact on attendance, with an actual decrease in park visitation during these days. However, it should be noted both “work from the park” days captured were some of the hottest days of the summer with highs of 92 and 95 degrees Fahrenheit. Despite this heat, the “work from park” days did see a significant increase in average time spent on site. Additionally, when compared to other normal Wednesdays the dining table use increased from 33% to 50%.

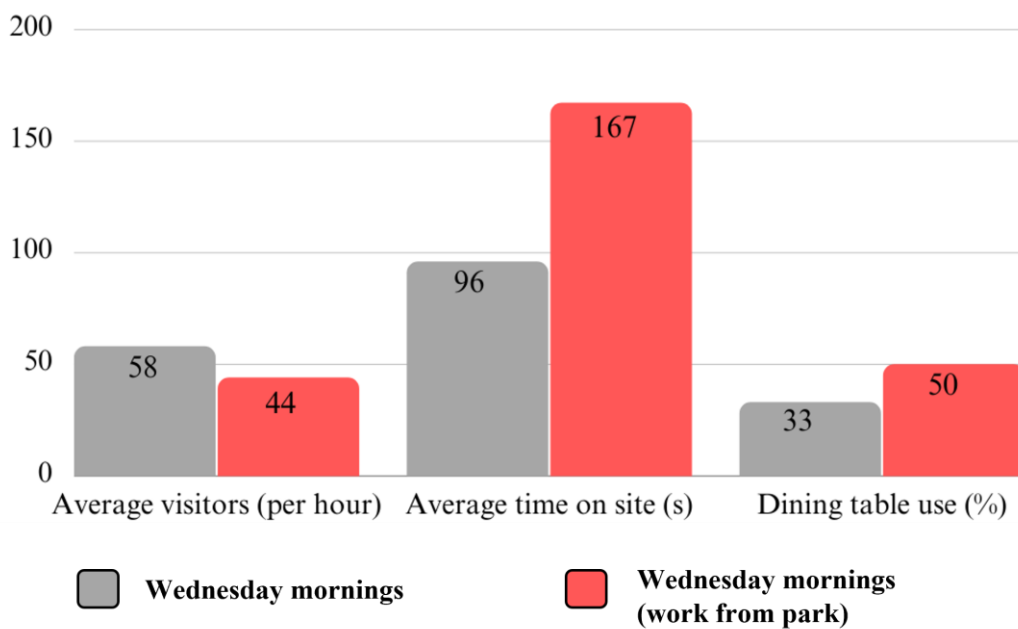


Figure 5.22 Visitation statistics for Wednesday mornings at Valade Park during June and July of 2022.

5.3.5 Real-time server

A GPU server, with an RTX 2080 (8 GB VRAM) was installed in the security building that services the DRFC park spaces. The server has access to the surveillance footage video streams via RTSP and can run the detection, tracking, and mapping modules at around 10 frames per second. The server is capable of processing two concurrent streams 24/7. Tracking data is organized by the hour. At the end of each hour the dataframe, (figure 4.1) produced by the tracker, is pushed to AWS S3 and analytics and visualizations are updated. Sociability data is also organized by hour; however, the SM only runs at night, compiling the SM JSON structure (Figure 5.7) for the entire day before pushing the data to AWS S3. The real-time server was deployed on streams at Valade Park during 2022 and on streams from Archer greenway during 2023.

5.4 Conclusion

In addition to a novel social classification schema, this chapter, to the author's knowledge, for the first time presents a fully autonomous schema for observing and classifying social behaviors in the public realm. As such the intellectual contributions are:

- The development of a novel schema for classifying social behavior in the public realm which builds on existing sociology frameworks and leverages insights gleaned from the stakeholder interviews in chapter 4.
- A fully autonomous system for monitoring, classifying, and mapping social behaviors of persons in the public realm.

Bluntly, the SM is a powerful addition to the sensing framework. By leveraging location, movement, and social interaction information the SM is able to provide complex activity labels

and provide deeper insights into the usage patterns of observed park spaces. The shortfalls of the SM are largely induced from errors of the tracker or detector. Even so, when compared against manually annotated ground truth datasets, the SM was able to correctly label and infer social relationships and dynamics with a high degree of accuracy.

The outputs of the SM enable park managers and operators to make informed management and operation decisions. Through visualizations of social behaviors, park stakeholders can identify which furniture and assets are driving interaction. Additionally, statistics on social interactions, time on site, and activity breakdowns, give deeper insights into the impacts of various park programming.

To test the impacts of the framework it was deployed at four different sites throughout the DRFC park spaces. Reports generated by the SM for each site led to various insights on park and program performance that would not be possible with other simple counting measures. For instance, some of the programs had little to no effect on traffic volume in the area. However, at each site observed, the programming significantly increased the time patrons spent on site during the program regardless if it increased volume of traffic or not. Additionally, during most programs, patrons were more social, engaged with others and formed new connections – showing the programs were successful in engaging the patrons and providing activities and opportunities to connect and build relationships. However, if the only data available was traffic counts, the programs might seem like poor investments, when instead the SM shows the programs are engaging patrons and driving relationships, they perhaps just need more advertising or exposure to drive traffic.

Data from the experiments also brought insight to the social behaviors of patrons visiting the site. Those coming alone but wishing to stay and passively participate gravitate to the edges

of the space, where they can feel safe and comfortable yet still apart of the scene. Additionally, we were able to see what furniture drives fleeting interactions, often it's a shaded resting spot or a place with a view where visitors may ask another to take their picture. When deployed on more sites and for greater periods of time the SM can provide even deeper insights and tell more compelling stories, informing park stakeholders on where best to place furniture, where to advertise for specific programs, and when to host certain activities or programs.

Chapter 6. Quantifying Sociability in Public Open Spaces

In previous chapters we have detected persons, labelled their base activity (*e.g.*, ‘sitting’, ‘cycling’, ‘jogging’), tracked them through the camera’s field of view, mapped their movements to a 3D world coordinate system, and have further classified their social activities (*e.g.*, ‘enduring stroll through plaza’). By doing so we generate robust data for each tracked person which may be used for deep analysis of trends of park spaces. However, the data may be too complex or verbose for park managers looking to quickly assess the performance of various park assets or social programs. Especially when looking at data over a period of months or years and across a multitude of sites. To ensure a significant time investment by an engineer or data analyst is not needed to extract powerful information from data output from our toolset, an additional module is needed. In this chapter, to address this need, we derive two indices’ to numerically score the performance of a site. Using insights gained from the structured interviews presented in chapter 4, we identify measurable desired outputs from social programming and park assets, and build indices’ which capture the extent to which these desired outputs are being met.

The indices' are presented and explained in Section 6.1: Methods. In Section 6.2: Experiments, the indices' are integrated with the sensing framework and deployed at three different sites in the Detroit Riverfront park spaces. Results are then presented and discussed for each site. Lastly, in Section 6.3: Conclusion, the utility of the indices'' is discussed as well as alternate forms and approaches for fine tuning of the indices'.

6.1 Methods

In chapter 4, pulling from the structured interviews, we expanded the definition of sociability in the context of public open spaces, and identified desired (and measurable) outputs from social programming and design interventions that would contribute to increasing a space's sociability. In brief, the respondents (an array of 15 persons heavily involved in public park management and operation) believed increasing traffic, time spent on site, time spent with others, and utilization of social areas could significantly increase the sociability of a site. Furthermore, respondents believed indications of sociability encompass an environment that facilitates a diverse range of activities, spontaneous interactions between strangers, and persons sharing assets.

Two indices' are used to capture the performance of the site and assess the degree to which these desired outcomes are being met. First, an activity index (AI) is used to score physical activity in the space. Are people present? Is the space supporting a variety of activities? How long are patrons staying? Next a social index (SI) scores the social activity of the space. Are people utilizing social areas and assets (*e.g.*, dining areas, fire pits)? Are people spending time together on site, engaging with others? The SI and AI both utilize user selected weighting coefficients so persons using the tool can finetune the indices'' focus on capturing the factors

that matter most to them. Additionally, the SI and AI will have user set variables such as ideal traffic count in an hour to calibrate the indices'' to a specific site – enabling the tool to extract useful information for large popular sites (*e.g.*, riverfront carousels, food truck lots, etc.) as well as smaller less popular sites. The SI and AI are designed to score a scene on an hourly basis, for a given hour how active and social was the space?

A subset of the interview subjects (n=3) from chapter 4 were given an opportunity to demo early versions of the AI and SI and offer feedback. The process was iterative, the chosen subset was able to demo the AI and SI tools three times over the course of two months. During each demo the participants were shown clips of various scenes and were given a description of all the activities and traffic observed during each scene. The demo participants were then able to see the SI and AI score for each scene, along with a breakdown of the value of each component of the equation. During this demo participants stated if they felt the scores accurately reflected the social activity in the scene and if they produced expected scores. Additionally, demo participants were asked about the form of the SI and AI and how they were calculated and offered feedback on if they felt the equations were missing any components or accurately captured the characteristics of a social and active space. The latest iteration of the SI and AI are presented below and had the highest agreement with demo participant expectations of all previous iterations.

6.1.1 Activity Index

The AI, which focuses on a space's current activity (social and non-social), pulls from three sources to assess performance: Traffic count, average time on site, and diversity of activities. The equation for scoring the AI of a space for a given hour is given below and will

range from 0-100.

$$AI = \alpha \times \underbrace{\min\left(100, \frac{TC}{ITC} \times 100\right)}_{\text{Traffic count}} + \beta \times \underbrace{\min\left(100, \frac{UA}{IUA} \times 100\right)}_{\text{Diversity of activities}} + \gamma \times \underbrace{\min\left(100, \frac{TCAA}{TC} \times 100\right)}_{\text{Time on site}} \quad (6.1)$$

The alpha (α), beta (β), and gamma (γ) terms are user selected coefficients that weigh the influence on the score from the three different sources. TC is the observed traffic count for the hour, while ITC is the ideal traffic count in an hour – set uniquely for each space based on historical traffic and park manager input. UA is the number of unique activities (*e.g.*, enduring stroll through space, people watching on bench, cycling) observed in the space, while IUA is the ideal number of unique activities supported by the space – also set uniquely for each space based on historical data and park manager input. The final part of the equation is a measure of how long patrons are staying on site. For each site the average time it takes a person to walk through the space (without engaging with any assets or stooping to talk) is estimated is estimated as AWT . $TCAA$ is then the observed count of patrons who spent time in the site longer than 1.1 times the walk through average, AWT . The equation is ratio-metric, measuring the ratio of observed traffic to the ideal, how much of that traffic spent longer on site than the AWT , and the observed number of unique activities compared to the ideal.

The sum of alpha, beta, and gamma, must equate to one. To simplify the process for users, the values of 0.5, 0.3, and 0.2 are pre-chosen. Users are then prompted to rank the importance of the three sources of influence in regards to the specific site (do they care more about time spent on site, traffic count, or the number of unique activities). These values are then assigned to alpha, beta, and gamma, based on the ranking provided by the user. Additionally, the user is asked to give an ideal number of traffic for an hour – during peak performance. They are

also asked to provide an ideal number of unique activities they believe space supports. These values are then assigned to *ITC*, and *IUA*. By querying the user for site specific numbers and weights, the index can adapt to a variety of spaces and can be finetuned to accurately assess the activity performance of site given a site's unique properties.

6.1.2 Social index

Similar to the AI, the SI, a measure of the space's current social activity, pulls from three sources to assess performance: volume of persons in social areas, time patrons spend with others, and the ratio of pro-social behaviors. The equation for scoring the AI of a space for a given hour is given below and will range from 0-100.

$$\begin{aligned}
 SI = & \alpha \times \underbrace{\min\left(100, \frac{TC_s}{ITC_s} \times 100\right)}_{\text{Traffic in social areas}} + \beta \times \underbrace{\min\left(100, \frac{TCAA_s}{ITC_s} \times 100\right)}_{\text{Time spent with others}} \\
 & + \gamma \times \underbrace{\left(\frac{Ps + E + F}{Pn + Ps + E + F}\right) \times 100}_{\text{Ratio of pro-social behaviours}}
 \end{aligned}
 \tag{6.2}$$

The alpha (α), beta (β), and gamma (γ) terms are user selected coefficients that weigh the influence on the score from the three different sources. TC_s is the observed number of persons spending longer than 10 seconds engaging with social assets (*e.g.*, fire pit) or in a social area (*e.g.*, dining area) for the hour, while ITC_s is the ideal count of persons in social areas in an hour – set uniquely for each space. The second part of the equation is a measure of how many patrons are spending time with others on the site. For each site the average time it takes a person to walk through the space (without engaging with any assets or stooping to talk) is estimated as *AWT*. $TCAA_s$ is then the observed count of patrons who spent time with another person in the site longer than 1.1 times the walk through average, *AWT*. In this setup, couples quickly moving

through the space and not engaging with others or any assets will not be counted toward increasing the social score. The final part of the equation analyzes the social classifications (Chapter 5) of all observed patrons in the hour, and calculates a ratio of pro-social behaviors (Fleeting (F), Enduring (E), Passive social (Ps)) to all behaviors (F , E , Ps , Passive non-social (Pn)). The equation is ratio-metric, measuring the ratio of observed traffic using social assets to the ideal, how much of that traffic spent time engaging with others longer than the AWT, and the amount of pro-social behaviors compared to the sum.

Similar to the AI, the sum of alpha, beta, and gamma, must equate to one. To simplify the process for users, the values of 0.5, 0.3, and 0.2 are pre-chosen. Users are then prompted to rank the importance of the three sources of influence in regards to the specific site (do they care more about utilization of social assets and areas, time spent with others, or the ratio of patrons engaging in pro-social behaviors). These values are then assigned to alpha, beta, and gamma, based on the ranking provided by the user. Additionally, the user is asked to give an ideal number of patrons utilizing social assets and areas for an hour – during peak performance which is then assigned to $ITCs$. Again, similarly to the AI, the SI captures prominent factors discussed in the interviews and allows for users to finetune to the index to adapt to specific spaces and incorporate personal preferences. This adaptability and customization ensure the utility of the index and that it's meeting the needs for the user and unique park space.

6.2 Experiments

The SI and the AI were integrated with the sensing framework and deployed on three different cameras along the Detroit Riverfront. The indices'' were used to analyze data from the summer and fall of 2022 and captured various park programming events. Each camera captured a

different area of the park systems. The three sites are quite different from each other and offer diverse assets and services. Analysis and findings from the indices'' for each site are presented below.

6.2.1 Freight Yard

The camera at the Freight Yard captures a portion of the Dequindre Cut, an exercise beltway that cuts through the city. The beltway has a bike specific lane and walking lanes and features art pieces and workout equipment along the 1.65 mile stretch. The camera is situated in the Freight Yard area which contains nine repurposed shipping containers serving food, beverages, retail items, local art, and at times a live DJ. In the camera's field of view (Figure 5.15), is a stretch of the Dequindre cut as well as a few chairs and benches in the dining area of the Freight Yard.

The indices'' were used to measure the performance of the site on Fridays and Sundays in the month of July. During this month the indices'' also captured the "Summer Sundays" social programming event held on July 17th, 2022. Example data is provided in figure 6.2. In general, the AI for the space was consistently above average on the weekends, as the space supports multiple forms of activities – cycling, skateboarding, jogging, dog-walking, and socializing at the tables. However, compared to the AI the SI was typically low, telling park managers most traffic is along the Dequindre cut and less traffic is utilizing the social spaces and areas provided by the Freight Yard.

There were two cases where the SI was much higher than the other days, despite overall traffic being similar. The first is during the "Summer Sunday" program, where live music and games encourage patrons to visit and enjoy the Freight Yard space. When compared to the following Sunday, the "Summer Sunday" increased overall traffic by 25%, however this does not

tell the whole story. When looking at the SI we see the “Summer Sunday” programming increased the SI of the space by 280% - quickly informing park managers the program successfully improved the social quality of the space. The SI also saw a sharp increase on the final Friday of the month. While no organized programs were running in the space during this evening the social scene was more active than other Fridays in the month showcasing the spontaneous nature of some social scenes. This data shows the importance of having furniture and assets that support and encourage the emergence of these social scenes, so they organically happen more often.



Figure 6.1 Scenes from the Freight Yard camera on a 7/24/22 (A), 7/17/22 (B), 7/22/22 (C), and 7/29/22 (D).

6.2.2 Valde Park

The indices” were also used to analyze a plaza in Valde Park – a riverfront space that includes multiple dining tables and is near sand volleyball courts, a playground, and concessions.

Data from the month of July in 2022 was analyzed using the AI and SI generated from the sensing framework (Figure 6.3). For this example, alpha, beta, and gamma were assigned 0.5, 0.3, and 0.2 respectively, for both the AI and SI. The SI was heavily dependent on time of day. Morning hours of weekdays were predictably the lowest scoring. Late afternoons saw increased AI – primarily driven by the increase in traffic flow and greater diversity of activities (*e.g.*, people exercising in space). As the day progressed into evenings the SI greatly increased, looking at chart in Figure 6.3, the breakdown of the AI and SI by source shows the increase in SI on Monday evenings is driven most by the increase of traffic in the social areas – in this hour people started utilizing the dining areas more. Saturday’s, as expected, consistently had the highest scores. However, the biggest increase in the AI on Saturday is due to the increased amount of time patrons were spending in the space. We see the same trend in the SI, there is a dramatic increase in the amount of patrons spending significant time with others in the space – leading to high SI scores.

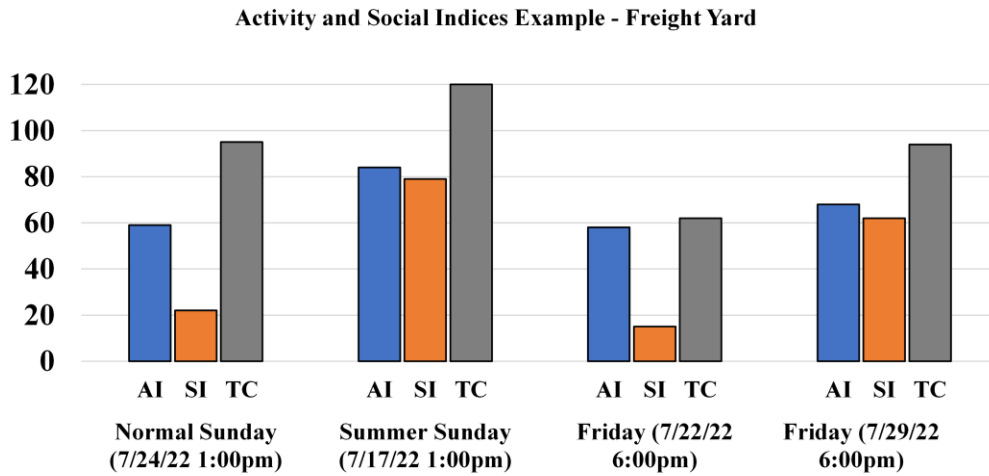


Figure 6.2 Traffic count (TC), Social index (SI), and Activity index (AI) of specific times in July of 2022.

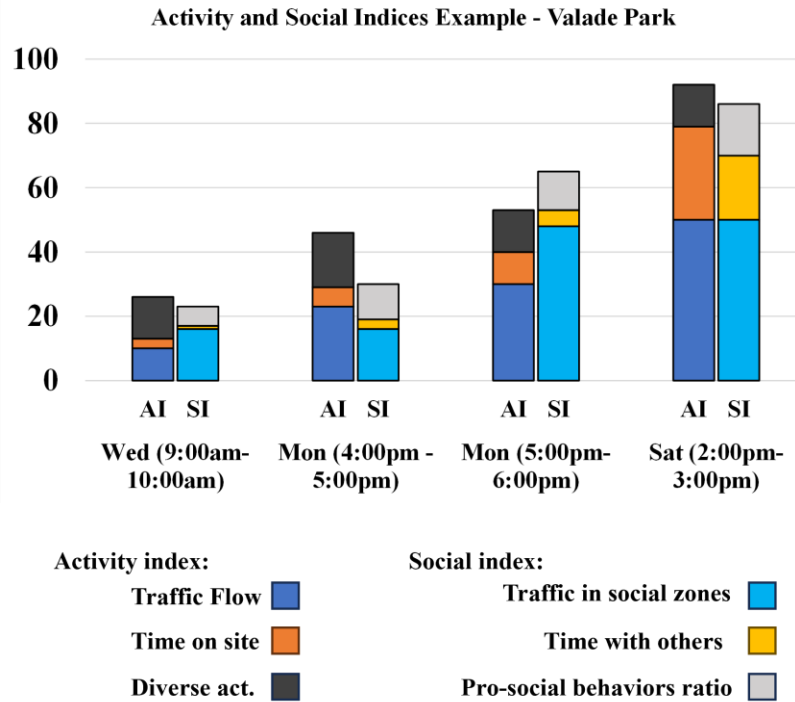


Figure 6.3 SI and AI examples on data from Valade Park in July of 2022.

6.2.3 Cullen Plaza

Cullen Plaza is a feature-rich area of the park. Sitting on the east riverfront the space provides patrons with views of water, while also providing food and drinks through a tiki bar and café, and a carousel for children and families to enjoy. Patrons can rent bikes here as well as board riverboat tours through Diamond Jack’s. The SI and AI were tested to evaluate the performance of the site during July 2022.

For this example, alpha, beta, and gamma were assigned 0.5, 0.3, and 0.2 respectively, for both the AI and SI. As expected, Saturdays were most popular and had consistently higher SI and AI scores. A breakdown of the scores by area of influence shows in addition to higher traffic, patrons spent more time on site, with a majority staying longer than the AWT. Similarly, for the SI, on the weekend, the scores increased both from a higher number of patrons using

social assets but also from the amount of time patrons were spending engaging with others. The hours before a Diamond Jack's riverboat tour were incredibly popular, even on weekdays leading to scores similar to peak weekend times. Weekday mornings were the quietest except for a few occasions where the place was particularly busy. Certain weekday mornings host the "Riverwalkers" program, an official program sponsored by the DRFC where people exercise and walk through the park areas together. During these mornings, traffic is usually slightly higher than average, however on July 11th the SI and AI both doubled the average. Investigation of this showed multiple families passed through the area and took photos in front of the Riverview. Families did not stay long, however, and quickly moved on. This is reflected in the scores, a majority of the increase came from the higher traffic numbers, while time on site and with others only grew marginally. Mornings like July 11th might be an anomaly or there could be an outside touring agency making stops at the plaza. Examples breakdowns of some of the scores are shown below along with image stills of the space during the shown example times.

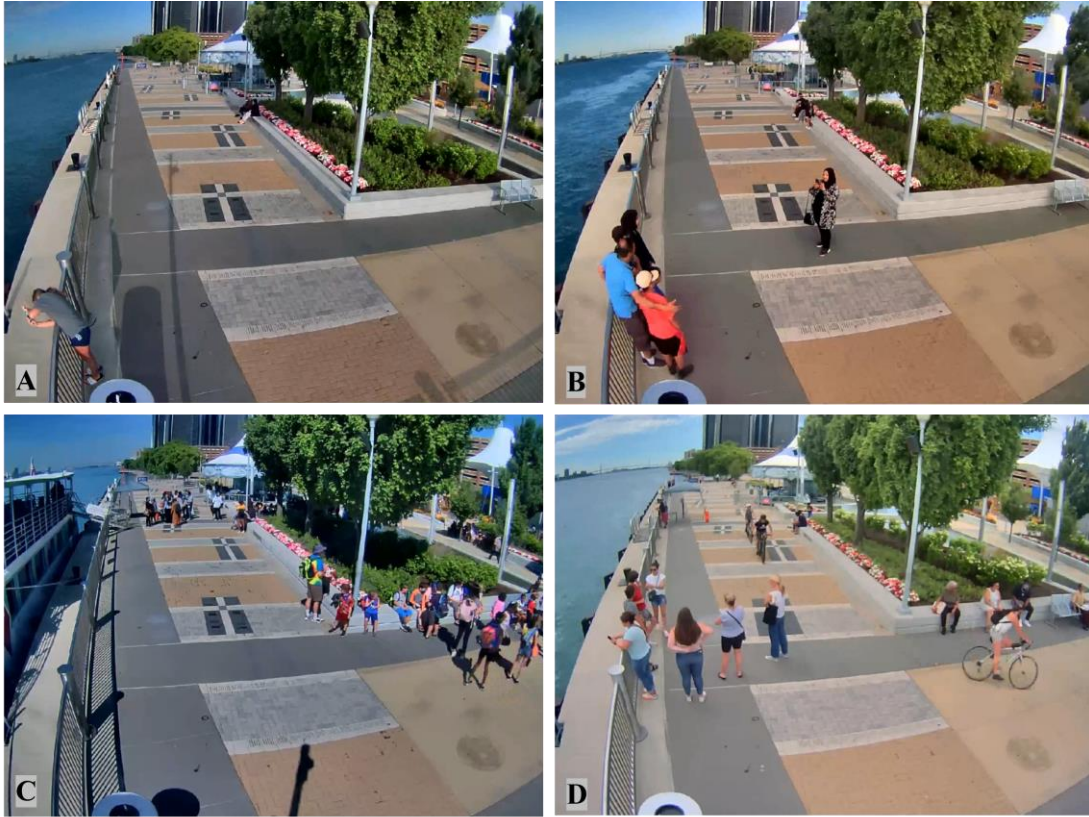


Figure 6.4 Images from Cullen Plaza during. (A) 7/20/22 8:00am. (B) 7/11/22 8:00am. (C) 7/7/22 10:00 am. (D) 7/16/22 4:00 pm.

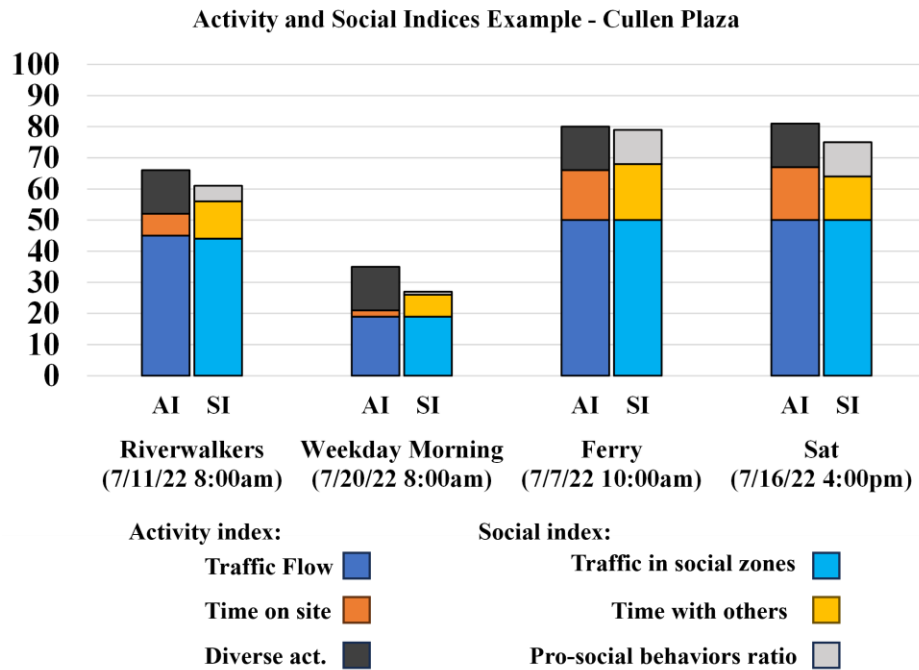


Figure 6.5 SI and AI examples on data from Cullen Plaza in July of 2022.

6.3 Conclusion

With the additions of this chapter, the sensing framework is able to autonomously quantify social use and value. The intellectual contributions here are:

- A novel system of metrics on sociability to enhance the decision-making process around park management and design.

The presented indices are as simple to use as an hourly traffic estimate, allowing managers to efficiently assess peak times and areas, but give richer details that are left uncaptured by traffic counts. For example, when using the SI and AI, you can see that “Summer Sunday’s” programming is quite effective at increasing the area’s sociability. While traffic count was only increased by roughly 25% when compared to normal Sundays, the SI nearly tripled – conveying the program effectively converted the usual ‘passing through’ traffic from the

Dequindre cut to Freight Yard visitors. When assessing which investments are “worthwhile” and what programs to fund this type of information is vital. The AI and SI provide immediate feedback, allowing managers to quickly see the fruits of investments and programs and identify areas for improvement.

While the data generated by the sociability module, presented in chapter 5, is more comprehensive, showcasing hotspots of activity and breaking down social engagement by area, at times it may be too complex or ‘bloated’. The activity and social indices’ are a necessary addition to the framework, providing scalability and utility to park managers working with a multitude of sites assessing across long time spans. The SI and AI quickly and succinctly highlight peaks and valleys and point to managers what times and zones should be analyzed deeper using the detailed outputs from the sociability module (chapter 5).

Quantifying something as qualitative as “sociability” is a challenging task, and it is of the belief of the author there is no true objective measure. The flexibility and customization of the indices’ is intentional, allowing for park managers to adjust the index to capture the identified components of sociability (highlighted in chapter 4) they deem most relevant. Through the alpha, beta, and gamma coefficients managers can tune the index to their needs. Perhaps certain programs are not aimed at increasing traffic instead are purposed solely to increase patron engagement with others, or other investments might be targeted at increasing the range of activities for patrons to partake in. The coefficients allow managers to best fit the tool to their current needs. While pre-chosen values of 0.5, 0.3, and 0.2 were chosen for this implementation, they may be chosen as any other combination of values (if the sum is equal to one). Pre-chosen values simplify the setup process for users, but other visual based methods may provide the same

simplicity while allowing for greater customization, such as an interactive pie chart, where users may portion the pie slices (the three sources of influence) to however they like.

6.3.1 Validation

The current iteration of the SI and AI was refined after incorporating user feedback through the demos with a subset of the interview participants. However, two different approaches are proposed to ensure the AI and SI go through a more rigorous validation process. First the demonstrations are slated to be expanded both content wise and audience wise and transformed into workshops. All participants from the chapter 4 interviews, as well as other identified stakeholders will be invited to participate. Participants will go through a process where they individually score the social activity of recordings of multiple, various park scenes. Participants will also be asked to provide reasoning and explanation for their scores. They will also be shown the scores generated by the SI and AI algorithm and surveyed on if they feel the tool is over or under evaluating different aspects of the scene. The participant scores will be compared to the autonomously generated scores. These comparisons along with the survey responses will be used to further refine and validate the AI and SI. Second the web-dashboard may be expanded to include weekly snippets from park scenes, and query dashboard users how they would score the scene. This continuous feedback tool will generate larger amounts of data to use to validate and refine the form of the AI and SI equations.

Chapter 7. Discussion

This chapter briefly describes extensions of the framework to other applications, with preliminary experiments and results for applying the framework for human health and performance monitoring. Additionally, this chapter discusses the broader impacts of the work, the necessity of inclusive models, and how extensions of the work can be used for deeper analysis and increase park accessibility and improve accommodations for those with disabilities. Lastly, the chapter addresses issues around privacy and community trust while also discussing the limitations of the presented work.

7.1 Extensions of the framework

Using camera-based sensing systems to detect, track, and map persons and objects as they move through space has numerous applications. The detecting, tracking, and mapping toolchain presented in chapters 2 and 3 may improve or replace sensing systems in different fields and areas. For example, the toolchain was specifically augmented and tested on human health monitoring applications. Data from the camera-based classification and tracking system can provide additional context to other data streams on human health and performance to enable more complex monitoring and analysis.

Advances in wireless sensor technologies have led to a significant increase in human health monitoring. Commercially available wearable biometric sensors are popular for users intending to track their fitness, health, and general wellbeing. By combining data streams of a person's biometrics from wearable sensors with visual classifications of their pose, activity, and location, we can enhance the capabilities of health and performance monitoring.

Section 7.1 details the specific integration of our detection, tracking, and mapping framework with wearable sensors on soldiers in urban warfare settings. A soldier's health and status in dangerous situations can be better assessed when fusing data from the soldier's biometrics and visual data from a body cam, drone, or other camera in the area. The sensing framework (detection, tracking, and mapping – chapters 2 & 3) was used to test early iterations of such a comprehensive human health and performance monitoring system. Specifically, an early data fusion and matching framework was developed to test the feasibility of matching tracks produced from the camera sensing system to the correct biometric data from the wearables.

7.1.1 Methods

7.1.1.1 WAR image gallery

The human health and performance tracking framework presented uses sets of warfighter-specific image gallery to train the state-of-the-art computer vision models for detection and tracking used in chapters 2 and 3. This image set is termed Warfighter Activity Recognition (WAR) image gallery. There are a multitude of open-source annotated image



Figure 7.1 Examples from WAR image gallery.

galleries [43], [44] available for training CNN models. These galleries contain tens of thousands of annotated images of animals, automobiles, persons, faces, accessories, and a variety of other common objects. However, the galleries lack representation of objects common in military applications, such as fatigued soldiers, firearms, combat utility vehicles, and helicopters. For these reasons a custom dataset is needed to properly train the CNN to reliably detect soldiers with heavy equipment and backpacks as well as other military specific objects. The WAR image gallery includes 2,000 images containing 16,000 instances of warfighters, rifles, tanks, helicopters, and combat utility vehicles (Fig. 1). The images are taken from public domain footage of various military exercises. Additionally, the warfighters are further classified by pose – delineating warfighter instances by upright, crouch, and prone positions. A secondary classifier structure, as presented in 2.1.4, is trained on the warfighter poses to classify warfighter position.

7.1.1.2 GPS mapping

The wearables are equipped with GPS, transmitting GPS estimates of the warfighter’s location at a frequency of 1 Hertz. For data fusion, the GPS estimates of the wearables will be matched to the location estimates of tracked warfighters coming from the camera-based sensing system. To do so, the mapping module not only needs to transform the 2D pixel coordinates of the tracked warfighter to a 3D world coordinate system, but that 3D world coordinate system also needs to be linked with GPS coordinates. For this application the same mapping module and algorithms used in chapter 2.1.2 are adopted. The only modification is the linking of the GPS coordinates. To estimate GPS positions of tracked individuals from the camera, the 3D world coordinate system is aligned with latitude and longitude, and the GPS location of the camera is recorded and set as the origin. When reporting object locations, the mapping module will generate the change in distance along the latitude and longitude from the origin. Using estimates

of conversions from changes in meters to changes in latitude and longitude degrees[89], which are a function of the geographic position, the mapping module can produce GPS estimates of objects tracked by the camera.

7.1.1.3 GPS matching

For each individual in a camera’s field of view wearing the biometric sensors there are two GPS streams. One GPS estimate comes from the wearables. The second GPS stream comes from the camera sensing system that has the tracking data - classifications on the warfighter’s activity (jogging, prone, etc.) and if they are armed – as well as the GPS estimates from the mapping module. To fuse the two streams and provide deeper context to health monitoring of the individuals the data needs to be correctly matched – which wearable streams correlate to which tracks? To accomplish this association the fusion framework leverages the Hungarian sorting algorithm[50], similar to the matching algorithm presented in chapter 2.2 but with a different cost function. The haversine distance formula is used to calculate the distance between two GPS estimates[90]:

$$d(i, j) = 2r \left(\sqrt{\sin^2\left(\frac{\varphi_i - \varphi_j}{2}\right) + \cos(\varphi_i)\cos(\varphi_j)\sin^2\left(\frac{\gamma_i - \gamma_j}{2}\right)} \right)$$

(7.1)

Where r is the earth’s radius, $\varphi_{i,j}$ are the latitudes of the GPS estimates from wearable i and track j , and $\gamma_{i,j}$ are the longitudes of the GPS estimates from wearable i and track j . To generate the cost between matching a wearable data stream i with a computer vision track j at time k , the following formulas are used:

$$cost(i, j) = mean(DC \times W)$$

(7.2)

$$DC = [d(i, j)_k, d(i, j)_{k-1}, d(i, j)_{k-2}, \dots, d(i, j)_{k-n}]$$

(7.3)

$$W = \left[\frac{1000}{1000}, \frac{1000-1}{1000}, \frac{1000-2}{1000}, \dots, \frac{1000-n}{1000} \right]$$

(7.4)

Where DC is a vector of past haversine distances between track and wearable GPS estimates, with a max n of 1000, and W is a vector of weights to tune the influence of recent distances more than past. For each time k , a cost matrix is generated for each potential pairing and the Hungarian sorting algorithm computes the pairings with lowest total overall cost.

7.1.1.4 Experiments

An early iteration of the sensing fusion framework was tested at the MCity facility on the University of Michigan's campus. MCity is a controlled 16-acre intelligent testbed environment with intersections, traffic lights, building facades, a network of surveillance cameras, and a variety of other city infrastructure. For the tests, two GPU servers were installed on the network to process image data from the cameras. Additionally, a LoRa gateway was installed on site to push on-body sensor data to cloud servers. For two days, 10 volunteers, dressed in fatigues and armed with plastic training rifles ran military exercises and battle simulations in the city. The simulations were captured on camera and processed in real-time using on-site GPU servers. The volunteers were also outfitted with on-body biometric sensors and GPS units. JSON encoded detection data and the on-body sensor streams were pushed at a frequency of 1 Hz to a cloud server. Offline the data streams from the two sources, cameras and wearables, were fused utilizing the GPS matching method described in 7.1.1.3.

7.1.2 Results

7.1.2.1 WAR image gallery

The object detector was trained on a merged dataset combining civilian instances from COCO [43] with the instances from WAR. To create a validation test set, 15% of the instances from WAR were withheld. Table 7.1 documents the results of the model when trained on the merged data.

Table 7.1 WAR + COCO training results.

Class	Instances from WAR image gallery	Instances from COCO	Accuracy (Precision/Recall)
Person	13,000 (full fatigues + equipment)	262,465 (civilian)	0.96 0.79
Rifle	2,700	NA	0.99 0.93
Combat vehicle	480	NA	0.84 0.84
Civilian vehicle	5	53,840	0.94 0.96

Additionally, a secondary classifier was trained to classify detected persons by their position – prone, upright, or crouch. The light weight Resnet18 [39] model trained on ImageNet [44] was chosen for this task. A transfer learning approach [91] was used with frozen weights on every layer except the final fully connected layer - which was trained on the instances in WAR with labeled positions. Similarly, 15% of the training data is withheld for testing and validation.

Training results for each subclass are listed in Table 7.2.

Table 7.2 WAR position sub-classifier training results.

Class	Instances from WAR image gallery	Accuracy (Precision/Recall)
Upright	10,595	0.93 0.96
Prone	1,150	0.90 0.93
Crouch	1,255	0.76 0.64

7.1.2.2 GPS mapping

Three on-site cameras were calibrated before the tests to extract the necessary parameters needed for the mapping module. During the tests, a Google Maps API was used to plot estimated GPS coordinates of detected objects onto a map of MCity (Fig. 7.2). Furthermore, GPS coordinates of thirteen locations were manually recorded using a handheld GPS device. The GPS estimates from the thirteen points were compared against the estimates of the points from the mapping module. On average, the GPS camera estimates, and GPS measurements were within 1.3 meters of each other.

7.1.2.3 GPS matching

Fourteen different scenarios with varying numbers of subjects and choreographies were used to test the capabilities of the GPS matching algorithm. Each scenario lasted between 60-150 seconds and featured subjects jogging, walking, and mingling with each other. Data from both the GPS and video feed operated at 1Hz. The algorithm attempted to match tracked targets from the camera to wearable data streams at each time step. To test the matching algorithm further and increase the difficulty of the challenge some scenarios had a few subjects without wearable GPS



Figure 7.2 (A) GPS heatmap of six subjects jogging through MCity street - using GPS estimates from mapping module. (B) GPS heatmap of two subjects patrolling MCity street and crossing paths – using GPS estimates from mapping module.

devices. For each scenario the accuracy of the matching was evaluated using the following formula:

$$Accuracy = \frac{TP}{TP + FN + FP}$$

(7.4)

Where TP is the number of correct matches, FP is the number of errors where a target without a wearable GPS was matched to a wearable, and FN is the number of errors where a target with a wearable GPS went unmatched. Details of each scenario along with performance results are listed in table 7.3.

Table 7.3 GPS Matching

Scenario	Total subjects	With wearable	Pace	Choreography	Accuracy
<i>1</i>	2	1	walk	together	1.00
<i>2</i>	2	1	walk	together then separate	0.94
<i>3</i>	2	1	jog	separate then together	1.00
<i>4</i>	2	2	walk	together	0.89
<i>5</i>	2	2	walk	together then separate then return	0.97
<i>6</i>	3	2	walk	together then separate then return	1.00
<i>7</i>	3	2	jog	together	1.00
<i>8</i>	3	2	walk	one separate joining the other two	0.84
<i>9</i>	4	2	walk	together	0.81
<i>10</i>	4	2	walk	together then separate	1.00
<i>11</i>	4	2	jog	together	0.67
<i>12</i>	5	2	walk	together	0.65
<i>13</i>	5	2	walk	random disperse	1.00
<i>14</i>	5	2	jog	together	0.68

7.2 Broader Impacts

Social infrastructure represents an essential infrastructure supporting the social interaction and cohesion of a community [7]. Design, management, and operation of social infrastructure directly impacts a community's social sustainability and resilience [6] – which is especially pertinent as issues from climate change continue to arise. This research advances scientific knowledge in how computer vision-based sensing systems can be used to observe and model the performance of social infrastructure empowering cities to more rationally prioritize investment and target specific outcomes. Especially for the shrinking cities of the American Midwest, social infrastructure is one of the most important elements needed to start and maintain transformation into resurgent post-industrial cities [92]. This research's engagement with community stakeholders, especially those with the ability to sustain research outcomes well past the research timeline, benefits the entirety of Detroit and its metro-region. The research worked directly with the social programming and event staff of the DRFC to analyze findings and tailor activation events that best benefit the community and draw attention to the park services and amenities. With many Detroit communities living below the poverty line, engaging them in social infrastructure governance and ensuring they receive quantifiable benefits from community assets helps achieve social equity in infrastructure design and management.

Furthermore, beyond the impacts on Detroit and the involved park spaces during the research, other communities can benefit from the work. Through publication, documentation, and open source the described framework can be used by other cities, and governing bodies of public spaces. For example, by leveraging relationships developed with the DRFC, the research

has been presented on multiple occasions to other non-profits and organizations associated with The High Line Network. The High Line Network [93] is a nonprofit organization that consists of social infrastructure projects based on the reuse of abandoned urban infrastructure. The network has 37 projects (including DRFC's Dequindre Cut) and convenes a community of park managers from across the US. This research has been regularly presented to this inter-city knowledge network to disseminate findings and share ideas.

Additionally, beyond having an impact on a community's social sustainability the presented sensing framework (Figure 1.2) is relevant in the topic of environmental sustainability. According to the UN's sustainable development goals report, 55% of the world's population living currently in urban areas is expected to rise to 70–75% by 2050 [94], [95]. To help combat the escalating rate of urbanization and all the challenges and issues that come with it – climate change, increased consumption, stress on various infrastructure systems and energy resources - cities are seeking various data-driven technologies and innovations [96]–[100]. In a push to increase sustainability cities are adopting eco-friendly practices - such as promoting public transportation, implementing energy-efficient infrastructure, encouraging green spaces, and adopting green-tech innovations [101], [102]. By prioritizing sustainability, cities can mitigate environmental impacts, reduce carbon emissions, and enhance the overall well-being and resilience of their residents. This sensing framework can help achieve these goals through resource conservation. The toolset and resulting analysis of data gathered enables park managers and operators to be better stewards of their resources and funds. With this framework, they can make data driven decisions and more efficiently spend and disperse park resources.

7.2.1 Model bias and accessibility

The performance of a CNN in object detection is heavily predicated on the data used to train the CNN. To perform well in diverse conditions and accurately detect variations of an object the model needs to be trained on an exhaustive set of images, extensively showcasing the variations and types of objects the model may come across. For these reasons, not only is it possible, but is likely a CNN trained to detect people can have bias. Specifically, multiple research endeavors have shown commercially available CNN based algorithms to have racial and gender bias [103]–[106]. When datasets are predominately male and white, algorithms can struggle when working with female or non-white faces. A couple of higher-profile examples include: zoom virtual backgrounds blurring out black users, and twitter’s editing software that automatically edits faces of uploaded images to better highlight them often missing black or female faces [107]. The OPOS library was built utilizing public video feeds of park spaces in Detroit, Michigan – allowing for the curation of a racially and gender diverse dataset. Furthermore, person detection, especially in the application of lower-resolution surveillance cameras, is less sensitive to racial and gender bias, as images are distant full profiles – as opposed to facial recognition software where skin tone and gender features play a more prominent role in classification. However, the model and application space can be more sensitive to other types of bias – such as consistent detection of mobility challenged individuals.

When designing a sensing framework with the intent to capture how patrons move through and utilize space to provide better services and improve park design – it is imperative to ensure all user experiences are tracked. It is especially important to ensure park designs, programming, and features are accessible and that experiences of those with mobility challenges are not ignored. However, certain equipment such as wheelchairs, can obscure large parts of the patron and inhibit detection.

Despite the low frequency of mobility challenged individuals in the park feeds – leading to few examples for training, OPOS is still robust to occluded individuals. First, OPOS includes a “person_other” category composed of occluded individuals found in the training images. This class covers individuals that are mostly blocked by friends or are cut off from the camera. There are thousands of these instances, which help the CNN-based object detector (YOLO) of the sensing framework detect patrons with body parts blocked by mobility equipment. Furthermore, OPOS has other classes which help in this instance. For example, the ‘sitter’ has thousands of instances of patrons sitting at tables, on benches, fountain steps, and chairs, while the ‘scooter’ and ‘cyclist’ class has thousands of examples of patrons using various forms of bicycles and scooters. The training from these classes enables the detector to consistently detect those sitting in wheelchairs or other similar types of equipment. However, with such few validation and testing examples of patrons explicitly using said equipment we are unable to rigorously test and validate the framework’s performance on tracking those with mobility challenges.

Therefore, a natural and compelling next step would be the curation of a dataset with thousands of examples of patrons using mobility equipment such as wheelchairs, walkers, and mobility scooters. In doing so, we accomplish two things. First, such a dataset would allow extensive testing of the framework and provide substantial evidence on the framework’s capabilities in tracking patrons with mobility equipment. By doing this, any potential bias the model may have may be exposed and fixed. Second, with such a dataset a specific sub-class could be created – training the model to label all patrons using such equipment. With this capability, the framework could be targeted at specific applications where the movement patterns and behaviors of mobility challenged people are especially pertinent. Additionally, specific

analysis and data on how programs or design choices impact those with mobility challenges could be computed for any application of the framework.

A few relevant datasets currently exist, such as mobilityaids [108], a dataset used to train hospital robots in identifying patients in wheelchairs. Additionally, a few datasets of wheelchairs can be found on open-source websites. These datasets would be a starting point, and together with images from different scenes (distant surveillance cameras as opposed to hospital cameras) could yield enough data to train a mobility equipment sub-class. While such tasks were not completed during the dissertation - given time and resource constraints, as well as the fact there were zero implications of the framework having a bias, it is the authors view they are worthwhile and will be a part of future plans and research involving this framework.

7.3 Challenges and limitations

7.3.1 Sociability interviews and community feedback

Due to time and resource constraints the sociability interviews focused only on park managers and operators. While plenty of insights were gained from the interviews regarding perceptions of sociability, social activities, and relationships between space and community, the range of interviewed perspectives was limited. Vertical diversity in management structure was achieved, with CEOs, managers, and volunteers being interviewed, however, the patron voice was left out. Amassing a sample size representative of the diverse population of Detroit park users was too large a task for the scope and resources of the research. Interviewing the patron population could lead to fresh insights and perspectives on perceptions of what makes a place sociable and desired outputs of social programming and park features. Extensions of this work

would include additional rounds of interviews and polling to further refine the sociability classification schema, and the social and activity indices”.

7.4 Conclusion

The sensing framework presented in this dissertation pushes a new paradigm shift in how social infrastructure is designed, managed, and operated. The flexibility and depth of the framework delivers simple and immediately impactful metrics but can also yield complex analysis when desired. By revolutionizing how sociability in public open spaces can be autonomously observed and quantified, we can better understand the intimate connection between the physical built form and the social organization people. With newfound understanding and analytics, public spaces can be designed to optimize and better encourage the types of social interactions and behaviors that build social capital and increase community wellness, social sustainability, and community resilience. Instantaneous feedback on social programming and design interventions will drastically improve decision making processes of stakeholders and deliver higher quality services for communities.

While this research focused on city parks as one form of social infrastructure, the work is applicable to many other physical settings that brings people together [109], [110]. The social infrastructures best suited to the proposed approach include public health care facilities with open spaces, shopping malls, commercial plazas, and public transportation hubs that have surveillance cameras. The types of social infrastructure that are less suitable to use the proposed approach might include spaces which are narrow, spaces with limited surveillance camera coverage, or areas where cameras pose the risk of violating the trust of communities.

Beyond social infrastructure management, the data produced by the system also allows more traditional asset managers associated with the management of built environment assets a quantification of public use of assets allowing loads to be assessed. For example, quantifying how much people use a specific physical asset may suggest more investment in inspecting and maintaining that asset will be needed. Coupled with physical response data from the asset itself, a more complete picture of the loads and responses of the system can be derived.

Going back to the example of the baby blue piano in Lonsdale Quay and the concept of “triangulation” (discussed in Chapter 1), you can see that people, more specifically, the human experience is at the heart of this research, driving the motivation and purpose of the sensing framework. How might we better design, manage, and plan our social infrastructure to better facilitate social organization and build social capital?

Appendices

Appendix A: Sociability Interview Protocol

Exploring sociability in public open spaces and how it relates to performance and desired user experience
Park manager/owner/investor interview structure

Overview and Objectives:

To better design, operate, and manage public open spaces it is imperative for owners and investors to understand how patrons use their spaces and derive benefits from them. With recent advances in wireless sensing and computing technologies, urban sensing strategies can go beyond traffic counting and capture deeper and more complex activities and social behaviors. An urban sensing framework designed around capturing 'sociability' – by tracking activities, interactions, and social behaviors in the park can empower park managers, owners, and investors to make data driven decisions and improve the experiences of park patrons. **To this end, these interviews seek to work with park owners, investors, and managers to understand and identify desired patron activities and behaviors which are perceived to increase sociability.** A sensing toolset that can capture desired outputs of social programming and park features will enable park managers to better assess their performance and drive investment in features and programming which are achieving their goals. Furthermore, these interviews seek to understand how park managers, owners, and investors wish to engage with sociability data to ensure the utility of the web dashboard -- by featuring impactful sociability analytics and visualizations.

Interview Objectives:

1. Identify desired outputs (activities and behaviors) from social programming and how they relate to sociability
2. Identify desired outputs from park features (benches, fountains, fire pits, etc.) and how they relate to sociability
3. Identify common activities in public open spaces and how they are perceived to relate to sociability
4. Identify public social behaviors perceived to contribute to sociability
5. Identify physical features and structures perceived to be relevant to sociability or help contribute to sociability
6. Identify desired web dashboard features and visualizations
7. Identify key sociability-related analytics
8. Ascertain importance of sociability and user experience to park managers/owners/investors

Sociability: Pulling from past research, a sociable public space has been described as a space where people can carry out their activities in relative comfort and safety while interacting, engaging in spectacles and ceremonies, or just simply sitting or waiting. Sociable spaces facilitate social and leisure activities, whether conducted individually or as a group, and give people a place to connect and gather. In the context of this study, the sociability of a space describes the current extent to which the space is able to facilitate the emergence of a social space. A sociability index would describe the current level of social interaction, engagement, and activities supported by the space.

<p>Background and importance of sociability 5-10 Minutes</p>	<p>Q1: What is your official title and position in [insert subjects affiliated organization/group]? <i>Linked objectives: NA</i></p> <p>Notes on response:</p>
	<p>[Clarifying follow-up question(s)]: Notes on response:</p>
	<p>Q2: Could you describe in detail the main duties and responsibilities of your role? <i>Linked objectives: NA</i></p> <p>Notes on response:</p>

	<p>[Clarifying follow-up question(s)]:</p> <p>Notes on response:</p>
	<p>Q3: [Read section on sociability and the team's definition and use of the term in the context of our study]. Would you agree with this definition of sociability in the context of public spaces? <i>Linked objectives: NA</i></p> <p>Notes on response:</p>
	<p>[Clarifying follow-up question(s)]: Would you define sociability differently? Would you add anything to the definition?</p> <p>Notes on response:</p>
	<p>Q4: On a scale of 1-5, one being not important at all and five being of utmost importance, how would you rate the importance of sociability in the context of public spaces? <i>Linked objectives: 8</i></p> <p>Notes on response:</p>
	<p>[Clarifying follow-up question(s)]:</p> <p>Notes on response:</p>
	<p>Q5: As a [insert role] is sociability among your top 3 priorities to maintain? If not, how far down? If so, is it your top priority? <i>Linked objectives: 8</i></p> <p>Notes on response:</p>

	[Clarifying follow-up question(s)]:
	Notes on response:

<p>NExploring sociability in public open spaces <i>20-25 Minutes</i></p>	<p>Q6: As a [insert role] what activities would you like to see most happen in your park spaces? <i>Linked objectives: 1, 2</i></p> <p>Notes on response:</p>
	[Clarifying follow-up question(s)]:
	Notes on response:
	<p>Q7: What are target goals for your social programming? <i>Linked objectives: 1</i></p> <p>Notes on response:</p>
	[Clarifying follow-up question(s)]:
	Notes on response:
	<p>Q8: What are target goals for how patrons move through your space? <i>Linked objectives: 1</i></p> <p>Notes on response:</p>
	[Clarifying follow-up question(s)]:
	Notes on response:
	[Clarifying follow-up question(s)]:
	Notes on response:

	<p>[Clarifying follow-up question(s)]: Are there any programs/interventions/features purposefully designed to move patrons from one space to another?</p> <p>Notes on response:</p>
	<p>Q9: What impacts on park use and behavior does a typical feature (fire-pit, fountain, play equipment) need to create in order to be considered successful? <i>Linked objectives: 2</i></p> <p>Notes on response:</p>
	<p>[Clarifying follow-up question(s)]:</p> <p>Notes on response:</p>
	<p>Q10: Describe a scene in a park setting with a high degree of sociability. <i>Linked objectives: 3,4,5</i></p> <p>Notes on response:</p>
	<p>[Clarifying follow-up question(s)]: In the described scene, what is most important, or what do you think contributes most to sociability? Could you rate on a scale of 1-5 the various activities/behaviors mentioned on how important they are regarding sociability?</p> <p>Notes on response:</p>
	<p>Q11: What physical activities provide evidence of a space being sociable? <i>Linked objectives: 3</i></p> <p>Notes on response:</p>

	[Clarifying follow-up question(s)]:
	Notes on response:
	Q12: What social behaviors provide evidence of a space being sociable? <i>Linked objectives: 4</i>
	Notes on response:
	[Clarifying follow-up question(s)]:
	Notes on response:
	Q13: Rate the following activities, on a scale of 1-5, on how "sociable" you believe them to be, one being not sociable at all and five being extremely sociable:
	<ol style="list-style-type: none"> 1. Leisurely stroll through park space alone: 2. Leisurely stroll through park space with others: 3. Running/Jogging through park space alone: 4. Running/Jogging through park space with others: 5. Sitting on a bench alone: 6. Sitting on a bench with others: 7. Dining alone on park tables: 8. Dining with others on park tables: 9. Loitering alone: 10. Loitering/conversing with friends: 11. Loitering/conversing with strangers: 12. Organized group play(volleyball nets, yoga in green space, etc.): 13. Biking/skating/scootering through park space alone: 14. Biking/skating/scootering through park space with others: 15. Engaging/viewing park scenery (art piece, gardens, fountain, river view, etc.) alone: 16. Engaging/viewing park scenery (art piece, gardens, fountain, river view, etc.) with others: 17. Walking pet alone: 18. Walking pet with others: 19. Engaging with park performer (listening to musician, watching magician, etc.):
	[Follow up question(s)]: Is there anything missing on this list that you would consider to be highly sociable (rating of 4 or higher)?
	Notes on response:

	<p>Q14: What park infrastructure do you believe drives park use and increases sociability? <i>Linked objectives: 5</i></p>
	<p>Notes on response:</p>
	<p>[Clarifying follow-up question(s)]:</p>
	<p>Notes on response:</p>

<p>Data on public use and behavior in park spaces <i>5-10 Minutes</i></p>	<p>Q15: In your opinion, what are the most important data you could gather about your park? <i>Linked objectives: 6,7</i></p>
	<p>Notes on response:</p>
	<p>[Clarifying follow-up question(s)]:</p>
	<p>Notes on response:</p>
	<p>Q16: What activities do you believe are important to observe and track? <i>Linked objectives: 7</i></p>
	<p>Notes on response:</p>
	<p>[Clarifying follow-up question(s)]:</p>
	<p>Notes on response:</p>

	<p>Q17: What social behaviors do you believe are important to observe and track? <i>Linked objectives: 7</i></p>
	<p>Notes on response:</p>
	<p>[Clarifying follow-up question(s)]:</p>
	<p>Notes on response:</p>
	<p>Q18: How can data about your park help you manage, operate, and design your park spaces? <i>Linked objectives: 7</i></p>
<p>Notes on response:</p>	

<p>Web dashboard <i>5-10 Minutes</i></p>	<p>Q19: How would you like to view and engage with sociability data? (Charts, graphs, toplines, interactive?) <i>Linked objectives: 6</i></p>
	<p>Notes on response:</p>
	<p>[Clarifying follow-up question(s)]:</p>
	<p>Notes on response:</p>
	<p>Q20: Given a specific scene, what would you like to know most regarding sociability? (Trends, categorized by activity, social layer, location.) <i>Linked objectives: 6, 7</i></p>
	<p>Notes on response:</p>

	[Clarifying follow-up question(s)]:
	Notes on response:
	Q21: What are important/desired features of a sociability dashboard? (ability to export raw data, compatibility with mobile screens, etc.) <i>Linked objectives: 6</i>
	Notes on response:
	[Clarifying follow-up question(s)]:
	Notes on response:

Appendix B: Sociability Interview Results

Exploring sociability in public open spaces and how it relates to performance and desired user experience
Park manager/owner/investor interview structure

Overview and Objectives:





















































To better design, operate, and manage public open spaces it is imperative for owners and investors to understand how patrons use their spaces and derive benefits from them. With recent advances in wireless sensing and computing technologies, urban sensing strategies can go beyond traffic counting and capture deeper and more complex activities and social behaviors. An urban sensing framework designed around capturing 'sociability' – by tracking activities, interactions, and social behaviors in the park can empower park managers, owners, and investors to make data driven decisions and improve the experiences of park patrons. **To this end, these interviews seek to work with park managers, owners, and investors to understand and identify desired patron activities and behaviors which are perceived to increase sociability.** A sensing toolset that can capture desired outputs of social programming and park features will enable park managers to better assess their performance and drive investment in features and programming which are achieving their goals. Furthermore, these interviews seek to understand how park managers, owners, and investors wish to engage with sociability data to ensure the utility of the web dashboard -- by featuring impactful sociability analytics and visualizations.



Interview Objectives:



















1. Identify desired outputs (activities and behaviors) from social programming and how they relate to sociability
2. Identify desired outputs from park features (benches, fountains, fire pits, etc.) and how they relate to sociability
3. Identify common activities in public open spaces and how they are perceived to relate to sociability
4. Identify public social behaviors perceived to contribute to sociability
5. Identify physical features and structures perceived to be relevant to sociability or help contribute to sociability
6. Identify desired web dashboard features and visualizations
7. Identify key sociability-related analytics
8. Ascertain importance of sociability and user experience to park managers/owners/investors

















Sociability: Pulling from past research, a sociable public space has been defined as an environment that enables individuals to comfortably and securely pursue their activities while engaging in social interactions, participating in events, ceremonies, and spectacles, or merely sitting and waiting. These spaces foster social and recreational endeavors, whether pursued individually or collectively, and provide individuals with a venue for bonding and gathering. In the context of this study, the sociability of a space refers to its present capacity to facilitate the formation of a social environment. A sociability index would describe the current level of social interaction, engagement, and activities supported by the space.

Background and importance of sociability 5-10 Minutes	Q1: What is your official title and position in [insert subjects affiliated organization/group]? <i>Linked objectives: NA</i> Job titles of subject: Executive : 👤👤👤👤👤 (5) -CEO (x2) -Executive director (x2) -President Management : 👤👤👤👤👤👤👤👤👤👤 (10) -Director of Programs (x2) -Public spaces manager (x2) -Program manager (x3) -Community engagement (x2) -Volunteer manager Affiliated Parks/Organization: Detroit Riverfront : 👤👤👤👤👤 (5) St. Louis Arch : 👤👤👤👤 (4) Detroit Parks Coalition : 👤👤 (2) Belle Isle : 👤 (1) Huron Metro Parks : 👤 (1) Willamette Falls, Oregon City: 👤 (1) Memphis Parks Riverfront : 👤 (1)
	[Clarifying follow-up question(s)]: Notes on response:
	Q2: Could you describe in detail the main duties and responsibilities of your role? <i>Linked objectives: NA</i> Mentioned duties/responsibilities: • Develop social programs : 👤👤👤👤👤 (5) • Secure funding : 👤👤👤👤👤 (5)

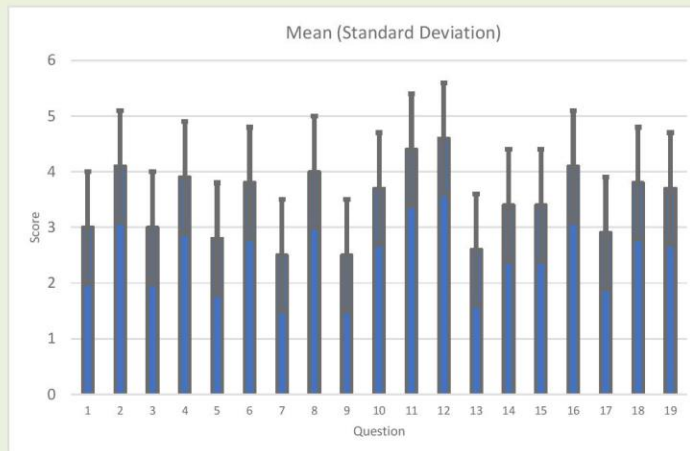
	<ul style="list-style-type: none"> • Develop Vision :     (4) • Find Partners :     (4) • Attract visitors into space :   (2) • Assure public access :   (2) • Hire staff :   (2) • Data collection on users :   (2) • Training :  (1)
	<p>[Clarifying follow-up question(s)]: Notes on response:</p>
	<p>Q3: What do you think of when I say “sociability” in the context of public open spaces?[Read section on sociability and the team’s definition and use of the term in the context of our study]. Would you agree with this definition of sociability in the context of public spaces? <i>Linked objectives: NA</i></p>
	<p>Responses:</p> <ul style="list-style-type: none"> • Interacting with others and building community :         (8) • Being around others / cohabitating / gathering :         (8) • Inclusive Spaces/ Diverse demographics :      (5) • Welcoming atmosphere :     (4) • Safe environment :  (1) • Accessible / ADA :  (1)
	<p>[Clarifying follow-up question(s)]: Would you define sociability differently? Would you add anything to the definition?</p> <p>Notes on response: All participants thought the provided definition was fair with three offering suggestions on edits to the definition.</p> <ul style="list-style-type: none"> • Add parts to the definition that specifically addresses safety and inclusivity of space   (2) • Add a part regarding interaction with strangers specifically  (1)
	<p>Q4: On a scale of 1-5, one being not important at all and five being of utmost importance, how would you rate the importance of sociability in the context of public spaces? <i>Linked objectives: 8</i></p> <p>Notes on response: All participants answered ‘5’</p> <ul style="list-style-type: none"> • Sociability is CRITICAL    (3) • Sociability is Foundational   (2) <p><i>“Building community is the #1 priority.” -Subject 2368</i></p>
	<p>[Clarifying follow-up question(s)]: Notes on response:</p>

	<p>Q5: As a [insert role] is sociability among your top 3 priorities to maintain? If not, how far down? If so, is it your top priority?</p> <p><i>Linked objectives: 8</i></p> <p>Notes on response: # 1 priority :  (10) Top 3 but not # 1 :  (5)</p>
	<p>[Clarifying follow-up question(s)]: Notes on response:</p>

<p>Exploring sociability in public open spaces 20-25 Minutes</p>	<p>Q6: As a [insert role] what activities would you like to see most happen in your park spaces?</p> <p><i>Linked objectives: 1, 2</i></p> <p>Notes on response:</p> <ul style="list-style-type: none"> • Playful activities/recreation activities :  (8) • Health and wellness :  (5) • Engaged with nature :  (5) <ul style="list-style-type: none"> - Fishing, interacting with garden, Enjoying river view, bird watching • Engaged with others :  (4) • New connections :  (3) • Nature conservation :  (2) • Accessible activities :  (2)
	<p>[Clarifying follow-up question(s)]: Notes on response:</p>
	<p>Q7: What are target goals for your social programming?</p> <p><i>Linked objectives: 1</i></p> <p>Notes on response:</p> <ul style="list-style-type: none"> • Build community :  (8) • Make the place approachable :  (8) • Increase patron time on site :  (7) • Encourage mental wellness activities :  (5) • Drive traffic :  (3) • Expose people to new things :  (2) • Create spontaneous interactions :  (2) • Outdoor education :  (2) • Build relationship with space :  (2) • Bring in/support local business :  (1) • Highlight diverse cultures and minorities :  (1) <p><i>"Let's give more people easy opportunities to be fit to live healthy lives. Let's give easy opportunities for people to protect their mental health. Let's give people, you know, easy opportunities to have fun. I think all of those are really important... I think to the extent that those activities also encourage people</i></p>

<p><i>to meet strangers - Then I think that's even better." --Subject 1989</i></p>	
<p>[Clarifying follow-up question(s)]: Notes on response:</p>	
<p>Q8: What impacts on park use and behavior does a typical feature (fire-pit, fountain, play equipment) need to create in order to be considered successful? <i>Linked objectives: 2</i></p>	
<p>Notes on response:</p> <ul style="list-style-type: none"> • Facilitate conversation and connectivity :  (11) • Increase comfort/sense of belonging :  (5) • increase accessibility :  (5) • Does it fit a need? :  (5) -sit/respice/shade • People are using it :  (4) • Facilitate connection to nature :  (2) • Encourage people to spend more time in space :  (2) • Self sustaining :  (2) • Facilitate interactions between strangers :  (1) 	
<p>[Clarifying follow-up question(s)]: Notes on response:</p>	
<p>Q9: Describe a specific situation in a park setting with a high degree of sociability. <i>Linked objectives: 3,4,5</i></p>	
<p>Notes on response:</p> <ul style="list-style-type: none"> • Diverse activities (10) :  (10) • People are comfortable and feel welcome :  (5) • Generational gatherings (children, parents, grandparents) :  (6) • Strangers conversing :  (4) • Happiness/joy/laughter :  (4) • Cultural activity (singing/dance, etc.) :  (3) • Outdoor games :  (2) 	
<p>[Clarifying follow-up question(s)]: Notes on response:</p>	
<p>Q10: Rate the following activities we can detect from our videos, on a scale of 1-5, on how "sociable" you believe them to be, one being not sociable at all and five being extremely sociable: Mean (STDEV)</p> <ol style="list-style-type: none"> 1. Leisurely stroll through park space alone: 3.0 (1.2) 2. Leisurely stroll through park space with others: 4.1 (0.6) 	

3. Running/Jogging through park space alone: **3.0 (1.4)**
4. Running/Jogging through park space with others: **3.9 (0.7)**
5. Sitting on a bench alone: **2.8 (1.4)**
6. Sitting on a bench with others: **3.8 (0.6)**
7. Dining alone on park tables: **2.5 (1.5)**
8. Dining with others on park tables: **4.0 (0.8)**
9. Loitering alone: **2.5 (1.5)**
10. Loitering/conversing with friends: **3.7 (1.0)**
11. Loitering/conversing with strangers: **4.4 (0.6)**
12. Organized group play(volleyball nets, yoga in green space, etc.): **4.6 (0.5)**
13. Biking/skating/scootering through park space alone: **2.6 (1.3)**
14. Biking/skating/scootering through park space with others: **3.4 (0.9)**
15. Engaging/viewing park scenery (art piece, gardens, fountain, river view, etc.) alone: **3.4 (1.1)**
16. Engaging/viewing park scenery (art piece, gardens, fountain, river view, etc.) with others: **4.1 (0.8)**
17. Walking pet alone: **2.9 (1.3)**
18. Walking pet with others: **3.8 (0.6)**
19. Engaging with park performer (listening to musician, watching magician, etc.): **3.7 (0.9)**



Average score for activities of people alone: 2.8 (1.3)

Average score for activities of people with others: 4.0 (0.7)

-Activities with other on average scored 1.2 points higher and had lower standard deviation suggesting greater agreement on high scores for groups

-Average score for activities of movement through space (with others): 3.8 (0.7)

-Average score for activities of people staying on site (with others): 4.1 (0.7)

-Slight increase in scores for activities requiring people to stay on site

Additional Comments:

While a majority of respondents generally gave higher scores to those engaging in activities with others versus alone a couple of respondents had differing opinions and fresh insights – stating they thought when people come alone to a space it's an indicator of the space being welcoming. The idea is if someone is comfortable enough to sit in the space alone it is reflective of the space feeling safe.

"But I think that you know the solo suggests some amount of comfort with what's going on...it's a reflection that the space makes them feel comfortable and safe to be able to take a walk by themselves....Is that person clearly a member of a vulnerable group in some way or other? And that that would even, to me, expand the sociability of the place" --Subject 0037













[Follow up question(s)]: Is there anything missing on this list that you would consider to be highly sociable (rating of 4 or higher)?

Notes on response:

- Sponsored programming : 📍📍📍📍📍📍📍📍 (8)
- Non-dominant cultural events : 📍 (1)
- Social dancing/singing : 📍 (1)
- Volunteering events : 📍 (1)

Q11: What park infrastructure do you believe drives park use and increases sociability?
Linked objectives: 5

Notes on response:
















- Restrooms :  (8)
- Nature :  (6)
- Parking :  (5)
- Play equipment :  (5)
- Clean, safe – lighting and trashcans :  (4)
- Equipment rentals :  (3)
- Public gathering space :  (2)
- ADA accessibility :  (2)
- Large event spaces/infrastructure :  (2)
- Wifi :  (1)
- Food service :  (1)
- Way finding :  (1)

[Clarifying follow-up question(s)]:
Notes on response:



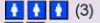


















Data on public use and behavior in park spaces
5-10 Minutes

















Q12: In your opinion, what are the most important data you could gather about your park?
Linked objectives: 6,7

Notes on response:

- time on site/duration of use :  (11)
- origins :  (10)
- heatmap/most used locations :  (7)
- demographics/age :  (7)
- interactions with furniture :  (6)
- traffic/volume of use :  (5)
- return visits :  (4)
- are they connecting :  (3)
- public perceptions on safety of space :  (2)
- purpose of visit :  (2)
- air and sound quality :  (1)
- health of fauna :  (1)
- entrance points :  (1)
- time of day people visit :  (1)
- mode of transportation to park :  (1)

[Clarifying follow-up question(s)]:
Notes on response:

	<p>Q13: What activities do you believe are important to observe and track? <i>Linked objectives: 7</i></p> <p>Notes on response:</p> <ul style="list-style-type: none"> walking/biking :  (7) using play equipment as designed :  (4) scooters :  (3) alternative modes of mobility (wheelchairs) :  (2) picture taking :  (2) dining :  (2) Bringing pets? :  (2) mindful moments exercises :  (1) fishing :  (1) littering :  (1) <p>-Scooters came up in the discussions as potential negative behaviors/activities. Multiple spaces get complaints from Patrons regarding rental electric scooter use (Bird, Lime, etc.). Respondents mentioned it would be nice to track to see how valid the complaints are and to better understand how they are being used to help when designing a solution.</p> <p>[Clarifying follow-up question(s)]: Notes on response:</p>
	<p>Q14: What social behaviors do you believe are important to observe and track? <i>Linked objectives: 7</i></p> <p>Notes on response:</p> <ul style="list-style-type: none"> spontaneous interactions with strangers :  (8) Are people with others? :  (7) sharing (sharing amenities/assets) :  (2) different demographics connecting :  (2) are people happy? :  (2) indicators of dissatisfaction :  (1) sleeping in park :  (1) do people help others :  (1) <p>"Does the space and the way it's managed encourage people to interact with strangers. I'll put it that way. And particularly again with strangers who don't look like them, whether it's age, or whether it's, you know, income." –subject 1989</p> <p>[Clarifying follow-up question(s)]: Notes on response:</p>
<p>Web dashboard 5-10 Minutes</p>	<p>Q15: How would you like to view and engage with sociability data? (Charts, graphs, toplines, interactive?) <i>Linked objectives: 6</i></p> <p>Notes on response:</p> <ul style="list-style-type: none"> charts & graphs of day to day :  (8) interactive :  (7) highest trafficked area/zones – heatmap :  (6)

<ul style="list-style-type: none"> filters (weather/day of week) :  (5) areas that are over performing :  (3) emerging trends :  (3) pro-social behavior vs anti-social behavior :  (1) -(spectrum bar at top)
<p>[Clarifying follow-up question(s)]: Notes on response:</p>
<p>Q16: Given a specific scene, what would you like to know most regarding sociability? (Trends, categorized by activity, social layer, location.) <i>Linked objectives: 6, 7</i></p> <p>Notes on response:</p> <ul style="list-style-type: none"> are people engaging in pro-social behaviors? :  (10) locations of social interactions :  (7) most common activities for groups/social gatherings :  (3) time spent with others :  (1) breakdown by demographic :  (1)
<p>[Clarifying follow-up question(s)]: Notes on response:</p>
<p>Q17: What are important/desired features of a sociability dashboard? (ability to export raw data, compatibility with mobile screens, etc.) <i>Linked objectives: 6</i></p> <p>Notes on response:</p> <ul style="list-style-type: none"> easy to digest – simple :  (9) export high quality figures :  (5) export raw data :  (3) practical :  (2) unique logins :  (2) secure :  (1) import data from elsewhere :  (1) <p><i>"something that's like really easy to comprehend – on a digestible level for a park manager. Right? I think that it's just a general difference between like Academia, and you know practical life is that like we need to be able to quickly understand. So infographics, or something that illustrates data in a real kind of fast and simple way" – Subject 0029</i></p>
<p>[Clarifying follow-up question(s)]: Notes on response:</p>

Bibliography

- [1] W. Whyte and P. Underhill, *City: Rediscovering the Center*. University of Pennsylvania Press, Incorporated, 1988.
- [2] E. Talen and C. Ellis, “Beyond Relativism: Reclaiming the Search for Good City Form,” *J. Plan. Educ. Res.*, vol. 22, no. 1, pp. 36–49, Sep. 2002, doi: 10.1177/0739456X0202200104.
- [3] K. Magis, “Community Resilience: An Indicator of Social Sustainability,” *Soc. Nat. Resour.*, vol. 23, no. 5, pp. 401–416, Apr. 2010, doi: 10.1080/08941920903305674.
- [4] Y. Ben-Porath, “The F-Connection: Families, Friends, and Firms and the Organization of Exchange,” *Popul. Dev. Rev.*, vol. 6, no. 1, p. 1, Mar. 1980, doi: 10.2307/1972655.
- [5] S. L. Cutter, B. J. Boruff, and W. L. Shirley, “Social Vulnerability to Environmental Hazards *: Social Vulnerability to Environmental Hazards,” *Soc. Sci. Q.*, vol. 84, no. 2, pp. 242–261, Jun. 2003, doi: 10.1111/1540-6237.8402002.
- [6] C. Baldwin and R. King, *Social Sustainability, Climate Resilience and Community-Based Urban Development What About the People?*, 1st ed. Routledge, 2018.
- [7] E. Klinenberg, *Palaces for the People: How Social Infrastructure Can Help Fight Inequality, Polarization, and the Decline of Civic Life*. Crown, 2018.
- [8] S. T. West, K. A. Shores, and L. M. Mudd, “Association of available parkland, physical activity, and overweight in America’s largest cities,” *J. Public Health Manag. Pract. JPHMP*, vol. 18, no. 5, pp. 423–430, 2012, doi: 10.1097/PHH.0b013e318238ea27.
- [9] D. C. Geng, J. Innes, W. Wu, and G. Wang, “Impacts of COVID-19 pandemic on urban park visitation: a global analysis,” *J. For. Res.*, vol. 32, no. 2, pp. 553–567, 2021, doi: 10.1007/s11676-020-01249-w.
- [10] D. Scott, “Economic Inequality, Poverty, and Park and Recreation Delivery,” *J. Park Recreat. Adm.*, 2013, Accessed: Sep. 19, 2023. [Online]. Available: <https://www.semanticscholar.org/paper/Economic-Inequality%2C-Poverty%2C-and-Park-and-Delivery-Scott/88d4b1d17bf2b85fd82ca9a096fa6665fe8d83b8>
- [11] J. Gehl, *Life Between Buildings*. Danish Architectural Press, 1971.
- [12] J. Gehl, *Cities for People*. Island Press, 2010.
- [13] J. Gehl and brigitte svarre, *How to Study Public Life*. 2013.
- [14] J. Jacobs, *The Death and Life of Great American Cities*. Knopf Doubleday Publishing Group, 1961.
- [15] D. Trudeau, “New Urbanism as Sustainable Development?,” *Geogr. Compass*, vol. 7, Jun. 2013, doi: 10.1111/gec3.12042.
- [16] E. Talen, “Sense of Community and Neighbourhood Form: An Assessment of the Social Doctrine of New Urbanism,” *Urban Stud.*, vol. 36, no. 8, pp. 1361–1379, Jul. 1999, doi: 10.1080/0042098993033.

- [17] A. Zautra, J. Hall, and K. Murray, "Community Development and Community Resilience: An Integrative Approach," *Community Dev.*, vol. 39, no. 3, pp. 130–147, Jul. 2008, doi: 10.1080/15575330809489673.
- [18] L. Anderson, "The Public Realm: Exploring the City's Quintessential Social Territory By Lyn H. Lofland New York: Aldine de Gruyter, 1998 305 pp. \$52.95 (cloth), \$25.95 (paper)," *Symb. Interact.*, vol. 22, no. 3, pp. 285–287, 1999, doi: 10.1016/S0195-6086(99)80092-7.
- [19] V. Mehta, "Streets and social life in cities: a taxonomy of sociability," *URBAN Des. Int.*, vol. 24, no. 1, pp. 16–37, Mar. 2019, doi: 10.1057/s41289-018-0069-9.
- [20] W. Whyte, *The Social Life of Small urban Spaces*. Conservation Foundation, 1980.
- [21] J. C. Semenza, T. L. March, and B. D. Bontempo, "Community-Initiated Urban Development: An Ecological Intervention," *J. Urban Health*, vol. 84, no. 1, pp. 8–20, Jan. 2007, doi: 10.1007/s11524-006-9124-8.
- [22] S. Karuppannan and A. Sivam, "Social sustainability and neighbourhood design: an investigation of residents' satisfaction in Delhi," *Local Environ.*, vol. 16, no. 9, pp. 849–870, Oct. 2011, doi: 10.1080/13549839.2011.607159.
- [23] M. C. Le, M.-H. Le, and M.-T. Duong, "Vision-based People Counting for Attendance Monitoring System," in *2020 5th International Conference on Green Technology and Sustainable Development (GTSD)*, Nov. 2020, pp. 349–352. doi: 10.1109/GTSD50082.2020.9303117.
- [24] J.-W. Kim, K.-S. Choi, B.-D. Choi, and S. Ko, "Real-time Vision-based People Counting System for the Security Door," Jul. 2002. Accessed: Sep. 11, 2023. [Online]. Available: <https://www.semanticscholar.org/paper/Real-time-Vision-based-People-Counting-System-for-Kim-choi/24cd7fd7c429b4827cbcd606d79a7c352f2c726>
- [25] C. Ma, C. Wan, Y. W. Chau, S. M. Kang, and D. R. Selviah, "Subway station real-time indoor positioning system for cell phones," in *2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Sapporo: IEEE, Sep. 2017, pp. 1–7. doi: 10.1109/IPIN.2017.8115912.
- [26] M. Bertolusso, M. Spanu, M. Anedda, M. Fadda, and D. D. Giusto, "Vehicular and Pedestrian Traffic Monitoring System in Smart City Scenarios," in *2021 IEEE 7th World Forum on Internet of Things (WF-IoT)*, Jun. 2021, pp. 60–64. doi: 10.1109/WF-IoT51360.2021.9595188.
- [27] C. Slobogin, "Public Privacy: Camera Surveillance of Public Places And The Right to Anonymity," *SSRN Electron. J.*, Feb. 2003, doi: 10.2139/ssrn.364600.
- [28] S. Zhang, Y. Feng, A. Das, L. Cranor, and N. Sadeh, "Understanding People's Privacy Attitudes Towards Video Analytics Technologies," presented at the PrivacyCon, Washington D.C., 2020. Accessed: Aug. 03, 2023. [Online]. Available: <https://www.semanticscholar.org/paper/Understanding-People%E2%80%99s-Privacy-Attitudes-Towards-Zhang-Feng/b1af765c1a8dcc1693b813259f5b7e9dfde1a548>
- [29] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Kauai, HI, USA: IEEE Comput. Soc, 2001, p. I-511–I-518. doi: 10.1109/CVPR.2001.990517.
- [30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Jun. 2005, pp. 886–893 vol. 1. doi: 10.1109/CVPR.2005.177.

- [31] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.
- [32] W. Liu *et al.*, “SSD: Single Shot MultiBox Detector,” in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., in *Lecture Notes in Computer Science*. Cham: Springer International Publishing, 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0_2.
- [33] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2980–2988. doi: 10.1109/ICCV.2017.322.
- [34] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2015. Accessed: Oct. 14, 2022. [Online]. Available: <https://papers.nips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html>
- [35] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation.” arXiv, Oct. 22, 2014. doi: 10.48550/arXiv.1311.2524.
- [36] R. Girshick, “Fast R-CNN,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1440–1448. doi: 10.1109/ICCV.2015.169.
- [37] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition.” arXiv, Apr. 10, 2015. doi: 10.48550/arXiv.1409.1556.
- [38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2012. Accessed: Jun. 26, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html
- [39] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
- [40] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature Pyramid Networks for Object Detection,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 936–944. doi: 10.1109/CVPR.2017.106.
- [41] J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” arXiv.org. Accessed: Jun. 26, 2023. [Online]. Available: <https://arxiv.org/abs/1612.08242v1>
- [42] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement”.
- [43] T.-Y. Lin *et al.*, “Microsoft COCO: Common Objects in Context,” in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., in *Lecture Notes in Computer Science*. Cham: Springer International Publishing, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.
- [44] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL: IEEE, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.

- [45] P. Sun, R. Hou, and J. P. Lynch, “Measuring the Utilization of Public Open Spaces by Deep Learning: a Benchmark Study at the Detroit Riverfront,” in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass Village, CO, USA: IEEE, Mar. 2020, pp. 2217–2226. doi: 10.1109/WACV45572.2020.9093336.
- [46] Y. Ma, S. Soatto, J. Košecká, and S. S. Sastry, *An Invitation to 3-D Vision*, vol. 26. in *Interdisciplinary Applied Mathematics*, vol. 26. New York, NY: Springer New York, 2004. doi: 10.1007/978-0-387-21779-6.
- [47] Z. Zhang, “A flexible new technique for camera calibration,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000, doi: 10.1109/34.888718.
- [48] C. D. Fryar, Q. Gu, C. L. Ogden, and K. M. Flegal, “Anthropometric Reference Data for Children and Adults: United States, 2011-2014,” *Vital Health Stat. 3.*, no. 39, pp. 1–46, Aug. 2016.
- [49] N. Wojke, A. Bewley, and D. Paulus, “Simple online and realtime tracking with a deep association metric,” in *2017 IEEE International Conference on Image Processing (ICIP)*, Beijing, China: IEEE Press, Sep. 2017, pp. 3645–3649. doi: 10.1109/ICIP.2017.8296962.
- [50] H. W. Kuhn, “The Hungarian method for the assignment problem,” *Nav. Res. Logist. Q.*, vol. 2, no. 1–2, pp. 83–97, 1955, doi: 10.1002/nav.3800020109.
- [51] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple online and realtime tracking,” in *2016 IEEE International Conference on Image Processing (ICIP)*, Sep. 2016, pp. 3464–3468. doi: 10.1109/ICIP.2016.7533003.
- [52] S. Zagoruyko and N. Komodakis, “Wide Residual Networks.” arXiv, Jun. 14, 2017. doi: 10.48550/arXiv.1605.07146.
- [53] L. Zheng *et al.*, “MARS: A Video Benchmark for Large-Scale Person Re-Identification,” in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., in *Lecture Notes in Computer Science*. Cham: Springer International Publishing, 2016, pp. 868–884. doi: 10.1007/978-3-319-46466-4_52.
- [54] N. Wojke and A. Bewley, “Deep Cosine Metric Learning for Person Re-identification,” in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV: IEEE, Mar. 2018, pp. 748–756. doi: 10.1109/WACV.2018.00087.
- [55] M. Jiang and X. Fan, “RetinaMask: A Face Mask detector,” *ArXiv*, May 2020, Accessed: Jun. 26, 2023. [Online]. Available: <https://www.semanticscholar.org/paper/RetinaMask%3A-A-Face-Mask-detector-Jiang-Fan/6d5adc9ae499fcaa62dd1f02b45afce877427253>
- [56] S. Ge, J. Li, Q. Ye, and Z. Luo, “Detecting Masked Faces in the Wild with LLE-CNNs,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 426–434. doi: 10.1109/CVPR.2017.53.
- [57] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal Visual Object Classes (VOC) Challenge,” *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010, doi: 10.1007/s11263-009-0275-4.
- [58] K. Bernardin and R. Stiefelhagen, “Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics,” *EURASIP J. Image Video Process.*, vol. 2008, no. 1, pp. 1–10, Dec. 2008, doi: 10.1155/2008/246309.
- [59] A. Milan, L. Leal-Taixé, I. Reid, and S. Roth, “MOT16: A Benchmark for Multi-Object Tracking,” Mar. 2016.

- [60] “Density Estimation for Statistics and Data Analysis,” Routledge & CRC Press. Accessed: Jun. 26, 2023. [Online]. Available: <https://www.routledge.com/Density-Estimation-for-Statistics-and-Data-Analysis/Silverman/p/book/9780412246203>
- [61] F. Mirahadi and B. McCabe, “EvacuSafe: Building Evacuation Strategy Selection Using Route Risk Index,” *J. Comput. Civ. Eng.*, vol. 34, no. 2, p. 04019051, Mar. 2020, doi: 10.1061/(ASCE)CP.1943-5487.0000867.
- [62] Y. Yu *et al.*, “Automatic Biomechanical Workload Estimation for Construction Workers by Computer Vision and Smart Insoles,” *J. Comput. Civ. Eng.*, vol. 33, no. 3, p. 04019010, May 2019, doi: 10.1061/(ASCE)CP.1943-5487.0000827.
- [63] B. Xiao and Z. Zhu, “Two-Dimensional Visual Tracking in Construction Scenarios: A Comparative Study,” *J. Comput. Civ. Eng.*, vol. 32, no. 3, p. 04018006, May 2018, doi: 10.1061/(ASCE)CP.1943-5487.0000738.
- [64] T. Cheng, G. C. Migliaccio, J. Teizer, and U. C. Gatti, “Data Fusion of Real-Time Location Sensing and Physiological Status Monitoring for Ergonomics Analysis of Construction Workers,” *J. Comput. Civ. Eng.*, vol. 27, no. 3, pp. 320–335, May 2013, doi: 10.1061/(ASCE)CP.1943-5487.0000222.
- [65] K. Kang *et al.*, “T-CNN: Tubelets With Convolutional Neural Networks for Object Detection From Videos,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2896–2907, Oct. 2018, doi: 10.1109/TCSVT.2017.2736553.
- [66] D. Roberts and M. Golparvar-Fard, “End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level,” *Autom. Constr.*, vol. 105, p. 102811, Sep. 2019, doi: 10.1016/j.autcon.2019.04.006.
- [67] X. Luo, H. Li, H. Wang, Z. Wu, F. Dai, and D. Cao, “Vision-based detection and visualization of dynamic workspaces,” *Autom. Constr.*, vol. 104, pp. 1–13, Aug. 2019, doi: 10.1016/j.autcon.2019.04.001.
- [68] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, “Fully-Convolutional Siamese Networks for Object Tracking,” in *Computer Vision – ECCV 2016 Workshops*, vol. 9914, G. Hua and H. Jégou, Eds., in Lecture Notes in Computer Science, vol. 9914. , Cham: Springer International Publishing, 2016, pp. 850–865. doi: 10.1007/978-3-319-48881-3_56.
- [69] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, “Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking”.
- [70] Y. Zhang *et al.*, “ByteTrack: Multi-object Tracking by Associating Every Detection Box,” in *Computer Vision – ECCV 2022*, vol. 13682, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., in Lecture Notes in Computer Science, vol. 13682. , Cham: Springer Nature Switzerland, 2022, pp. 1–21. doi: 10.1007/978-3-031-20047-2_1.
- [71] P. Dendorfer *et al.*, *MOT20: A benchmark for multi object tracking in crowded scenes*. 2020.
- [72] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 3354–3361. doi: 10.1109/CVPR.2012.6248074.
- [73] M. Kristan *et al.*, “A Novel Performance Evaluation Methodology for Single-Target Trackers,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 99, Mar. 2015, doi: 10.1109/TPAMI.2016.2516982.
- [74] J. Luiten *et al.*, “HOTA: A Higher Order Metric for Evaluating Multi-object Tracking,” *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 548–578, Feb. 2021, doi: 10.1007/s11263-020-01375-2.

- [75] D. Das, “Urban Quality of Life: A Case Study of Guwahati,” *Soc. Indic. Res.*, vol. 88, no. 2, pp. 297–310, Sep. 2008, doi: 10.1007/s11205-007-9191-6.
- [76] K. Zakariya, N. Z. Harun, and M. Mansor, “Spatial Characteristics of Urban Square and Sociability: A Review of the City Square, Melbourne,” *Procedia - Soc. Behav. Sci.*, vol. 153, pp. 678–688, Oct. 2014, doi: 10.1016/j.sbspro.2014.10.099.
- [77] K. Lynch, *Good City Form*. MIT Press, 1984.
- [78] F. Tibbalds, *Making People-Friendly Towns: Improving the Public Environment in Towns and Cities*. Taylor & Francis, 2012.
- [79] E. Mossop, *Public Space: Civilising the City*. Fine Art Publishing, 2001. Accessed: Jul. 11, 2023. [Online]. Available: <https://opus.lib.uts.edu.au/handle/10453/131509>
- [80] S. Jalaladdini and D. Oktay, “Urban Public Spaces and Vitality: A Socio-Spatial Analysis in the Streets of Cypriot Towns,” *Procedia - Soc. Behav. Sci.*, vol. 35, pp. 664–674, Jan. 2012, doi: 10.1016/j.sbspro.2012.02.135.
- [81] F. Alves, S. Santos Cruz, A. Ribeiro, A. Silva, J. Martins, and I. Cunha, “Walkability Index for Elderly Health: A Proposal,” *Sustainability*, vol. 12, p. 7360, Sep. 2020, doi: 10.3390/su12187360.
- [82] M. O. Leavitt, “2008 Physical Activity Guidelines for Americans”.
- [83] N. Dempsey, G. Bramley, S. Power, and C. Brown, “The social dimension of sustainable development: Defining urban social sustainability,” *Sustain. Dev.*, vol. 19, no. 5, pp. 289–300, 2011, doi: 10.1002/sd.417.
- [84] “Urban Form and Social Sustainability: The Role of Density and Housing Type.” Accessed: Sep. 21, 2023. [Online]. Available: <https://journals.sagepub.com/doi/epdf/10.1068/b33129>
- [85] F. Pezoa, J. L. Reutter, F. Suarez, M. Ugarte, and D. Vrgoč, “Foundations of JSON Schema,” in *Proceedings of the 25th International Conference on World Wide Web*, in WWW ’16. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, Apr. 2016, pp. 263–273. doi: 10.1145/2872427.2883029.
- [86] T. Leeper, “AWS S3 Client Package.” 2020.
- [87] “Amazon Web Services: Elastic Beanstalk.” [Online]. Available: <https://aws.amazon.com/elasticbeanstalk/>
- [88] S. Hossain, “Visualization of Bioinformatics Data with Dash Bio,” presented at the Python in Science, Jan. 2019, pp. 126–133. doi: 10.25080/Majora-7ddc1dd1-012.
- [89] “Maps and Mapping | U.S. Geological Survey.” Accessed: Aug. 01, 2023. [Online]. Available: <https://www.usgs.gov/science/faqs/maps-and-mapping>
- [90] G. V. Brummelen, *Heavenly Mathematics: The Forgotten Art of Spherical Trigonometry*. Princeton University Press, 2013.
- [91] M. Shaha and M. Pawar, “Transfer Learning for Image Classification,” in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Mar. 2018, pp. 656–660. doi: 10.1109/ICECA.2018.8474802.
- [92] G. Thün and K. Velikov, “Rethinking the Delivery of Social Infrastructure: A Systems Based Methodology and Proposition,” presented at the The City and Complexity – Life, Design and Commerce in the Built Environment, University of London, London, England, 2020.
- [93] A. Rigolon and J. Németh, “‘We’re not in the business of housing:’ Environmental gentrification and the non-profitization of green infrastructure projects,” *Cities*, vol. 81, pp. 71–80, Nov. 2018, doi: 10.1016/j.cities.2018.03.016.
- [94] U. Nations, *The Sustainable Development Goals Report 2022*. United Nations, 2022.

- [95] “The Sustainable Development Goals Report 2021,” UN DESA Publications. Accessed: Aug. 02, 2023. [Online]. Available: <http://desapublications.un.org/publications/sustainable-development-goals-report-2021>
- [96] Faris. A. Almalki *et al.*, “Green IoT for Eco-Friendly and Sustainable Smart Cities: Future Directions and Opportunities,” *Mob. Netw. Appl.*, Aug. 2021, doi: 10.1007/s11036-021-01790-w.
- [97] S. Makani, R. Pittala, E. Alsayed, M. Aloqaily, and Y. Jararweh, “A survey of blockchain applications in sustainable and smart cities,” *Clust. Comput.*, vol. 25, no. 6, pp. 3915–3936, Dec. 2022, doi: 10.1007/s10586-022-03625-z.
- [98] K. Saravanan and G. Sakthinathan, *Handbook of Green Engineering Technologies for Sustainable Smart Cities*. CRC Press, 2021.
- [99] S. E. Bibri, A. Alexandre, A. Sharifi, and J. Krogstie, “Environmentally sustainable smart cities and their converging AI, IoT, and big data technologies and solutions: an integrated approach to an extensive literature review,” *Energy Inform.*, vol. 6, no. 1, p. 9, Apr. 2023, doi: 10.1186/s42162-023-00259-2.
- [100] M. Gourisaria, G. Jee, H. Gm, D. Konar, and P. Singh, “Artificially Intelligent and Sustainable Smart Cities,” 2022, pp. 237–268. doi: 10.1007/978-3-031-08815-5_14.
- [101] H. Ahvenniemi, A. Huovila, I. Pinto-Seppä, and M. Airaksinen, “What are the differences between sustainable and smart cities?,” *Cities*, vol. 60, pp. 234–245, Feb. 2017, doi: 10.1016/j.cities.2016.09.009.
- [102] M. Angelidou, A. Psaltoglou, N. Komninos, C. Kakderi, P. Tsarchopoulos, and A. Panori, “Enhancing sustainable urban development through smart city applications,” *J. Sci. Technol. Policy Manag.*, vol. 9, no. 2, pp. 146–169, Jan. 2017, doi: 10.1108/JSTPM-05-2017-0016.
- [103] J. Buolamwini and T. Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,” in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, PMLR, Jan. 2018, pp. 77–91. Accessed: Aug. 02, 2023. [Online]. Available: <https://proceedings.mlr.press/v81/buolamwini18a.html>
- [104] N. Furl, P. J. Phillips, and A. J. O’Toole, “Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis,” *Cogn. Sci.*, vol. 26, no. 6, pp. 797–815, Nov. 2002, doi: 10.1016/S0364-0213(02)00084-8.
- [105] H. E. Khiyari and H. Wechsler, “Face Verification Subject to Varying (Age, Ethnicity, and Gender) Demographics Using Deep Learning,” *J. Biom. Biostat.*, vol. 07, no. 04, 2016, doi: 10.4172/2155-6180.1000323.
- [106] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain, “Face Recognition Performance: Role of Demographic Information,” *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 6, pp. 1789–1801, Dec. 2012, doi: 10.1109/TIFS.2012.2214212.
- [107] “Gender and Racial Bias in Computer Vision.” Accessed: Aug. 02, 2023. [Online]. Available: <https://glair.ai/post/gender-and-racial-bias-in-computer-vision>
- [108] A. Vasquez, M. Kollmitz, A. Eitel, and W. Burgard, “Deep Detection of People and their Mobility Aids for a Hospital Robot,” in *2017 European Conference on Mobile Robots (ECMR)*, Sep. 2017, pp. 1–7. doi: 10.1109/ECMR.2017.8098665.
- [109] C. Whitzman, “Social Infrastructure in Tall Buildings: A Tale of Two Towers: Carolyn Whitzman, MA, MCIP,” in *Tall Buildings and Urban Habitat*, CRC Press, 2001.
- [110] J. A. Temple and A. J. Reynolds, “Benefits and Costs of Investments in Preschool Education: Evidence from the Child-Parent Centers and Related Programs.” Rochester, NY,

2007. Accessed: Sep. 21, 2023. [Online]. Available:
<https://papers.ssrn.com/abstract=1143662>