AUTOMATED VISION-BASED GENERATION OF EVENT STATISTICS FOR DECISION SUPPORT

A Dissertation
Presented to
The Academic Faculty

by

Gbolabo Ogunmakin

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
May 2016

AUTOMATED VISION-BASED GENERATION OF EVENT STATISTICS FOR DECISION SUPPORT

Approved by:

Dr. Patricio Vela, Advisor
School of ECE
Georgia Institute of Technology

Dr. Raheem Beyah
School of ECE
Georgia Institute of Technology

Dr. Linda Wills
School of ECE
Georgia Institute of Technology

Dr. Claudia Rebola
School of Industrial Design
Rhode Island School of Design

Dr. Ayanna Howard
School of ECE
Georgia Institute of Technology

Date Approved: April 04, 2016

To my parents, my sisters, and my fiancée.

ACKNOWLEDGEMENTS

First, I'd like to thank my advisor, Dr. Patricio Vela, for his support, guidance, and patience throughout my doctoral research. Thanks to my committee members, Dr. Linda Wills, Dr. Ayanna Howard, Dr. Raheem Beyah, and Dr. Claudia Rebola. Special thanks to Dr. Beyah for being a great mentor and encouraging me to push through in difficult times.

Thanks to the IVALab, past and present members. Thanks especially to Miguel Serrano, Luisa Fairfax, Bella Nguyen, and Hassan Kingravi. I couldn't have done it without you guys. Special thanks to Hassan for always being there whenever I needed him no matter what the situation was, I really appreciate it.

I'd like to thank my friends Nashlie Sephus, Temi Olubanjo, Barbara Nsiah, Kelly Miller, and Hope Akpan for all the support and encouragement. Special thanks to Hope for taking care of me during a very difficult time, I am forever indebted to you.

Lastly, I'd like to thank my family for their unconditional support and motivation. Special thanks to my parents, Olamiju Ogunmakin and Grace Ogunmakin, for always making sure that I had everything I needed to finish my doctoral studies. Special thanks to my sisters, Dolapo Ogunmakin and Bolanle Ogunmakin for always being there for me when I needed them. The biggest thanks goes to my fiancée, Busayo Abiola-Oke. I'm grateful for her love, her support, and her belief in me. Thanks to everyone who's ever helped me out through my PhD journey, we did it!

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . . . **iv**

LIST OF TABLES . . . . . **vii**

LIST OF FIGURES . . . . . **viii**

SUMMARY . . . . . **xi**

I INTRODUCTION **1**

1.1 Objectives and Contributions 2

1.2 Literature Review 6

1.2.1 Video Surveillance 7

1.2.2 Background Modeling 8

1.2.3 Target Re-identification 9

1.2.4 Target Tracking 10

1.2.5 Construction Site Work Sampling 12

1.2.6 Senior Housing Surveillance 17

II SYSTEM OVERVIEW **23**

2.1 Detector 23

2.2 Tracker 28

2.3 Target Re-Identification 30

2.4 Activity Status Estimation 34

2.5 Event Detection Processor 35

2.5.1 Work Sampling Processor 36

2.5.2 Interaction Processor 38

III CONSTRUCTION SITE WORK SAMPLING **39**

3.1 Introduction 39

3.2 Methodology 40

3.3 Results 44

3.3.1 Clough Undergraduate Learning Commons (CLC) Construction Site . . . 45
3.3.2 Engineered Biosystems Building . . . 48
3.3.3 Underground Parking Garage Josephsplatz . . . 52
3.3.4 Hospital Building translaTUM . . . 54
3.3.5 Discussion . . . 58
IV SENIOR HOUSING SURVEILLANCE . . . 63
4.1 Introduction . . . 63
4.2 Methodology . . . 63
4.2.1 Qualitative Observations . . . 64
4.2.2 Technology Interventions Implications . . . 66
4.2.3 Experimental Intervention . . . 67
4.3 User Interfaces . . . 69
4.4 Experimental Results . . . 73
4.4.1 Video Sequence 1 . . . 73
4.4.2 Video Sequence 2 . . . 75
4.4.3 Calvin Court . . . 75
4.4.4 User Interface Correction . . . 82
4.5 Discussion . . . 83
V CONCLUSION . . . 85
REFERENCES . . . 86

LIST OF TABLES

1	Sample event statistics table	37
2	Automated and manual tabulation of ground leveling task (minutes) .	45
3	Error % for the different construction sites	60
4	Data collection table	69
5	Total daily traffic and interactions on the 2^{nd} floor.	77
6	Total number of common area entrances and passes on the 2^{nd} floor. .	78
7	Total daily traffic and interactions on the 3^{rd} floor.	79
8	Total number of common area entrances and passes on the 3^{rd} floor. .	80
9	Total daily traffic and interactions on the 7^{th} floor.	81
10	Total number of common area entrances and passes on the 7^{th} floor. .	81
11	Average time it takes to process 60 minutes of video.	83

LIST OF FIGURES

1 Process flow for traditional and proposed automatic surveillance system. 4

2 Example of biometric re-identification methods 10

3 Example of appearance based re-identification [1] 10

4 Examples of different types of tracking. 11

5 Outline of the framework 24

6 Effects of feedback on background model 26

7 Sample shadow removal results 29

8 Different key templates of the same target as it traversed the scene . 33

9 New key template selected when target changes orientation 33

10 Target as it goes through a static occlusion and gets re-identified correctly when it re-appears 34

11 Activity status of a dump truck getting filled by an excavator 35

12 Dump truck state estimates for a video segment of 6 minutes duration, plus activity states in a pie chart. 37

13 Process flow for the automatic surveillance system. 40

14 ROI for CLC views 1 and 2. 41

15 Estimated background models for Clough construction site (view 1). 41

16 Estimated background models for Clough construction site (view 2). 41

17 EBB entrance/exit region and ROI 42

18 Estimated background models for EBB construction site. 42

19 Plan view of the Josephsplatz construction site with camera view overlay (opaque trapezoid). 43

20 Estimated background models for Josephsplatz construction site. . . 43

21 The translaTUM construction site information, video geometry, and image view. 44

22 Estimated background models for translaTUM construction site. . . 44

23 Trajectories for the different machines tracked (CLC). 45

24 Loader text output (CLC). 46

25 Loader States (CLC). 46
26 Pie chart of the ROI for the Clough construction work site (view 2). 47
27 Dump Truck States (CLC). 48
28 Bulldozer States (CLC). 48
29 Ground leveling States (CLC). 48
30 Tracking results for the EBB construction site 48
31 Aggregate statistics for the EBB construction site. 49
32 Inter-arrival times between trucks (EBB). 50
33 Time spent in the filling ROI (EBB). 50
34 Time spent in the hosing ROI (EBB). 51
35 Tracking results for the Josephsplatz construction site 51
36 Dump truck states (Josephsplatz). 51
37 Loading time per truck (Josephsplatz). 52
38 Inter-arrival times between dump trucks (Josephsplatz). 52
39 Soil removal statistics for Josephsplatz site 53
40 Tracking results for day 8 of the translaTUM construction site. 54
41 Tracking results for day 11 of the translaTUM construction site. . . . 55
42 Time spent per dump truck in the scene while being filled for day 8 (translaTUM) (Log scale). 56
43 Time spent per dump truck in the scene while being filled for day 11 (translaTUM) (Log scale). 57
44 Inter-arrival times between trucks for day 8 (translaTUM) (Log scale). 58
45 Inter-arrival times between trucks for day 11 (translaTUM) (Log scale). 59
46 Aggregate statistics for translaTUM. 59
47 Performance factor over time (translaTUM) 60
48 Estimated cumulative soil removed per 20 minute interval (translaTUM). 61
49 Qualitative studies in retirement communities. 65
50 Designed technology intervention. 67
51 Equipment Layout 69

52 Simple ground truth GUI 70
53 Visualization/Correction GUI 72
54 Process flow for the automatic surveillance system. 73
55 Sample Outputs from sequence 1 74
56 Art Piece States 74
57 Target 1 States 75
58 Sample Outputs from sequence 2 76
59 Total traffic on the 2nd Floor 77
60 Total traffic on the 3rd Floor 80
61 Total traffic on the 7th Floor 82

SUMMARY

Many tasks require surveillance and analysis in order to make decisions regarding the next course of action. The people responsible for these tasks are usually concerned with any event that affects their bottom-line. Traditionally, human operators have had to either actively man a set of video displays to determine if specific events were occurring or manually review hours of collected video data to see if a specific event occurred. Actively monitoring video stream or manually reviewing and analyzing the data collected, however, is a tedious and long process which is prone to errors due to biases and inattention. Automatically processing and analyzing the video provides an alternate way of getting more accurate results because it can reduce the likelihood of missing important events and the human factors that lead to decreased efficiency.

Generating these statistics require an appropriate detection, tracking, and event representation framework that allows for identification of the targets of interest and determination of how long they spent in proximity of an object of interest. The challenge associated with this is keeping track of targets over a long period of time. Keeping track of targets over a long period of time is complicated due to occlusions and a constantly changing scene. Standard methods usually put together independent components that, while optimized for individual tasks, do not provide much feedback to each other.

The thesis aims to contribute to the area of using computer vision as a decision support tool by integrating detector, tracker, re-identification, activity status estimation, and event processor modules to generate the necessary event statistics needed by a human operator. The contribution of this thesis is a system that uses feedback

from each of the modules to provide better target detection, and tracking results for event statistics generation over an extended period of time.

To demonstrate the efficacy of the proposed system, it is first used to generate event statistics that measure productivity on multiple construction work sites. Results demonstrate that it's possible use the system to generate event statistics that show productivity on a construction site. Having these statistics available will be of great help to project managers when it comes to increasing their productivity and reducing cost. It will assist their decisions regarding the best way to utilize their resources.

The versatility of the proposed system is also demonstrated in an indoor assisted living environment by using it to determine how much of an influence a technology intervention had on promoting interactions amongst older adults in a shared space. The system is used to generate statistics regarding the usage of the shared space. Results show that the system reduces the time spent for analysis by a human operator to confirm or disprove their hypothesis regarding the intervention. This allows the operator to quickly make decisions regarding the changes necessary for the technological intervention to achieve the desired effect.

CHAPTER I

INTRODUCTION

Many tasks require some type of feedback in order to make decisions regarding the next course of action. Video surveillance can be employed as a tool to monitor events of interest for feedback regarding how implemented methods are working. Video has long been used in surveillance for applications such as activity recognition, event detection, human-computer interactions, productivity measurement, and safety, to name a few [2, 3, 4, 5, 6, 7, 8]. The number of cameras now available and the data collected makes manual review and analysis of the data an arduous and error-prone task. Traditionally, in surveillance settings, human operators have had to monitor the video streams and their effectiveness is mostly determined by how vigilant they are. These applications and the massive amount of data generated by continuously running video cameras drive the need to create intelligent visual surveillance systems. Having a manually operated surveillance system results in many events being miss-detected which can be caused by the excess number of videos to monitor, boredom and tiredness due to prolonged monitoring, lack of a-priori and readily accessible knowledge for what to look for, and distraction by additional responsibilities [9]. Automating the process reduces the amount of work operators have to do so it reduces the likelihood of missing important events. The main objective of an automatic video surveillance system is then to automatically detect, track, and analyze the activities of the objects of interest to generate necessary event statistics.

Automating the process, however, can be quite a difficult task, so this thesis aims to contribute to the area of automatic generation and analysis of event statistics by developing a system that integrates target detector, target tracker, target re-identification, activity status estimation, and event processor modules. Each module

is necessary for this system since they are all dependent on each other to provide accurate results. To perform event detection, the system must be able to detect and track targets over extended periods of time. The target detector module feeds its detection results to the tracker module which in turn feeds its results to the re-identification module, the activity estimation module and also back to the detector module for updating. The re-identification module uses the tracking results to determine when the target has changed significantly in order to re-initialize the tracker. The event processor module then combines the results from the different modules to generate the event statistics needed by the human operator. These statistics then enable the operator to make the necessary decisions for the task at hand.

1.1 Objectives and Contributions

The goal of this thesis is to create an analysis method which generates statistics that allows users to make decisions regarding their next course of action. The event features and statistics of interest necessary for decision support are the number of targets that entered the scene, the amount of time they spent in the scene, the amount of time they spent in a region of interest, and their proximity to the region of interest and other targets. Generating these statistics requires an appropriate target detection, target tracking, and event representation framework that allows for identification of the targets of interest and determination of how long they spent in the scene and the regions of interest.

The challenge associated with this is keeping track of targets over a long period of time. Long term visual tracking is a challenging task primarily due to target occlusion, illumination changes, scale changes, shadows, deformations, and target re-identification. Previous work attempts to address one or two of these drawbacks, but these solutions don't do it long enough to generate the event statistics needed for decision support. Video sequences used to demonstrate concepts in current literature

are typically less than 10,000 frames, and with a typical frame rate of 12 frames per second, this is approximately 14 minutes long [2].

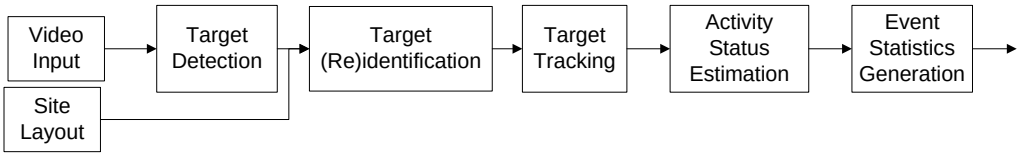
One reason why current systems face these drawbacks is that the flow of processing in most computer vision systems for surveillance is often hierarchical and unidirectional, as shown in Figure 1(a); most systems usually put together independent components that while highly optimized for individual tasks, do not provide much feedback to each other, and therefore do not work together in concert [3]. The goal of the proposed system is to address these problems by using feedback, as shown in Figure 1(b), from the different interconnected modules to induce robustness for estimation.

This goal will be achieved through the following contributions:

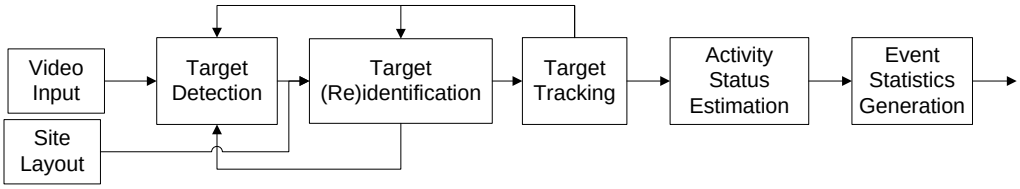
1. a generic system that uses feedback from each of the modules to provide better target detection and tracking results for event detection. The tracking module feeds back into the target detection module to update the background. The tracking module also provides feedback to the re-identification module in order to update the target's model. The re-identification module gives feedback to the tracking module regarding whether the target's model has sufficiently changed enough to warrant a new tracking model. It is also provides feedback to the target detection module on whether a detected target is a new or previously seen target.

2. generation of event statistics over a long period of time using identification of targets of interest, tracking of their movement within the scene, and their proximity to pre-determined objects. These objects could be other targets, or they could be static regions of the scene that have been provided by the user at run-time.

The input to the system will be the videos to be processed and a site layout which



(a) Traditional systems with no feedback



(b) Proposed system with feedback to different modules

Figure 1: Process flow for traditional and proposed automatic surveillance system.

defines the entrance and regions of interest. The entrances are used by the target detection module to create a tracking and re-identification model whenever it detects objects in that location, while the region of interest is used by the event detection module to generate statistics for the videos. The statistics of interest are identified by the user of the system when they provide the site layout.

The scope of the work done in this thesis is limited to application domains that require knowledge of the proximity of targets to a region of interest or other targets over an extended period of time for decision support. The use of the generic system is demonstrated in two application domains: construction sites and assisted living.

First, the efficacy of the proposed system is demonstrated on multiple construction sites. Automatically generating event statistics is extremely useful to project managers since productivity measurement and improvement still remains a difficult and expensive task for construction companies due to the need for manual data collection and analysis from the construction site. The scope of this thesis with regards to construction work sites are activities that support some kind of work which can

be detected or quantified by current surveillance algorithms. For now, this involves collaborative work between machines, such as excavators and dump trucks engaged in an earth-moving operation, since the activity states of these machines are more easily tracked.

Earthwork processes are often subject to unanticipated delays [10], which are likely to propagate through the entire remaining schedule and adversely impact progress, productivity, and costs [11]. The presented work impacts research into site operations by enabling the automated monitoring and tracking of on-site resources. Video-based monitoring and processing algorithms provide a non-intrusive, easy, inexpensive, and rapid mechanism for generating a body of operational information and knowledge. If made available to project managers, the information and knowledge would enable inquiry into construction operations that is currently not possible [12].

Analyzing the event features and statistics extracted by the system allows project managers to measure progress and productivity. The primary focus is analyzing the states of the Region of Interest (ROI), figuring out how long it takes in between dump trucks dumping and getting filled, and how many bucket loads it takes to fill the dump trucks by the excavator. Interactions between machines on the worksite are the most important factors to productivity since they determine the speed tasks on the worksite are performed. The ability to automatically analyze the event statistics reliably leads to less expensive means of productivity measurement and it also allows the project manager to efficiently manage resources which also reduces cost. The work done in this thesis differs from everything else in literature because longer video sequences covering majority of the construction site are used. To date, we are unaware of any vision-based solutions that provide productivity estimation of construction site operations through the measurement of progress and activity states over an extended period of time. The results demonstrate that it is possible to generate event statistics that show productivity associated to earthwork processes on construction sites.

Second, the proposed system is used to measure the effects of a technology intervention designed to promote interactions amongst older adults in a retirement community. Researchers have found that loneliness and social isolation is a cause of great discomfort among older adults in retirement communities. To combat this problem, a technology intervention was designed, and video cameras were set up to monitor the area surrounding the intervention to determine its effects. The long term goal is to achieve quantitative analysis of long-term responses to designed technological interventions by older adults in retirement communities. Specifically to demonstrate positive socialization impacts through these interventions.

The system is used to monitor the number of people present in the scene to detect if and how many of them go into the common area. The system tracks people's trajectories within the scene to understand how they localize themselves in the monitored space. The system also quantifies the interactions among older adults with and without the technology intervention. The amount of older adults who come out of their room and the amount of interactions that take place will show whether or not the technology intervention is having the desired effect. The scope of this thesis with regards to the senior housing surveillance is limited to interactions that can be determined using the proximity of targets to each other. The results demonstrates that the developed system can be used to quantify interaction statistics associated to interventions in a retirement community in less time than it takes for a human to manually annotate the same amount of time, thereby proving it as a valid decision support tool.

1.2 Literature Review

A literature review of the different modules in an automatic video surveillance system is given below.

1.2.1 Video Surveillance

Video surveillance in dynamic scenes, especially for humans, is currently a very active research topic in computer vision. It is preformed in retail outlets, traffic monitoring, banks, city centers, airports, building security. Typical video surveillance systems are monitored by human operators to determine the actions to take given a specific event of interest. Many studies, however, have shown the limits of human-based surveillance [9, 13]; they show that events are missed due to large amounts of data, bored and tiredness due to prolonged monitoring, lack of knowledge for what to look for, and distractions [4]. These studies reveal the need for an automated system that takes some of the burden off human operators while producing accurate results. An automated system would reduce the amount of work operators have to do and the likelihood of missing important events.

Collins et al. [14] developed a system for video surveillance and monitoring that uses a combination of temporal differencing and template tracking to accurately track a target in a video sequence. Multiple cameras were used to cooperatively track the object through the area of surveillance. Objects were detected and classified into semantic categories, which allow for temporal consistency contraints. Haritaoglu et al. [15] developed a system that employs a combination of shape analysis and tracking and construct models of people's appearances in order to detect and track groups of people as well as monitor their behaviors even in the presence of occlusion and in outdoor environments. Javed and Shah [16] developed an automated wide area surveillance system that detects, tracks, and classifies moving objects across multiple cameras. It uses single cameras to detect objects and then tracks using a voting scheme that utilizes color and shape cues to establish correspondence. The system combines tracking results from individual cameras with overlapping and/or non-overlapping field of views and it does not need calibration. The system can detect unusual activities happening in the surveilled area. Tian et. al [17] developed

a surveillance system that provides the capability to automatically monitor a scene, manage the surveillance data, perform event-based retrieval, receive real-time event alerts through the internet and extract statistical patterns of activity. The framework utilized in the aforementioned systems all share a similar structure. The structure is broken into the following stages: background modeling, target detection and identification, tracking, and understanding and description of behaviors [5, 6, 7, 8, 13, 14, 15]. As mentioned earlier, the framework is usually hierarchical and unidirectional. The following subsections give a review of the components that make up the system.

1.2.2 Background Modeling

Background subtraction is an important aspect of any computer vision system since it allows for detection of foreground objects of interest. To perform background subtraction, a background model is needed. The foreground objects in an incoming frame are detected by finding regions that deviate from the background model. Background models can be divided into the following categories: mixture of Gaussians (MOG) models, subspace learning models, fuzzy models, and robust PCA models [18].

The MOG model introduced by Stauffer and Grimson [19] is the most common approach used for background subtraction. For this model, each pixel's history is modeled using a mixture of K Gaussian distributions; the probability of belonging to the background is calculated and used to determine the foreground. The model's mean, variance, and weights are then updated using the detection results. Several researchers have modified this algorithm to make it more robust to situations such as: noisy images, camera jitter, time of day, bootstrapping, camouflage, foreground aperture, etc [18, 20, 21].

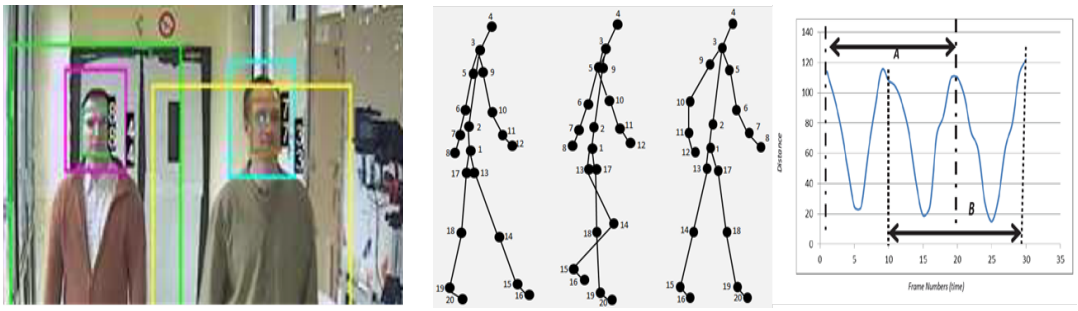
Subspace learning methods create a model using dimension reduction and foreground is detected by comparing the input image and its reconstruction [22]. The first subspace learning method using Principal Component Analysis (PCA) was proposed

by Oliver at al. [23]. Subspace learning is applied on N images to construction a background model, which is represented by the mean image and the projection matrix comprising the first p significant eigenvectors of PCA. Foreground detection is performed by thresholding the difference between the input image and the reconstructed background image (using the input image, mean image, and projection matrix).

Fuzzy models and robust PCA models are more recent background models which are more robust to challenges faced in dynamic scenes but they are too sophisticated and complex for the needs of this proposal. The MOG model is used in this proposal because it provides enough accuracy with fast speed for the data being analyzed. A more comprehensive review on these models is given in the survey by Bouwmans [18].

1.2.3 Target Re-identification

Once the targets have been detected, they need to be identified for re-identification and tracking purposes. Target re-identification is important in video surveillance because it allows for accurate computation of statistics of events in the video. Re-identification algorithms need to be robust to challenging situations such as changes in camera viewpoints and orientations, illumination changes, pose changes, and rapid change in clothes appearance. Existing work can be classified into biometrics based methods and appearance based methods. Biometric methods such as face [24, 25] and gait [26, 27, 28], have been used for re-identification but these method requires features that can only be extract in high resolution images. Appearance based methods are commonly used for person re-identification. Appearance based methods usually focus on the selection of discriminative features and the combination of various features to create invariant signatures [29, 30]. Yang's approach [30] is improved upon in this proposal since it has the same KPCA framework as the tracker used thereby saving computation time.



(a) Face [25] (b) Gait [28]

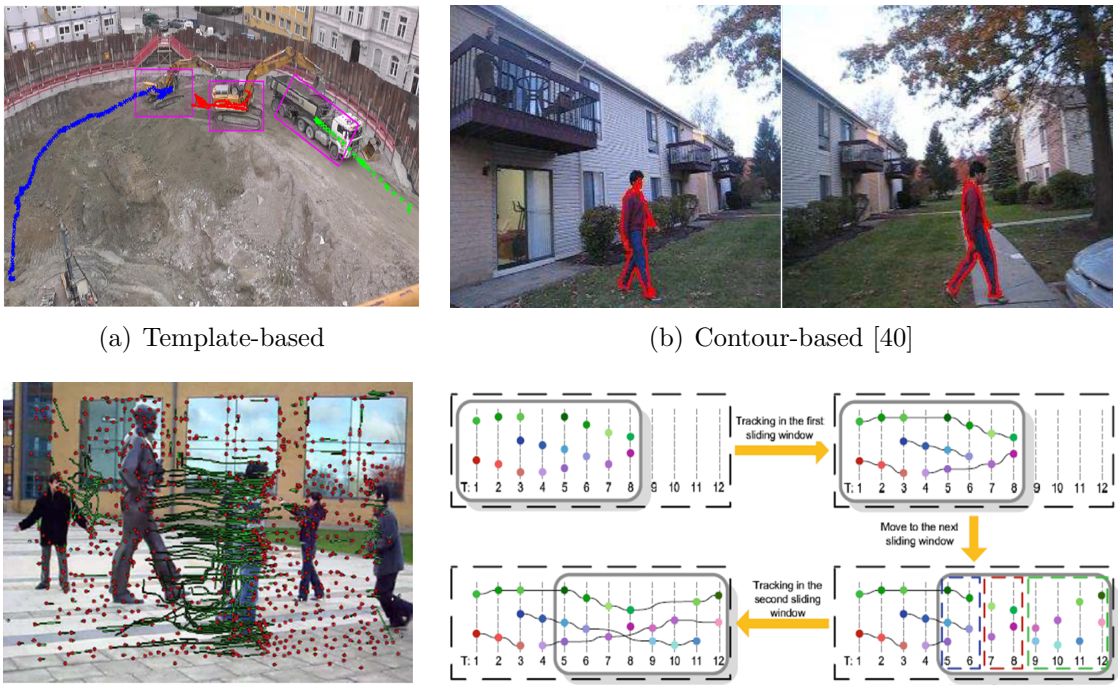
Figure 2: Example of biometric re-identification methods

Figure 3: Example of appearance based re-identification [1]

1.2.4 Target Tracking

Tracking algorithms can be split into four different categories: kernel based, feature based, contour based, and association based [31, 32]. Kernel-based tracking typically tracks using a target region stored as a template. It finds the target's new region by maximizing a similarity measure. Similarity measures such as the Bhattacharya coefficient [33], sum of squared differences [34], normalized cross correlation [35], and robust maximum mean discrepancy [36] have been used. Based on the motion of the target, a parametric motion model such as translation, affine, coformal affine, projective model is chosen. The simplest kernel based method is template matching. Template matching uses gradient descent to find the location of the target in a frame. Cheng [37] introduced mean shift for tracking and since then it has been improved by multiple researchers who incorporate different similarity measures [33, 38, 39].

Feature-based tracking algorithms extract the target's features and matches the



(a) Template-based (b) Contour-based [40]

(c) Feature-based [41] (d) Association-based [42]

Figure 4: Examples of different types of tracking.

features in subsequent frames . The displacement of the target is then estimated using the collective motion of the features. Edges [43, 44, 45], texture [46],corners [44], scale-invariant [47, 48], and affine-invariant [49] features are examples of features that have been used for tracking. Optical flow, Kalman filters, Extended Kalman filters, Unscented Kalman filters, Hidden Markov Models, boostrap filters, and particle filters are examples of feature trackers. Feature-based trackers depend on detecting features in subsequent frames so they fail when feature extraction become impossible due to the features missing.

Contour-based trackers are based on the temporal deformation of a contour according to the variations of an energy functional. Contour-based trackers fail whenever the target is similar to another target or the background. Contour based tracking can be done using an explicit representation or an implicit representation of the contour. It is done by iteratively evolving an initial contour. Most recent approaches

use an implicit representation of contours using level sets and iteratively evolve the contour by directly minimizing the contour energy function. The contour energy function can be motion-based [50] or segmentation-based [40]. The energy is defined with respect to the evolving contour and its minimum is attained when segmentation is achieved. The contour is commonly embedded as the zero level set of a higher-dimensional function, e.g. a signed distance function, and iteratively deformed until the energy reaches its minimum.

Association based trackers operate by finding the proper linking affinities based on multiple cues between object detection responses or tracklets [32, 42, 51, 52]. Association based trackers depend heavily on the object detection algorithm so they provide inaccurate results when the detection is not reliable. Although kernel-based trackers tend to fail whenever the target changes its orientation or pose, a kernel-based tracker is used in this proposal because of the potential to use multiple templates to represent the target. The last module of an automatic surveillance system is the understanding and description of behaviors and activities, the following two sections give a literature review on the behaviors and activities of interest in the two application fields.

1.2.5 Construction Site Work Sampling

Construction sites are locations where the building or assembling of infrastructure takes place, usually entailing a high usage of resources and manpower. Due to the high complexity of large scale construction projects, project schedules are sensitive to delays. Since these projects are usually an expensive undertaking, managers are always looking for ways to measure and improve labor productivity. Proper coordination of resources and manpower positively impacts on-site productivity, which in turn influences construction safety, costs, and schedule [53, 54]. To measure productivity at such construction work sites, one needs to attribute a quantitative measure to the amount of work performed. Awareness of labor productivity leads to improvements

on direct work rate [55] therefore, the existence of productivity analysis methods improves onsite operations and mitigates any adverse conditions that may impede progress. The ability to provide feedback on the duration of activities on the work site would allow project managers to be more aware of delays, determine their causes, and respond accordingly. Additionally, a project manager would have a better idea of efficiently managing resources if provided with metrics like the idle time of each machine, and time spent performing direct and support work. The construction company with the most efficient operations has a great chance of making more money and delivering a faster construction project to the project owner [56].

According to Gong and Caldas [57], traditional data collection methods for productivity analysis can generally be divided into four categories: 1) cost report and project schedule, 2) work measurement based on direct observation, 3) survey or interview based methods, 4) video-review based methods. The only common thread between these categories is that they require labor intensive, time-consuming, and expensive manual analysis. Currently work sampling, which falls under work measurement based on direct observation, is the method most used by construction companies to determine productivity. Work sampling is a method that evaluates the amount of time workers spend on direct work, support work, and no work. Traditionally, this is done by an observer following a pre-planned route and recording the activity, called a work sample, the instant they come in contact with whoever is performing the activity [56]. An additional drawback of this method is that the samples are acquired discretely and the observer might miss some activities as they moves about the pre-planned route. Due to the drawbacks mentioned above, researchers have been researching ways of automatically performing productivity analysis by monitoring the resources and manpower used on the work site.

Researchers have attempted to use Global Positioning Systems (GPS) [58, 59, 60, 61, 62], Radio Frequency Identification (RFID) [63, 64, 65], and Ultra Wideband

technology (UWB) [66, 67, 68] to perform productivity analysis and while they are capable of tracking the location of resources, there are several drawbacks. They all require the resources be manually tagged which can be quite intrusive, and problematic when there is a significant amount of resources to be tracked. GPS systems are not accurate indoors and RFID only operate within a fixed radius. These methods are also not able to provide information regarding the type of activities the resources are engaged in. Video surveillance is a natural approach used to address this issue since it provides a continuous data feed for the construction site. However, manually reviewing the video to determine how resources on the construction site are being utilized is an arduous and expensive task. Computer vision can be used to overcome these shortfalls by automating the process of performing work sampling.

Using computer vision as a tool for work site analysis is a fairly recent research area. Many current research efforts associated to progress and productivity monitoring seek to prove the hypothesis that it is possible to reliably track multiple resources with images (video and/or time-lapse) in order to reproduce the daily workflow activities associated to a construction site. The intent behind such monitoring and analysis is to automatically provide critical information, through computer-vision algorithms, on construction operations for improved decision making in construction engineering and management [69]. The information obtained from such automated systems generates knowledge about work site operations. In an information-based framework, much effort is spent acquiring and interpreting information. In a knowledge-based framework, efforts are allocated to making decisions based on the interpreted information. If successful, computer-vision based methods will transform the review of construction operations from being information-based to knowledge-based, thus saving human resources and improving decision effectiveness [70].

Using passive imaging cameras aimed at a worksite for performing resource tracking and activity monitoring relies on applying methods and tools from surveillance

research [14]. As detailed in Section 2, video surveillance systems require the connection of several modules, which perform object detection, identification, tracking, and reidentification (needed when an object leaves and returns the scene). Once these basic components are functional, an additional interpretation module may be added in order to identify behaviors or activities engaged by the tracked objects within the sensed scene [15]. These may be further decomposed into typical or unusual [16]. Further analysis of the behaviors and activities over time may be performed in order to identify key events, which can then be entered into a database or spreadsheet for reporting or query purposes. For example, the system in [17] performs event-based retrieval, provides real-time event alerts through the internet, and extracts long term statistical patterns of activity.

Detection. Resource detection on construction site videos is broadly categorizable into two approaches: specific object detection algorithms and general foreground estimation algorithms. Detection algorithms, usually relying on machine learning techniques, involve training to learn the unique signature of a given object. Algorithms include neural networks [71], support vector machines with specific feature models [72], random forests [73], and parts-based models [74]. Parts-based modeling approaches work best for articulated objects since their appearance geometry has high variation, which can be compensated through multiple, individual part detectors. Rigid targets with sufficiently discriminative appearance characteristics can be detected in real-time, however more complex targets require processing that prevents real-time processing. Since detection-based methods often seek specific targets, the resource type will typically be known from the detection itself. Foreground estimation algorithms tend to be simpler, as they generate a model of the expected scene and classify as continuous regions that do not match as target regions [75, 57]. Foreground estimation works well when the entrance into the scene of the object is controlled, or the object of interest is not occluded by other foreground objects [76].

In less controlled settings, strong results are obtained by combining the two techniques [77, 75, 71]. Since foreground detection methods do not classify the detected objects, the addition of a detection algorithm for these scenarios has the added benefit of rejecting irrelevent foreground detections and of identifying the resource category (if there is a detector for it).

Tracking. There have been comparative papers studying the performance of specific tracking algorithms on construction worksites [78, 70] with probabilistic and kernel-based methods showing strong performance. Kernel-based methods have been used since for tracking on-site resources [79]. [80] used a probabilistic kernel method to track multiple workers with overlapping trajectories, showing that these methods can be modified to handle occlusions. Follow-up research [81] extended these results to rigid construction machines observed from a distance. While most tracking papers are on tracking with a single camera view, [82] demonstrated 3D tracking of construction site resources using stereo video.

Activity Analysis. For a detected and tracked construction resource, further analysis of the object's visual stream provides important information regarding the role and contribution of the resource to the construction process. Deciphering this information falls within the category of activity analysis. Activity analysis on a construction site involves determining the action each target in the scene is engaged in over a period of time. Early activity analysis utilized sensors installed on the resource of interest [83]. However, since then vision-based strategies have been demonstrated by applying advances in computer vision. The vision-based research literature, both in construction and more generally, can be split into activity identification through analysis of specific spatio-temporal visual features of the resource or through analysis of the trajectory and proximity information.

The former category has mostly focused on articulated resources, such as personnel and machines [84, 85]. By decoding the articulation poses or target feature elements

over time, the activity category can be inferred [86, 87]. Work activities may be broken into effective work, ineffective work, or contributory work for productivity analysis [84].

For rigid objects, or those without discernable pose properties, alternative means are needed, leading to the latter category. The addition of a-priori information about the targets and their work packages, the location of regions of interest plus their meaning, and the trajectories of each target enables the decoding of activities through Markov models based on work process diagrams [57, 76]. Whole site analysis is possible for earthworking processes [81] since the quantity and types of machines are somewhat limited during this part of the construction phase, and the activities are inferrable from the interaction dynamics of the machines.

Productivity Estimation. Analysis of activity state estimates over time, when connected to specific work packages, provides productivity data for the work packages [77]. Over short time intervals, with specific work packages, productivity can be inferred through the activity states coupled with some minimal information regarding the task [88]. For longer time intervals, however, it is more useful to connect activity statistics to actual progress, which requires progress tracking. To date, we are unaware of any vision-based solutions that provide productivity estimation of construction site operations through the measurement of progress and activity states over time.

1.2.6 Senior Housing Surveillance

The older adult population (65 and older) of the United States has been dramatically increasing and is expected to increase further in the next decade [89]. According to estimations, older adults will represent 20% of the total population by 2030, and by 2050, the number of older adults is expected to outnumber children 14 and under for the first time in history. It is estimated that by 2050, 50% of older adults who require

care will not have children [90]. To date, 2.3% of older adults are housed in independent living retirement facilities [91]. These communities are an affordable option for older adults to become part of a community. However, even though these communities provide social exposure, depression and isolation are present [92]. Depression is strongly linked to morbidity for the older adult population [93], while social isolation is connected to decreased quality of life, depression, and morbidity [94, 95, 96, 97]. Thus, for the communities to be places where older adults can thrive, mechanisms must exist within the facility that successfully promote social interaction, increase the well-being of the population, and reduce the incidence of depression. Environmental [98], activities-based [99, 100], technological [101], and social robotics [102, 103] interventions are all seen and being investigated as mechanisms that can achieve these objectives. At the same time, contemporary research in these areas does not involve the population base, time duration, and quantitative results [102, 104, 105] found in more traditional areas of gerontological research, e.g. compared to [95, 97]. With the above observations and ideas in mind, this thesis covers the first steps towards a more comprehensive, quantitative, and long-term assessment of technology-based interventions in an independent living community. In particular, it presents a semi-automated surveillance system used to codify and quantify the social interactions found within public spaces in an independent living community. The semi-automated system aims to provide a summary of the events that take place within the public spaces to allow researchers/caregivers to determine how much of an effect their interventions are having, which in turn allows them to make decisions regarding the next steps.

Socialization and Older adults. Loneliness has been found to be a cause of great discomfort among retirement community residents. Social interaction and social support impact both quality of life and health [94, 106, 107], and lack of either leads to higher mortality rates amongst lonely older adults [108]. Shared common areas in retirement communities should provide spaces for interaction, yet they are

underutilized from the perspective of social interaction. Based on an observational study of residents in assisted living facilities Zimmerman et al [109] found that about half of awake residents make use of public spaces. Furthermore, the study noted that residents who are cognitively and functionally impaired are more likely to be in public spaces yet less likely to be engaged, while residents who are awake and alone in private spaces are less likely to be impaired yet more likely to have and develop medical conditions. Observational remarks by researchers in social robotics have also noted that while public spaces may be occupied, the occupants are not necessarily socially engaged [103, 110]. There is a disconnect between occupancy of shared spaces and socialization in shared spaces. If social interactions can increase an older adult's quality of life and their health as well, then means to encourage socialization are imperative, especially in shared common areas. Caregivers are constantly looking for ways to encourage and improve social interactions amongst their residents. They need to be able to evaluate the effects of of their chosen interventions on socialization.

Behavior Mapping in Environments. Evaluating utilization and socialization in and nearby common areas of retirement communities will require observation of the shared environment and the interactions of the population that utilize it. As mentioned in Chapter 1, human operators have traditionally had to actively man a set of monitors, or review video after the fact [103, 105, 110], to determine if specific events were occurring or to analyze subject behavior. To facilitate the collection of statistics, software such as Observer XT by Noldus provides professional and user-friendly event logging system for the collection, analysis, and presentation of observational data [111]. This system codes and describes behavior in an accurate and quantitative way, but the process is lengthy due to the need for manually annotating the observations. Manual approaches are not feasible for long term, all day surveillance. As such there is a need to develop system that can automatically detect human behavior and interaction in environments. Semi-automated video processing algorithms are

essential for rapidly providing summaries of daily interactions or activities occurring within the sensed space [112].

The proposed semi-automated behavioral mapping surveillance and interaction-processing system will serve as a tool to monitor the social interactions affected by the technology interventions. The goal of the system is to identify how people utilize the space, understand if and how technologies designed for, and placed in, the space promote socialization. A social interaction is a mutual or reciprocal action that involves two or more people and produces various characteristic visual/audio patterns.

Current Behavioral Mapping Technologies. Due to the challenges of video-based technology, research on the use of sensor systems to determine interactions within a group may involve active or tag-based sensors. These sensors are integrated into the environment to provide activity recognition [113, 114, 115]; such setups are sometimes called smart homes. As sample case, Lymberopoulos, et al. [116] used cameras, door sensors, and passive infrared sensors to create a spatiotemporal human activity model for activity detection. These approaches require extensive retrofitting of environments and are not feasible for existing independent living communities. Within the context of wearable sensors, the use of wrist-worn devices with on-board sensing (acceleration, voice, etc.) and local wireless communication enables the quantification of activities or social interaction [117, 118]. Potential issues with tag-based approaches, such as RFID, for daily surveillance of a building's residents are their intrusiveness (they are always on), potential economic cost, and/or lack of complete anonymity. Connecting a person to their tag automatically provides knowledge of their entire history, seamlessly removing anonymity.

Video-based methods provide a means to record the interactions, but require additional processing in order to provide activity and interaction statistics. Wu et al. [119] and Park et al. [120] use video data to perform activity recognition, with RFIDs tags being used as a supervised training strategy. As a processing method, Wu et

al. [119] used Dynamic Bayesian Networks (DBNs) to determine the most likely activity and object labels in their work. Wu et al. [121] use multiple cameras to collect spatial-temporal data and perform activity recognition. These works focus on individual activities rather than interaction between individuals. To detect social interaction, audio is an essential sensor modality. Chen et al. [122] evaluated various machine-learning algorithms with fused video and audio data for detecting social interactions; the algorithms evaluated were decision trees, naive Bayes classifiers, naive Bayesian networks, Adaboost, and logistic region. Hauptmann et al. [123] also use video and audio data for activity recognition. They use the mean shift tracker and support vector machines (SVM) to train the system to recognize activities. Machine learning tools typically require extensive manual annotation to work well, yet it can be the case that the supervised training does not provide sufficient detection accuracy to be considered reliable. Another class of research using video combines active range imaging sensors with passive visual imaging sensors for activity awareness [124, 125]. The fused color+depth images, called RGB-D, have been used to count people and to re-identify people [126, 127]. A current limitation of depth sensors is that most cost-effective indoor depth sensors are range limited (Microsoft Kinect's maximum range is about 15 feet).

Given the characteristics of available sensor technology and the desire to cost effectively install surveillance setups on multiple floors, and possibly across multiple building, passive video and audio sensors appear to be the best choice. Video only was chosen over RGB-D sensors due to the distances involved. For interaction detection, the work done in this thesis takes a simplified approach to detecting social interactions amongst individuals, which relies on geometric and mathematical descriptions of basic activities that can be tested using the tracked trajectories of the sensed participants. The video data is processed using the surveillance system described in Section 2. The last step of the video processing pipeline is an interactive post-processing correction

step, used to correct any improper trajectory correspondences when people leave and re-enter the scene.

The remainder of this thesis is organized as follows. Chapter 2 gives more details about the algorithms used in this work. Chapter 3 describes the work performed on construction work sampling and the results achieved. Chapter 4 describes the work performed on detecting interaction induced by a technology intervention. Chapter 5 concludes and discusses future work.

CHAPTER II

SYSTEM OVERVIEW

This section gives an overview of the proposed system. The problem setup involves a single monocular camera configured to view a scene where activities of interest could occur. These activities involve interactions between different targets or within a region of interest (ROI). The proposed automated system processes the video given a-priori information about the layout of the scene, and provides an automated report of the activity states of the targets and the state of each region of interest.

As shown in Figure 5, the first step in analyzing the video is modeling the background so that foreground detection can be performed to detect the targets of interest. Once a target has been detected, its appearance model is learned for re-identification and tracking. The tracking result for each frame is used to perform foreground segmentation for updating the background model, detecting new targets, and re-learning the target's appearance model. Once the whole video sequence is tracked, the results are passed to an event detection processor that outputs the event statistics such as the average time spent in the region of interest, the number of targets that entered the region of interest, the number of targets that entered the scene, how long they spent in each region of interest, each target's deviation from the average time spent in the region of interest, pie charts showing how targets spent their time, state plots for each individual target, and state plots for the region of interests. Below is a description of the different components of the proposed system used in the preliminary research to determine the feasibility of the approach.

2.1 Detector

The first step in any automated visual tracking system is detecting when new foreground objects appear. Once the foreground object has been detected, its appearance

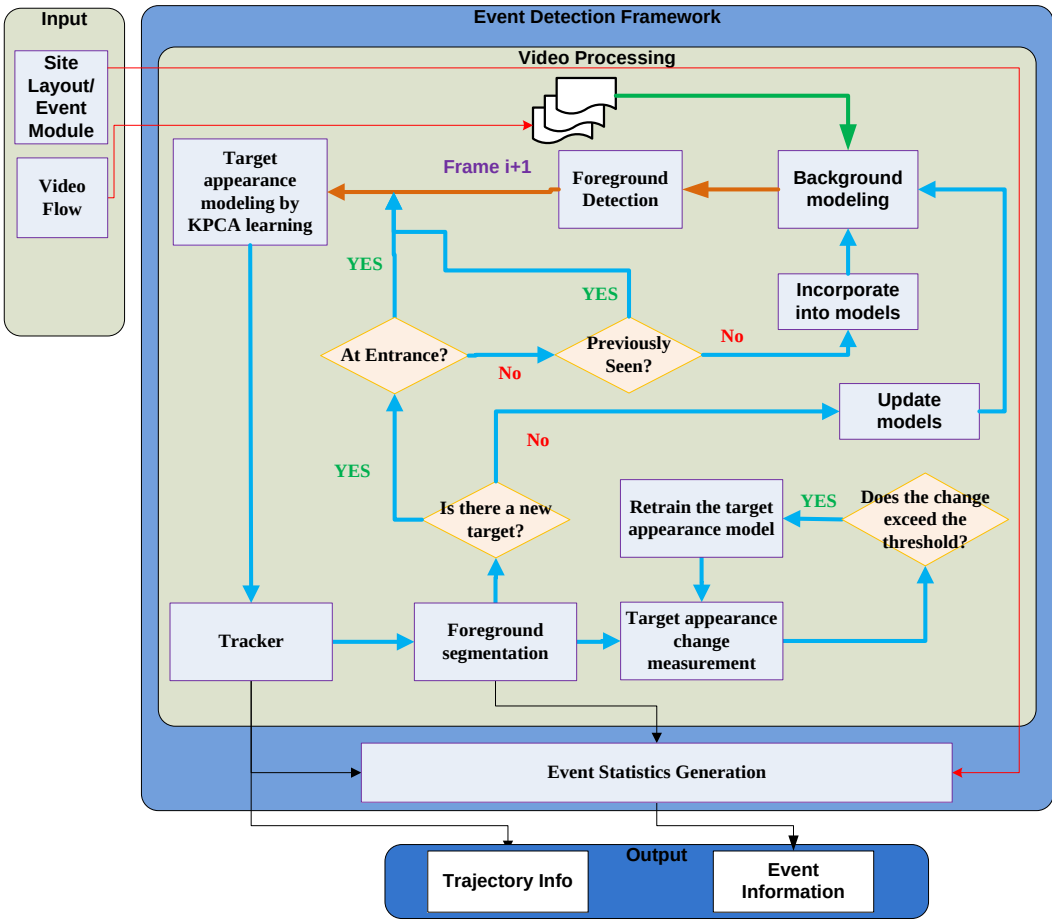


Figure 5: Outline of the framework

template is used to initialize a tracker. A foreground detector using Gaussian Mixture Models (GMM) is utilized for the work done in this thesis. GMMs store multimodal representations of background so that more complex dynamic scenes can be handled. To perform foreground detection using GMMs, each pixel is modeled separately by a mixture of k Gaussians [19]. Each k Gaussian distribution describes a background pixel. The probability of observing a pixel value, x, belonging to the background at time t is given by:

$$P(x_t) = \sum_{i=1}^{k} \omega_{i,t} \eta(x_t; \mu_{i,t}, \Sigma_{i,t}) \tag{1}$$

where k is the number of distributions, $\omega_{i,t}$ is the estimated weight of the i^{th} Gaussian in the mixture at time t, $\mu_{i,t}$ is the mean value of the i^{th} Gaussian in the mixture at time t, $\Sigma_{i,t}$ is the covariance matrix of the i^{th} Gaussian in the mixture at time t, it is assumed to be $\Sigma_{i,t} = \sigma_k^2\mathbf{I}$ and η is a Gaussian probability density function. If the probability of a pixel belonging to the background is less than the given threshold, it is considered foreground.

The prior weights of the k distributions at time t, $\omega_{k,t}$ are updated as follows:

$$\omega_{k,t} = (1-\alpha)\omega_{k,t-1} + \alpha(M_{k,t}) \tag{2}$$

where α is the learning rate, and $M_{k,t}$ is 1 for the matched model, and 0 for the remaining models.

The mean and variance of the matched distribution are updated as follows:

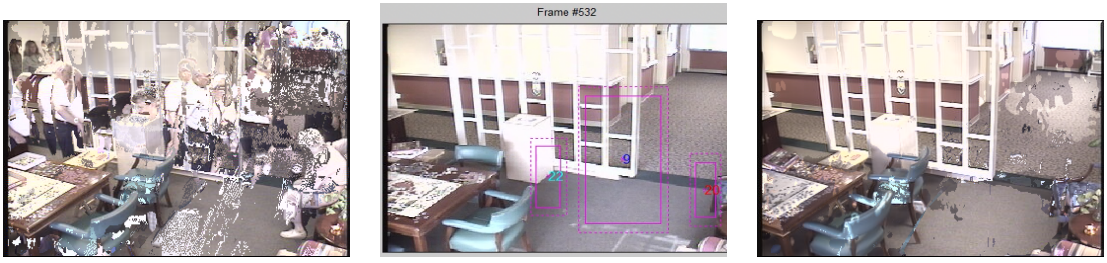
$$\mu_{i,t} = (1-\rho)\mu_{i,t-1} + \rho x_t \tag{3}$$

$$\sigma_{i,t}^2 = (1-\rho)\sigma_{t-1}^2 + \rho(x_t - \mu_t)^T(x_t - \mu_t) \tag{4}$$

where $\rho = \alpha\eta(x_t|\mu_k, \sigma_k)$.

The mean, variance and weights are only updated for the pixels not classified as foreground by the foreground segmentation step of the tracker. This allows the detector to incorporate anything that isn't tracked from the entrance region into the background model while assuring the tracked objects will never be included in the background model no matter how long they spend in the scene. Figure 6 shows the effects of using feedback to update the background over a long period of time. Without feedback, the background model becomes too noisy, leading to extraneous tracks, which leads to inaccurate statistics.

A three stage background estimation technique [128] is utilized at three different intervals whatever video is being processed to create the initial k distributions for the



(a) No feedback (b) Extraneous tracks (c) With feedback

Figure 6: Effects of feedback on background model

background model. For the first stage, each image from the learning, each image from the evaluated interval is divided into blocks, with each block location in the image being compared to all other image blocks in the same location. The set of all blocks at a location (i, j) defines the representative set $\mathbf{R}(i, j)$. The set $\mathbf{R}(i, j)$ contains only unique representative blocks $\mathbf{r}_k(i, j)$, where k is the image frame number. A block $\mathbf{r}_k(i, j)$ is kept after it has appeared for at least two consecutive frames. The uniqueness of a block is determined using the correlation coefficient and the mean of absolute differences with other blocks in the location. One element of $\mathbf{R}(i, j)$ is assumed to be the background block.

After generating the set $\mathbf{R}(i, j)$, the second stage initializes the background model by filling it in with the blocks from representative sets having just one block at that location. In the third stage uninitialized background blocks $\mathbf{r}(i, j)$ are filled with a representative block from the representative set $\mathbf{R}(i, j)$. The uniqueness of a block is determined by evaluating it correlation coefficient and mean absolute difference with another block. An empty background block will only be estimated if the background is available in at least 2 neighboring blocks of its 4-connected neighbors, which are adjacent to each other and also in the diagonal block located between them. The best candidate for the background is acquired by comparing the discrete cosine transforms of its superblock (block appended to its neighboring blocks). The candidate with a smoother distribution is selected as the best fit. The estimated background at the

specified intervals serves to initialize Gaussian mixture model (GMM).

To deal with shadows that might occur during the tracking process, a shadow removal method [129] that uses color constancy between pixel, color constancy within pixels, and temporal consistency between frames is utilized. The color constancy between pixel assumes that the intensity ratios between neighboring shadow pixels in the current image should be the same as those in the background image. The ratio map is computed using equation 5, and the error score, equation 6, for discriminating the pixel as shadow can be calculated by summing the absolute difference between d and d' over all pixels in a small neighborhood window ω centered at (x, y). The overall error score, equation 7, is the sum of the error score for each channel.

$$\begin{cases} d(x,y) = \ln \frac{I(x,y)}{I(x+1,y)} = \ln I(x,y) - \ln I(x+1,y) \\ d'(x,y) = \ln \frac{I'(x,y)}{I'(x+1,y)} = \ln I'(x,y) - \ln I'(x+1,y) \end{cases} \tag{5}$$

$$D(x,y) = \sum_{(i,j)\in\omega(x,y)} |d(i,j) - d'(i,j)| \tag{6}$$

$$\Psi(x,y) = \sum_{i\in R,G,B} D_i(x,y) \tag{7}$$

The color constancy within pixel assumes that the color of a pixel stays roughly the same when a shadow is casted on it. The brightness and color information of a pixel is acquired by transferring the color space from RGB to the normalized r-g using the following equations

$$\begin{cases} C_r(x,y) = \ln \frac{I_R(x,y)}{I_R(x,y)+I_G(x,y)+I_B(x,y)} \\ C_g(x,y) = \ln \frac{I_G(x,y)}{I_R(x,y)+I_G(x,y)+I_B(x,y)} \end{cases} \tag{8}$$

The error score for discriminating the pixel (x, y) as shadow is defined as

$$\Lambda(x,y) = |C_r(x,y) - C'_r(x,y)| + |C_g(x,y) - C'_g(x,y)| \tag{9}$$

where C contains the color information of the current image and C' contains the color information of the background region. If $\Lambda(x, y)$ is small, it means the color of the pixel does not change much, and it is more likely to be a shadow pixel.

Foreground objects with uniform color and its shadow on a uniform background cannot be distinguished using just color constancy between pixels. Foreground regions with color similar to its background region would be wrongly classified as shadow regions using just the color constancy within pixels. Assuming that shadow pixels tend to remain in a shadow region in the next frame, and also assuming that the foreground object moves slowly, temoporal consistency between frames can be used to get a clue for potential shadow regions. The error scores are fused together using the following recursive linear equation

$$\Theta_t(x, y) = \alpha \cdot \frac{\sigma^2_\Lambda(x, y) + \sigma^2_\Psi(x, y) \cdot \Lambda_t(x, y)}{\sigma^2_\Psi(x, y) + \sigma^2_\Lambda(x, y)} + (1 - \alpha) \cdot \Theta_t(x, y) \tag{10}$$

where α is a constant that controls the speed of the recursive update. $\sigma^2_\Psi(x, y)$ and $\sigma^2_\Lambda(x, y)$ are the temporal variances and are inversely proportional to the weights that control the importance of each factor. They are updated recursively in a manner similar to equation 3. $\Theta_t(x, y)$ is the total error score for discriminating the pixel as a shadow pixel. It is thresholded to determine whether a pixel belongs to a foreground object or to a cast shadow region. Figure 7 shows the effectiveness of the shadow removal technique.

2.2 Tracker

An improvement on the kernel covariance tracker proposed by Arif and Vela [36] is utilized for tracking detected targets. In the original kernel covariance tracker, the target's feature vector is a joint color-spatial vector, $u_i = [\mathcal{I}(x_i), x_i]^T$, where $\mathcal{I}(x_i)$ is the color data at location x_i. The target's model is learned by mapping its feature vector into the a higher dimensional feature space $\mathcal{H}$ using the Gaussian kernel, $\boldsymbol{k}(u_i, u_j) = \exp(\frac{-1}{2}(u_i - u_j)^T \Sigma^{-1}(u_i - u_j))$. The eigenvectors, $\alpha^k = [\alpha^k_i, \ldots, \alpha^k_N]$, and

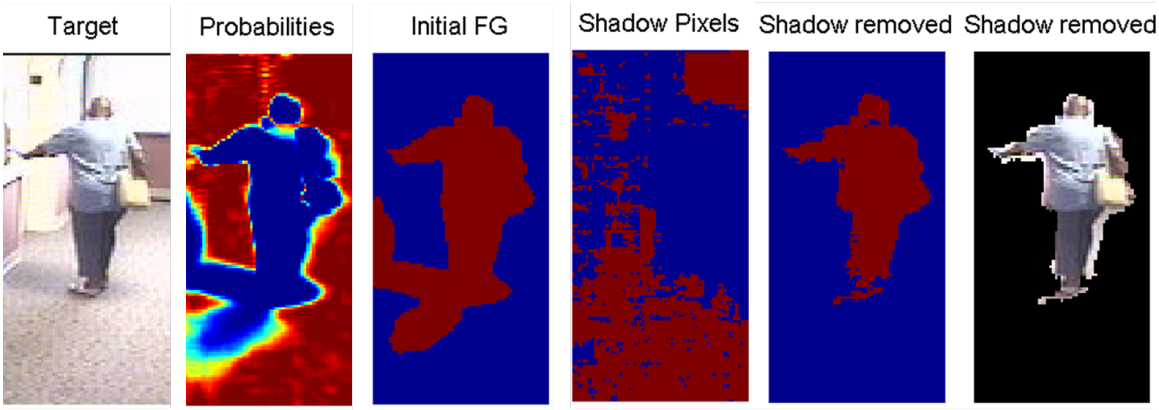


Figure 7: Sample shadow removal results

eigenvalues, λ^k, of the kernel matrix are computed and the eigenvectors are normalized by the eigenvalues. All the mapped points are then projected onto the normalized eigenvectors (11).

$$f^k(u_i) = \sum_{j=1}^{n} \frac{\alpha_i^k}{\sqrt{\lambda^k}} \boldsymbol{k}(u_i, u_j) \tag{11}$$

The target is then tracked by finding the region $\mathcal{R}$ that best matches the target. The similarity score of a region to the target is given by,

$$\mathcal{SC}(\mathcal{R}) = \sum_{i=1}^{n} \sum_{k=1}^{m} (f^k(u_i))^2, \tag{12}$$

where n is the number of feature vectors in the target's template and m is the number of eigenvectors retained. For every frame and each target, a gradient ascent procedure localizes the target by comparing the foreground image data with the targets' learned model in order to optimize the region similarity.

The process of computing the map from the feature vector space to the higher-dimensional feature space $\mathcal{H}$ utilized by the kernel covariance tracker is known as *kernel PCA* [130]. This method is the prototypical example of a powerful class of algorithms known as kernel methods: these methods are powerful, but they suffer from significant challenges when dealing with large amounts of data. In particular, kernel PCA, as seen above, requires the computation of an $n \times n$ kernel matrix:

diagonalizing this matrix (i.e. computing its eigendecomposition for the purposes of computing the eigenfunctions of the integral operator associated to the kernel for the nonlinear feature mapping) requires $\mathcal{O}(n^3)$ operations. Furthermore, projecting new data onto the approximate feature space requires $\mathcal{O}(nm)$ operations, where m is the number of retained eigenfunctions. There exist many methods for speeding up both the training and projections times separately but the technique known as shadow densities speeds both up at the same time [131, 132].

Shadow densities connect kernel PCA to the eigendecomposition of kernel smoothing operators. In particular, given a sampled data set $\{x_i\}_{i=1}^n$, the spectral decomposition of the kernel matrix K can be related to the kernel density estimate $\hat{p}(x)$ of the underlying probability density $p(x)$ generating the data. Shadow densities generate an estimate $\tilde{p}(x)$ of the kernel density estimate $\hat{p}(x)$ that has cardinality $r \ll n$, and then use the lower-cardinality estimate to construct an approximation to the feature space using only r points, resulting in $\mathcal{O}(r^3)$ training complexity, and only $\mathcal{O}(rm)$ computations for projection. Furthermore, unlike methods such as the Nyström method [133] and random Fourier features [134], the cardinality r is generated naturally by a user-provided parameter ℓ that controls the approximation to the kernel in an intuitive manner (large values of ℓ ensure better approximation at the cost of more basis functions). For the tracking purposes in this thesis, $\ell = 3$, which is guaranteed to produce good results both theoretically and empirically, is utilized.

2.3 Target Re-Identification

Target re-identification is performed using KPCA in a fashion similar to the approach of Jun et al. [30]. Once the target is detected, key templates for re-identification are acquired as the target is being tracked. Key templates are chosen to represent the target's change in appearance and pose due to movement, illumination changes, and scale changes. Key templates are selected based on how distinctively they represent

the target. The first key template of a new target is initialized when it first enters the scene. They are selected based on the combination modified Haussdorf distance and the Bhattacharrya coefficient between the tracked target's foreground and other key templates.

Once the target is detected, key templates for re-identification are acquired as the target is being tracked. Key templates are chosen to represent the target's change in appearance and pose due to movement, illumination changes, and scale changes. Key templates are selected based on how distinctively they represent the target. The first key template of a new target is initialized when it first enters the scene.

A two phase process is used to determine a key frame. The first phase uses the velocity to determine how well the target's shape fits the tracking model. Velocity is used to determine the direction the target is moving. This is important since we are using a template tracker, a change in orientation means the template no longer fits the target. A significant change in direction means the tracker will most likely lose the target. As a target moves in a different direction, its orientation changes, which signifies the need for a new key frame. The angle of change $\theta = cos^{-1}(\frac{y_c - y_p}{x_c - x_p})$ is thresholded to determine if significant change has occurred. If the speed is low, the key frame is not selected.

The first check is a measure of how well the orientation of the tracked target matches the orientation of the key frame used to initialize the tracker.

The first score is a measure of how well the target's new foreground, determined from the foreground detection, fits the tracking model. This is a measure of when the tracker will lose the target.Equation 12 is thresholded to determine when the tracking model no longer fits the target.

The second phase uses the color information to determine whether or not this is a good enough representation of the target to be used as a key frame. A fragmented GMM is used to determine this. The fragments are parts of the targets segmented

according to their color and location. Using fragments slightly encodes the spatial information but the color is the most important. It uses the probability of the target being similar to the candidate key frame to determine whether or not it is acceptable. Because objects go through occlusion some times, it might not fit the tracking model, the second phase determines that this is not actually a key frame we want. The Edison package [135] is used to segment the target into fragments.

The closed form of the L2 norm is used to determine whether or not targets fit each other. Fragments are compared using the equations below.

$$CS = 1 - \frac{\int f_1 f_2}{\| \int f_1^2 dx \| * \| \int f_2^2 dx \|} \tag{13}$$

Let target 1 be represented by GMM $f1 = \sum_{i=1}^{n} \omega_i \eta(x; \mu_i, \Sigma_i)$ and target 2 be represented by GMM $f2 = \sum_{i=1}^{m} v_j \eta(x; \mu_i, \Sigma_i)$.

$$\int f_1^2 dx = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\omega_i \omega_j}{\sqrt{(2\pi)^d |\Sigma_i + \Sigma_j|}} e^{-\frac{1}{2}(\mu_i - \mu_j)^T (\Sigma_i + \Sigma_j)^{-1} (\mu_i - \mu_j)} \tag{14}$$

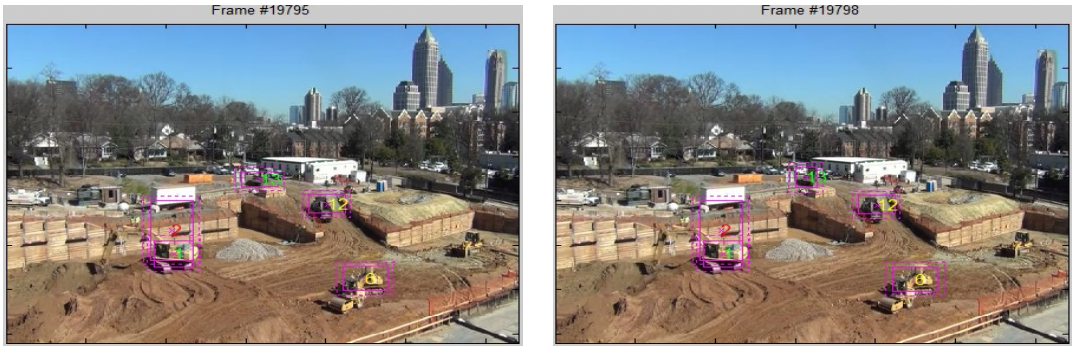
$$\int f_2^2 dx = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{v_i v_j}{\sqrt{(2\pi)^d |\Sigma_i + \Sigma_j|}} e^{-\frac{1}{2}(\mu_i - \mu_j)^T (\Sigma_i + \Sigma_j)^{-1} (\mu_i - \mu_j)} \tag{15}$$

$$\int f_1 f_2 dx = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\omega_i v_j}{\sqrt{(2\pi)^d |\Sigma_i + \Sigma_j|}} e^{-\frac{1}{2}(\mu_i - \mu_j)^T (\Sigma_i + \Sigma_j)^{-1} (\mu_i - \mu_j)} \tag{16}$$

If the distance is above a given threshold, the current template is stored as a key template. As the target traverses the scene, key templates are acquired as needed and concatenated. The target's re-identification model is then learned using shadow densities. Figure 8 shows an example of the key template selection for a target as it traversed the scene. Figure 9 shows an example of thee key template selection when the target changes orientation while tracking.

The target's dominant visual features are learned using KPCA as described in section 2.2. The training data is vectorized and mapped into a higher dimensional

Figure 8: Different key templates of the same target as it traversed the scene



(a) Frame 19795 (b) Frame 19798

Figure 9: New key template selected when target changes orientation

space using the Gaussian kernel. The eigenvectors and eigenvalues are computed and the eigenvectors with the largest eigenvalues are retained. When a target is detected at the entrance locations, its feature vectors are projected to the learned models of previously seen targets. The similarity score is given by Equation 12. If score, $\mathcal{SC}$, is above the given threshold, the target is declared new and it's first key template is initialized. If the score is below the given threshold, it is declared as the target with the minimum score. The target's tracker is then re-initialized and tracking continues.

This thesis also relies on the re-identification to solve the moving and static occlusion problem. Moving occlusions occur when two moving targets occlude each other in the camera's view. Static occlusion occurs when a target gets occluded by a background object. To detect a static occlusion, after localizing the target in frame, the bounding box for the target is checked against the detected foreground. If the bounding box does not contain a certain amount of foreground pixels, and the target is not

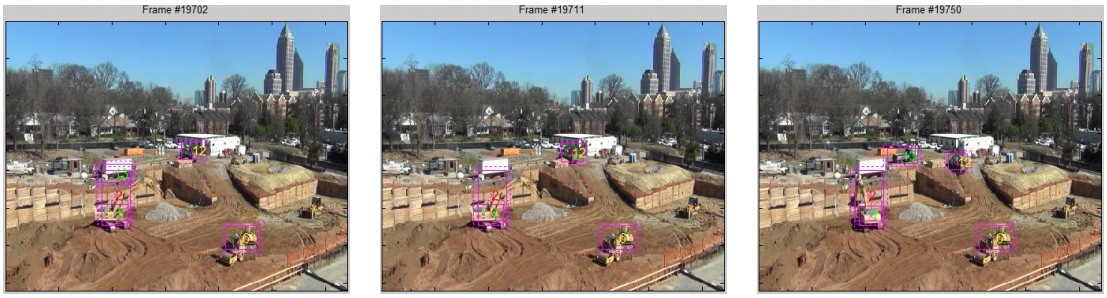


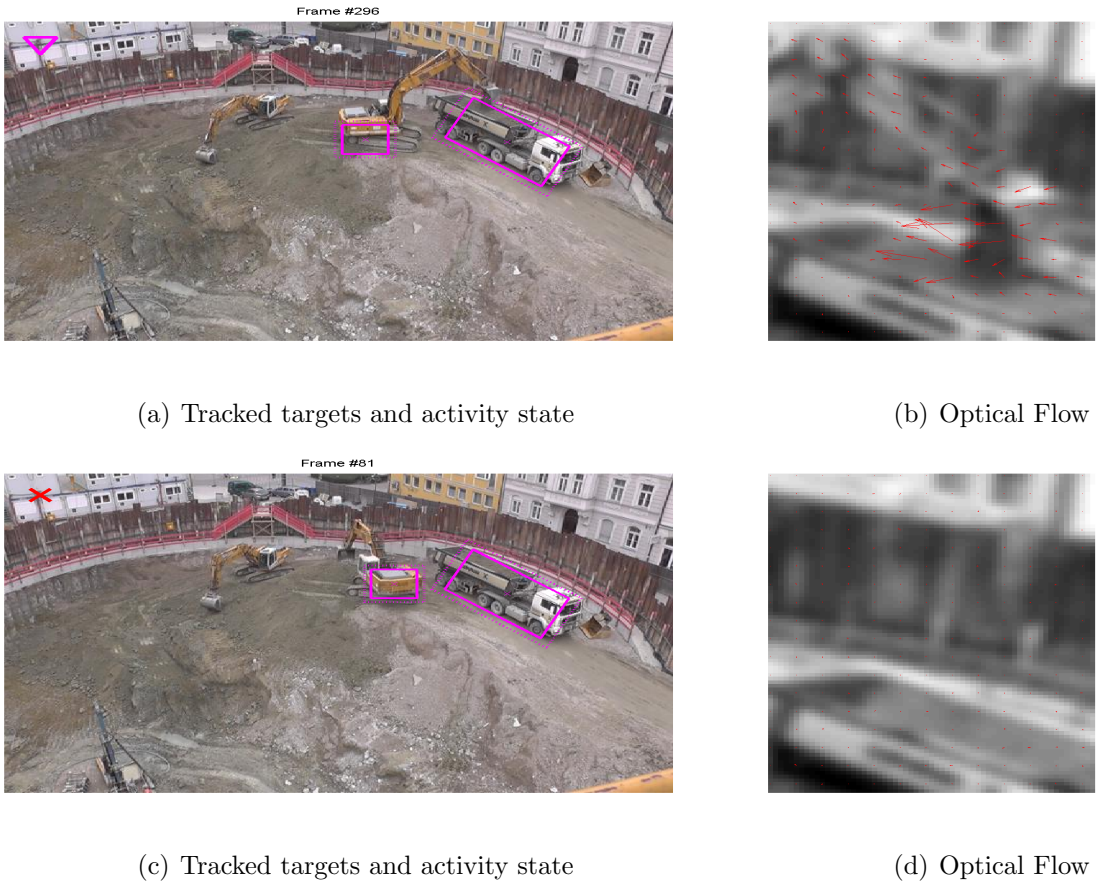
Figure 10: Target as it goes through a static occlusion and gets re-identified correctly when it re-appears

near an entrance/exit region, it is assumed the target is occluded and the tracker is temporarily suspended. The detector now expects the target to show up somewhere around the region where it became occluded. If a target detected around that region, the model of the occluded target will be compared to this target to see if it was the disappeared target. For moving occlusions, the proximity of each target's bounding box to other targets' bounding boxes is used to determine when they occlude each other. As the bounding box separates, or if a new target is detected around the region where the targets started occluding each other, the re-identification module is used to determine which tracker to associate the target to. Figure 10 demonstrates a dump truck going through occlusion, and being re-identified correctly after re-appearing.

2.4 Activity Status Estimation

The construction work sites require an activity status estimator to determine their activity status. The movement of the excavators spatio-temporal features provides an indication of excavator activity. The movement of the excavator in the proximity zone of a dump truck establishes when an excavator is filling a dump truck. This state can only be triggered when the two machines are in close proximity and the dump truck is static. The change in the dump truck's location, $\dot{\vec{x}}(t)$, over time is used to determine whether it is moving or idle. The excavator's arm movement is determined using optical flow. Once the system detects that the dump truck is static

in a region near the excavator, optical flow is performed on a window extracted above the dump truck's centroid. The optical flow vectors are thresholded to determine if there is enough movement if the excavator is filling the dump truck.



(a) Tracked targets and activity state

(b) Optical Flow

(c) Tracked targets and activity state

(d) Optical Flow

Figure 11: Activity status of a dump truck getting filled by an excavator

The results of this online activity status estimation will be used by the event processor to determine what happened in the video

2.5 Event Detection Processor

The proposed surveillance system is used to gather event statistics for work sampling on a construction site and an art installation in a retirement community.

2.5.1 Work Sampling Processor

Four work metrics were measured by the work sampling processor. The amount of time spent on direct work, support work, no work, and absent. Direct work is determined by the amount of time the trucks spend in a region of interest. Different regions of interest correspond to different work activities for different machines. For example, a dump truck's region of interest can be the same as a bulldozer's region of interest but when the dump truck is in that region of interest, its direct work is dumping dirt, and the bulldozer's direct work in that region of interest is leveling the dirt. Multiple construction machines that have a different designated functions are tracked and the work sampling processor produces the amount of time spent performing these different functions. If the truck becomes idle outside the regions of interest, then no work is being performed. If the truck moves outside of the region of interest, then it is performing support work. The change in the truck's location over time is used to determine whether it is moving or idle.The truck is performing direct work when it is inside the region of interest. If the truck is not visible in the scene, then it is considered absent.

The work sampling processor takes as input, tracking results, activity status estimation (if available), entrances/exits, and regions of interest and uses this to generate the statistics needed to determine the time spans of the work activities of the excavators and dump trucks. The metrics computed are the number of dump trucks that entered the scene (n_{trucks}), how much time they spent in the scene (t_{scene}), how much time they spent in the region of interest getting filled (t_{roi}), how many bucket loads of soil were placed in each dump truck ($n_{buckets}$), and how long the machines spent idle while in the scene (t_{idle}). Depending in the machine's ID, the processor will determine the work being performed.

The number of dump trucks that entered the scene, n_{trucks}, is computed by counting the number of trackers initialized for detected dump trucks. Their duration in

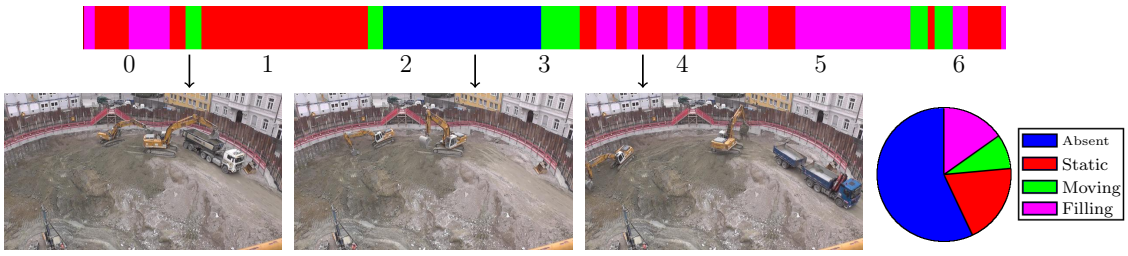


Figure 12: Dump truck state estimates for a video segment of 6 minutes duration, plus activity states in a pie chart.

Table 1: Sample event statistics table

Truck #	Entered	Moving	Static	Filling	Exited	Total	# of Buckets
3	26.75	.36	3.2	1.43	31.74	4.99	8
4	32.85	.59	1.35	2.06	36.85	3.99	9

the scene, t_{scene}, is obtained by subtracting the time stamp they left the scene from the time stamp they entered the scene. The time spent being loaded by the excavator, t_{roi}, is determined using the results from the activity estimation. It sums up the total amount of time the activity estimation detected that the dump truck was being loaded. The number of bucket loads, $n_{buckets}$, is also determined by the activity estimation module by counting how many times it detected the excavator bucket over the dump truck using optical flow (optical flow measures the apparent motion of pixels from one frame to the next). The time the machines spent idle in the scene, t_{idle}, is determined by checking how much movement has happened between frames. Movement below a threshold triggers the idle state.

The event processor tabulates the temporal statistics of the activities and also identifies events, such as filling cycles and outlier time spans. A sample activity timeline and summary pie chart is given in Figure 12 for a 6 minute segment of processed video. Table 1 shows a sample event statistics table that can be generated by the processor.

2.5.2 Interaction Processor

The interaction processor for the retirement community takes as input the tracking results, entrances, and a region of interest. It returns the number of targets that entered the scene, the number of times each target entered and left the scene, how many times they visited the art piece, how long they spent observing the piece, how many times they passed the art piece without stopping, how many targets interacted, which targets interacted, and how many times they interacted. Observations {moving, idle, in ROI, multiple entities}, are acquired from the tracking results. These observations are used to determine the states {observing, no interaction, passing, leaving, approaching} of the targets and the art piece. The change in the target's location, $\dot{\vec{x}}(t)$, over time is used to determine whether the target is moving or idle.

CHAPTER III

CONSTRUCTION SITE WORK SAMPLING

3.1 Introduction

This chapter focuses on a vision-based approach to the automatic estimation of productivity associated to excavation processes on a construction site. Earthwork processes are often subject to unanticipated delays [10], which are likely to propagate through the entire remaining schedule and adversely impact progress, productivity, and costs [11]. The presented work impacts research into site operations by enabling the automated monitoring and tracking of on-site resources. Video-based monitoring and processing algorithms provide a non-intrusive, easy, inexpensive, and rapid mechanism for generating a body of operational information and knowledge. If made available to project stakeholders, the information and knowledge would enable inquiry into construction operations that is currently not possible [12]. Longer term, vision-based research can serve as a valuable aid to project management by enabling tighter control and greater efficiency.

Video sequences used to demonstrate concepts in current literature are typically less than 10,000 frames, and with a typical frame rate of 12 frames per second, this is approximately 14 minutes long [2]. The videos used to demonstrate the proposed system are significantly longer than what is typically used in literature. This thesis shows that event statistics can be generated for longer video sequences by presenting a system that integrates different modules that provide feedback to one another, thus enabling more accurate results.

This chapter demonstrates that the designed system can generate robust statistics necessary to determine work activities and productivity levels. Demonstrating that an active vision system can effectively analyze and assess work-site progress will assist

project managers by reducing the time spent monitoring and interpreting project status and performance, thus enabling increased attention to the control of cost and schedule. By making project management and the workforce more aware of the performance status of their project and their work environment, potential savings to the industry are envisioned. Since benefits in construction often impact a broader theme of issues, they are likely to impact schedule, cost, safety, and quality at the same time.

3.2 Methodology

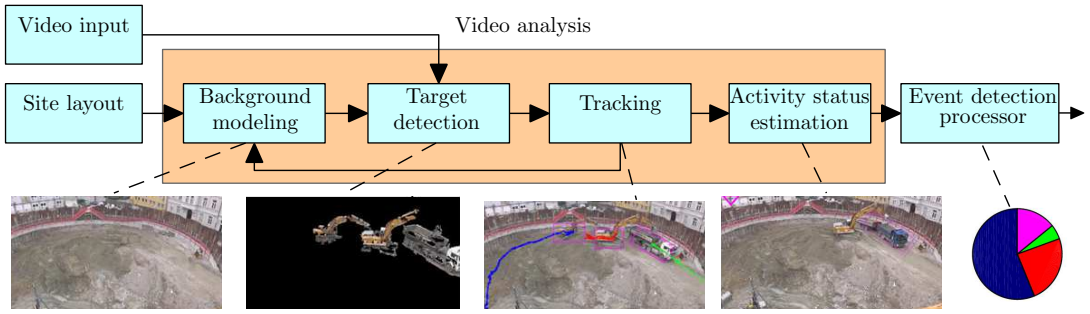


Figure 13: Process flow for the automatic surveillance system.

The proposed system, as seen in Figure 13, was used to process three different construction sites. The videos are from surveillance cameras mounted atop buildings facing the construction sites. The first video contains a loader moving items from one location to another. The second video contains a dump truck and a bulldozer. The third, fourth, and fifth video contains multiple dump trucks and excavators.

The first two videos were videos of the Georgia Tech Clough Undergraduate Learning Commons (CLC) building at different views, and they both contained more than 80,000 frames, or around 1 hour and 51 minutes. The loader moved material from one region to the other. The bulldozer and the dump truck were engaged together in the activity of ground leveling, where the dump truck hauled dirt to a pile, and

(a) Loader ROI

(b) Dump Truck and Bulldozer ROI

Figure 14: ROI for CLC views 1 and 2.

the bulldozer attempted to level the pile. Figure 14 show the regions of interest highlighted in yellow for the two videos. Figure 15 and Figure 16 shows the estimated backgrounds for the two videos.

Figure 15: Estimated background models for Clough construction site (view 1).

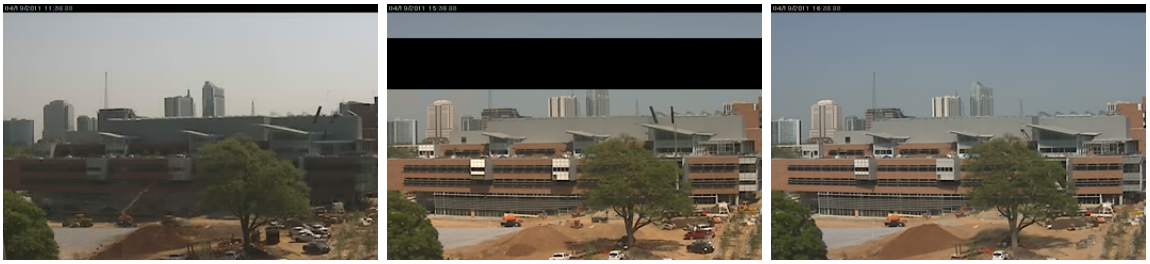


Figure 16: Estimated background models for Clough construction site (view 2).

The third video is of the Georgia Tech Engineered Biosystems Building (EBB).

The video is 3.5 hours long and it contains over 360,000 frames. Figure 17 shows the highlighted regions of interest. Figure 18 shows the estimated backgrounds for the video. The filling ROI is where the dump truck gets filled by the excavator. The hosing ROI is at the exit of construction site. The machines need to get hosed down so that they do not spread dirt all over the streets. The machines are engaged in earth-moving operations. The activity of status of the machines are static, moving, absent, filling, and hosing. The event statistics of interest are the number of trucks that entered the scene, the inter-arrival times, the time spent in the scene, and the time spent in the region of interest getting filled and get hosed down.

Figure 17: EBB entrance/exit region and ROI

Figure 18: Estimated background models for EBB construction site.

The fourth video is of the Josephsplatz parking garaged being built in Munich, Germany. Figure 19 shows the plan view of the construction site. The video is 6 hours

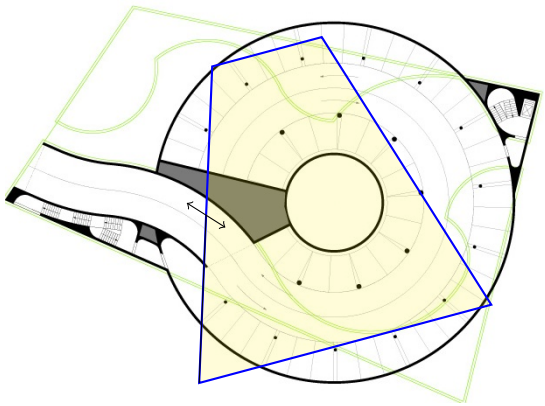


Figure 19: Plan view of the Josephsplatz construction site with camera view overlay (opaque trapezoid).

long and it contains over 700,000 frames. Figure 20 shows the estimated backgrounds for the video. The machines in these two videos are also engaged in earth-moving operations. The activity of status of the machines are static, moving, absent, and filling. The event statistics of interest are the number of trucks that entered the scene, the inter-arrival times, the time spent in the scene, the time spent in the region of interest getting filled, and the number of bucket loads (since the bucket size is provided with this video) removed from the work site.

Figure 20: Estimated background models for Josephsplatz construction site.

The fifth and sixth video are of the translaTUM hospital building located in Munich, Germany. The layout of the site is illustrated in Figure 21, with an overlay of the region depicting the video camera perspective (trapezoidal region) and double-arrows indicating the entrance/exit region of interest for the dump-trucks. Both

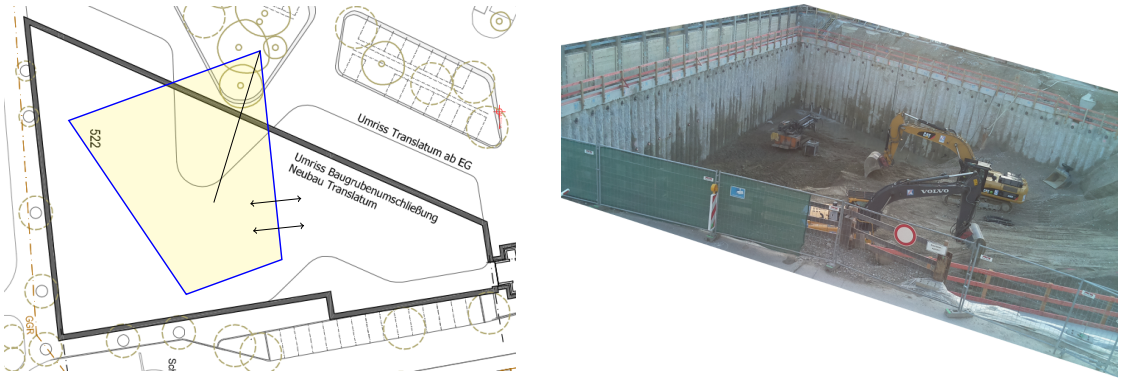


Figure 21: The translaTUM construction site information, video geometry, and image view.

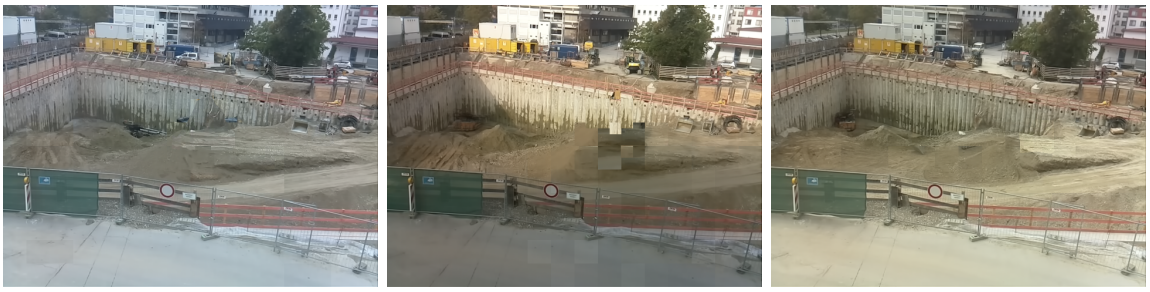


Figure 22: Estimated background models for translaTUM construction site.

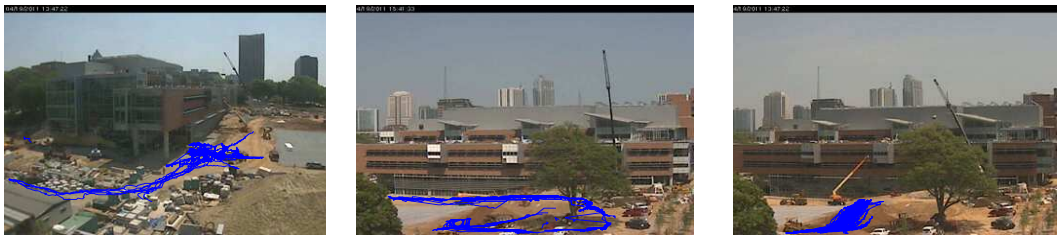
videos are 11 hours, 31 minutes, and 4 seconds long and they contain over 400,000 frames. Figure 22 shows the estimated backgrounds for the videos. The machines in these two videos are also engaged in earth-moving operations. The activity of status of the machines are static, moving, absent, and filling. The event statistics of interest are the number of trucks that entered the scene, the inter-arrival times, the time spent in the scene, the time spent in the region of interest getting filled, and the number of bucket loads removed from the work site.

3.3 Results

The system, Figure 13, is used to process the videos described above. Ground truth was collected by manually reviewing and annotating the videos.

3.3.1 Clough Undergraduate Learning Commons (CLC) Construction Site

The loader was tracked for 86,400 frames at a 5 frame interval. Of those frames, it was visible in 64,740 frames and absent for 21,720 frames. It was detected entering and leaving the worksite 5 times. It visited the ROI 18 times and spent 35,695 frames in the ROI. The system was able to accurately determine the amount of times the loader entered and exited the camera's field of view to pick up materials to be moved to another location. It was idle for 27,670 frames and it spent an average of 1,982 frames in the ROI during each visit and an average of 2,983 frames in between ROI visits. Figure 24 shows a more detailed output from the work sampling processor.



(a) Loader Trajectory (b) Dump Truck Trajectory (c) Bulldozer Trajectory

Figure 23: Trajectories for the different machines tracked (CLC).

Table 2: Automated and manual tabulation of ground leveling task (minutes)

Cycle	Leveling		Dumping		No Work	
	GT	Estimated	GT	Estimated	GT	Estimated
1	5.77	5.89	0.97	1.01	1.18	1.06
2	6.65	6.72	1.02	1.06	1.36	1.25
3	4.64	4.67	0.95	0.95	2.03	1.97
4	5.55	5.72	1.28	1.34	3.67	3.44
5	3.04	3.16	0.97	0.97	3.88	3.75
6	5.21	5.23	1.00	1.04	0.91	0.85
7	5.33	5.55	0.98	1.00	2.23	1.99
8	4.65	4.68	0.93	1.00	0.79	0.70
Error	1.87%		3.23%		6.93%	

The loader was detected 5 times
Entered at frame 1 and left at frame 1050
Entered at frame 15400 and left at frame 15605
Entered at frame 23855 and left at frame 24035
Entered at frame 33595 and left at frame 34665
Entered at frame 76350 and left at frame 76545

The loader was in the region of interest for 35695 frames
At intervals

1	205
5490	14785
14850	14950
18430	23680
33975	34035
38155	39345
40255	44510
57100	58450
60805	63740
65915	66635
70170	70995
71230	75415
75650	75810
76025	76115
80200	82120
82770	85315
85415	85850
86235	86400

The loader visited the worksite 18 times
The loader was not visible in 21720 frames
The loader was visible in 64740 frames
The loader was idle in 27670 frames
The loader was working in 35695 frames
The loader spent an average of 1982 frames in the ROI
The loader spent an average of 2983 frames in between ROI visits
The loader spent an average of 18461 frames out of view

Figure 24: Loader text output (CLC).

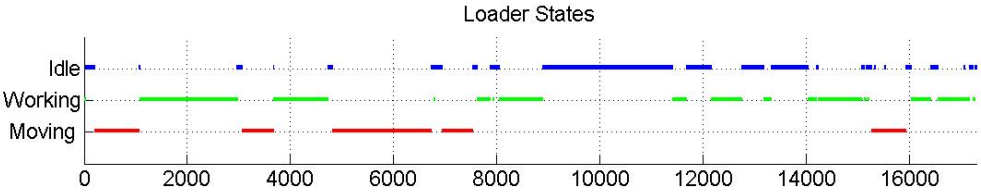
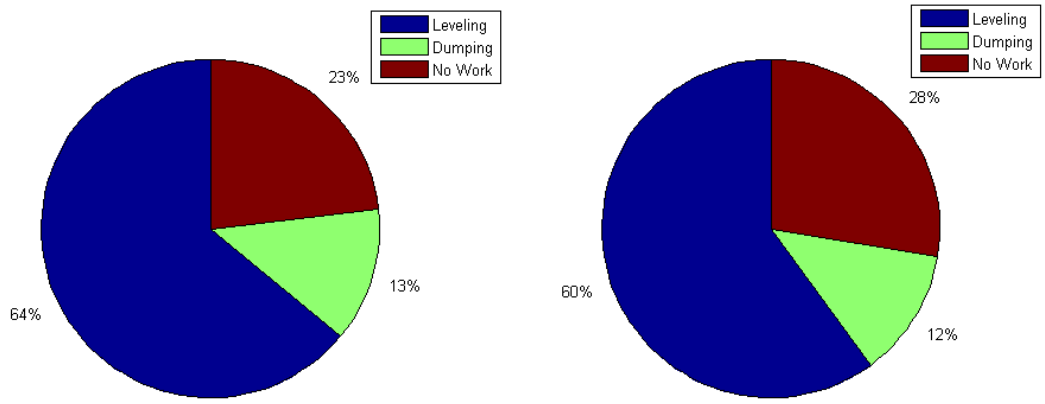


Figure 25: Loader States (CLC).

Figure 25 shows the states of the loader as a plot. Figure 27 and Figure 28 show the work states of the machines over time and Figure 29 shows the activity state in the ROI over time. The dump truck and bulldozer were tracked for 119,740 frames at a 5 frame interval. The bulldozer was detected 10 times and the dump truck



(a) Ground Truth for ROI

(b) Estimates for ROI

Figure 26: Pie chart of the ROI for the Clough construction work site (view 2).

was detected 8 times. The dump truck was detected more times than the bulldozer because the dump truck would exit the camera's field of view in order to turn and head back to pick up more dirt. The bulldozer was visible in 85,180 frames and the dump truck was visible in 48,420 frames. The bulldozer and the dump truck were in the ROI for 80,850 and 16,050 frames, respectively. The bulldozer and the dump truck visited the ROI a total of 16 and 14 times. The estimated average time spent leveling is 5.1 minutes, compared to a ground truth of 5.2 minutes, there was a 1.87% error. The estimated average time spent dumping is 1.01 minutes, compared to a ground truth of 1.05 minutes, there was a 3.22% error. The estimated average time with no work in between the cycles is 2.00 minutes, compared to a ground truth of 2.01 minutes, there was a 0.25% error. As expected, the bulldozer became idle as the dump truck approached the pile and it became active after the dump truck left. Table 2 shows the breakdown of the activities in the ROI. More time was spent leveling than dumping. The states alternated beween leveling and dumping with periods of inactivity between. The results showed that the bulldozer spent majority of its time on the scene, 95%, working while the dump truck spent majority of its time on the scene, 77%, idle.

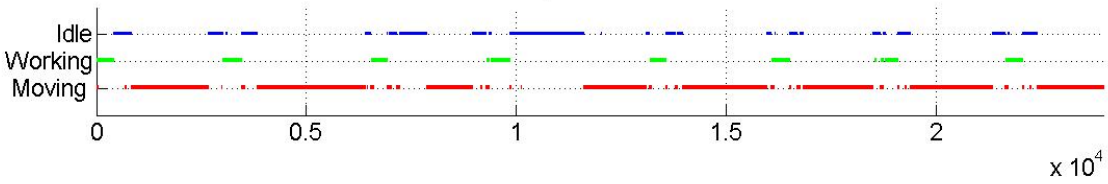


Figure 27: Dump Truck States (CLC).

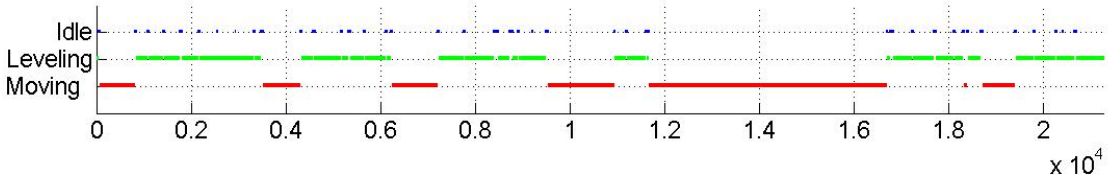


Figure 28: Bulldozer States (CLC).

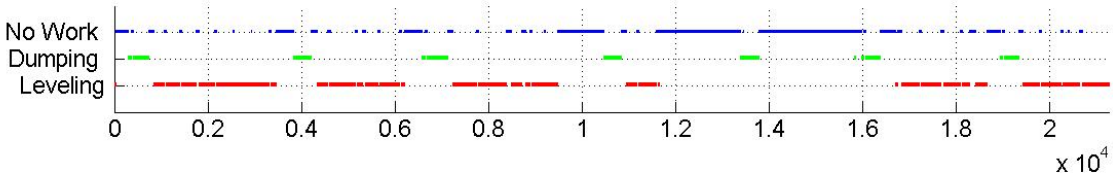


Figure 29: Ground leveling States (CLC).

3.3.2 Engineered Biosystems Building

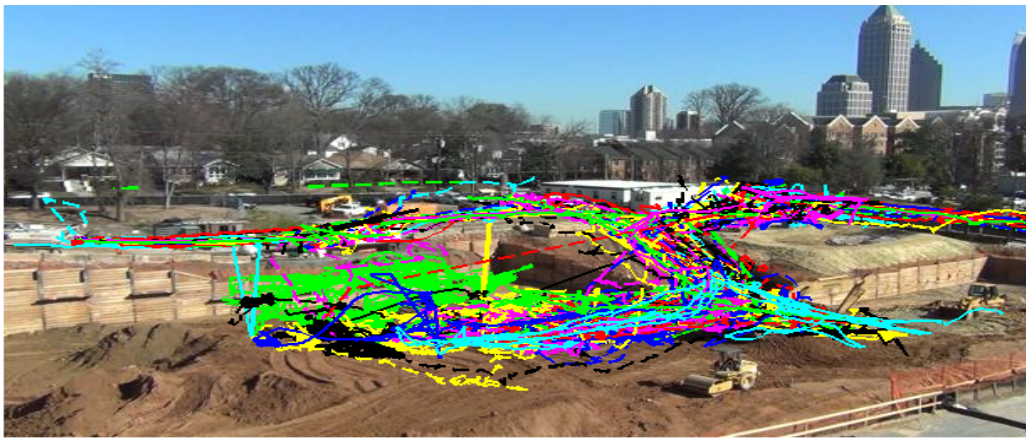


Figure 30: Tracking results for the EBB construction site

A total of 33 dump trucks were detected. Figure 30 shows the trajectories as they traversed the scene. Figure 33 and 34 shows the time they spent in the different ROIs compared to the manually observed ground truth. Excluding the outliers, the red and

magenta line indicates the estimated and ground truth average, respectively. They spent an estimated combined total of 188 minutes in the filling ROI, compared to the ground truth of 169 minutes, there is a 11.2% error. The average estimated time spent by each truck in the filling ROI is 4.8 minutes, compared to the ground truth 4.1 minutes, there is a 16.6% error. The error here is slightly high because there were a lot of moving occlusions. The dump trucks were lining up, waiting to get filled by the excavator around the filling ROI. As the trucks are moving around trying to get correctly lined up, they occlude one another for a short period of time. This results in delay in the system recognizing that a dump truck has left the ROI. The trucks spent an estimated combined total of 83 minutes in the hosing ROI, compared to the ground truth of 80 minutes, there is 4% error. The average estimated time spent by each truck in the hosing ROI is 2.4 minutes, compared to the ground truth 2.3 minutes, there is a 4% error.

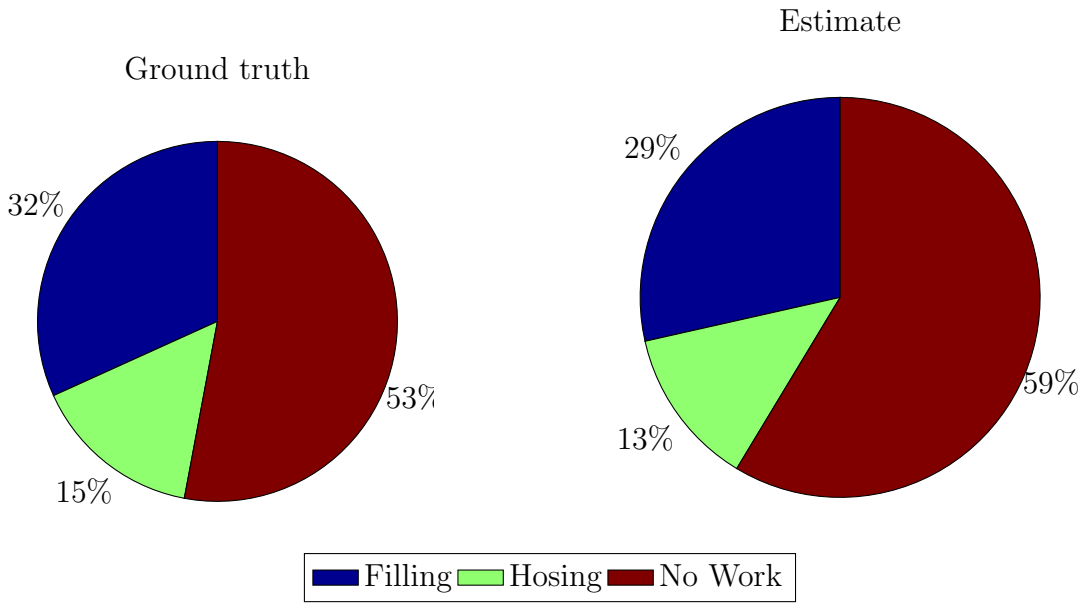


Figure 31: Aggregate statistics for the EBB construction site.

Figure 31 shows the percentage of time spent performing hosing, filling, or no work. It shows that the machines spend half of their time on the scene, waiting, not performing any work. This can be seen by also be seen by the inter-arrival

time between the trucks in Figure 32. The inter-arrival time is the time duration between when the previous dump truck enters and the next dump truck enters. The average estimated inter-arrival time is 3.81 minutes, compared to the ground truth 3.81 minutes, gives a 0.05% error which shows that system was able to accurately detect when the trucks entered the scene. The average inter-arrival time is less than the time each dump truck spends in the filling and hosing ROI. This means that there will be more waiting. The manager could take this information and use it to decide to either reduce the number of dump trucks employed, or get more excavators for performing the earth-moving operation.

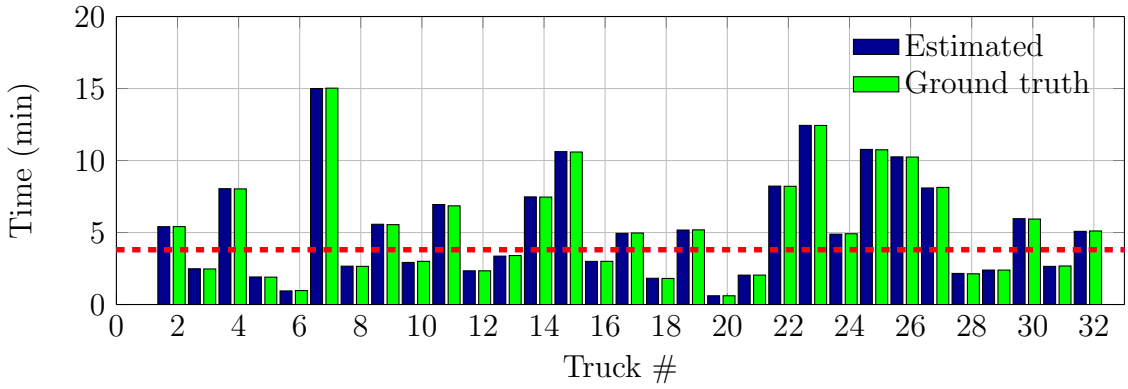


Figure 32: Inter-arrival times between trucks (EBB).

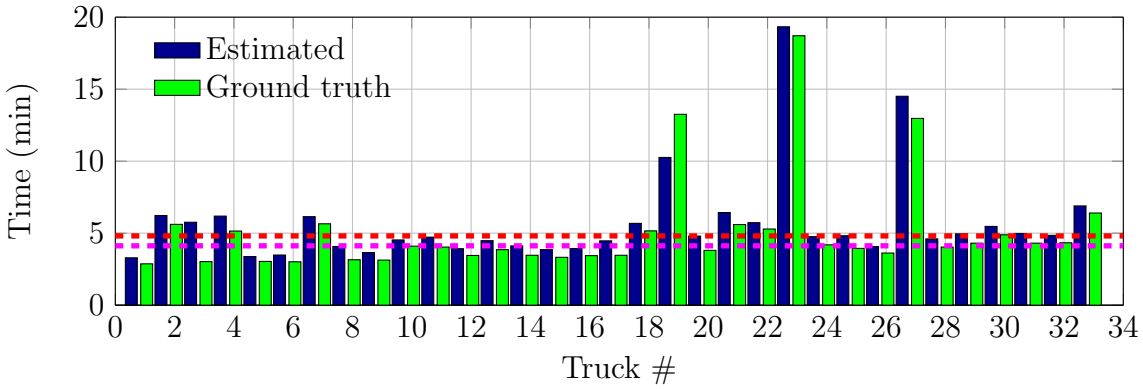


Figure 33: Time spent in the filling ROI (EBB).

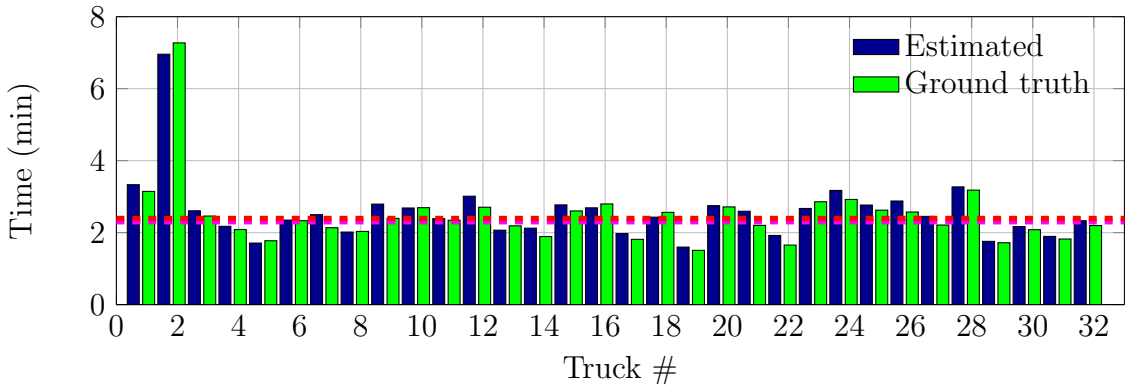


Figure 34: Time spent in the hosing ROI (EBB).

Figure 35: Tracking results for the Josephsplatz construction site

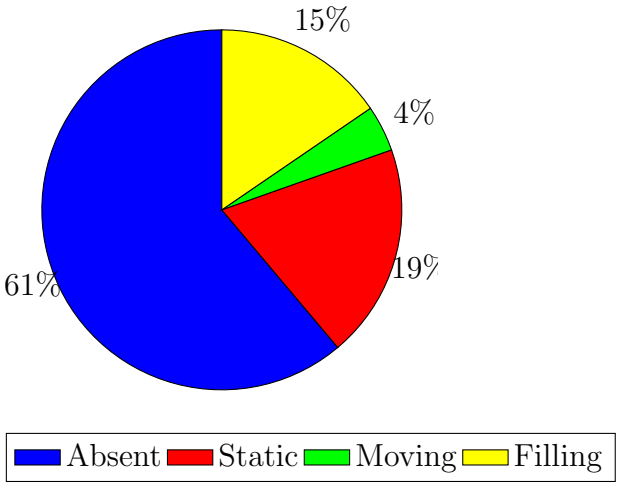


Figure 36: Dump truck states (Josephsplatz).

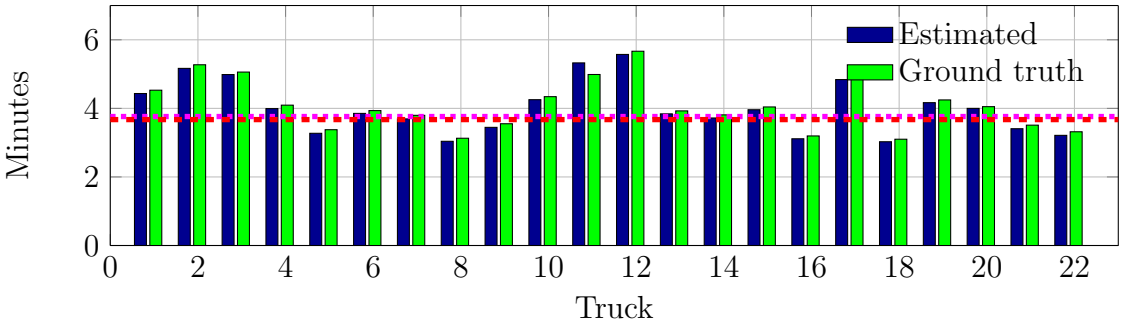


Figure 37: Loading time per truck (Josephsplatz).

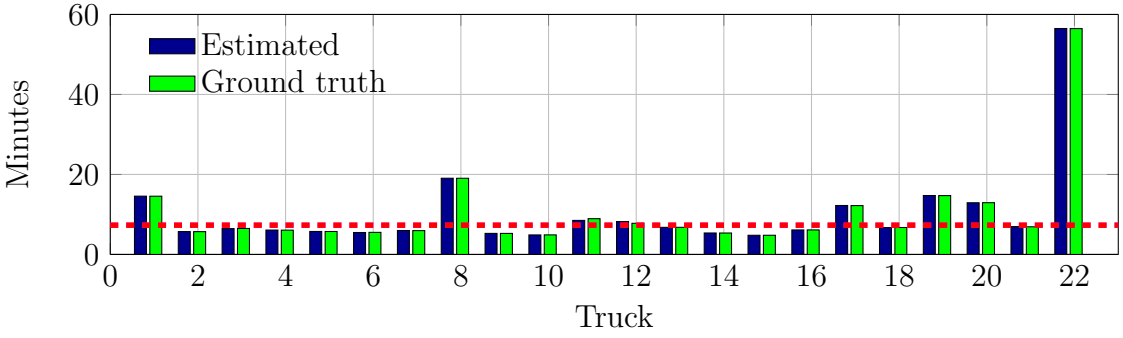


Figure 38: Inter-arrival times between dump trucks (Josephsplatz).

3.3.3 Underground Parking Garage Josephsplatz

For the recorded four hour period an excavated volume of $\Delta v = 418m^3$ was calculated. This yields a performance factor of $\tau_p = \frac{418m^3}{4h} = 104.5\frac{m^3}{h}$ The event processor combined the tracking and activity estimation results to generate the illustrated statistics. A total of $n_{trucks} = 22$ were detected in the processed video, with 0% error compared to the ground truth. A total of $n_{buckets} = 171$ bucket loads were detected compared to the ground truth of $n_{buckets} = 177$ bucket loads (3.4% error). The total volume of the excavator's bucket was $2.5m^3$. Using the ground truth $n_{buckets,gt} = 177$, results in a performance factor of $\tau_{v,gt} = \frac{2.5m^3 \cdot 177}{4h} = 110.63\frac{m^3}{h}$. Using the estimated $n_{buckets} = 171$, results in a performance factor of $\tau_v = \frac{2.5m^3 \cdot 171}{4h} = 106.88\frac{m^3}{h}$.

Calculating the corrected volume of soil excavated per bucket, using the ground truth, results in $v_{bucket\_cor,gt} = \frac{418m^3}{177} = 2.36m^3$. Calculating the corrected volume of soil excavated per bucket, using the estimation, results in $v_{bucket\_cor} = \frac{418m^3}{171} = 2.44m^3$.

Figure 37 shows the total time each dump truck spent in the scene for loading, compared to the manually observed ground truth. The red line indicates the estimated average loading time per dump truck, excluding outliers, which was 3.67 minutes, while the pink line indicates the ground truth average time per dump truck which was 3.76 minutes, for a 2.4% error between the estimated results and ground truth. There were five identified outliers taking 4.8 minutes or more. The estimated total amount of time the dump trucks were in the scene was 88.3 minutes out of 240 minutes of video (39% of video duration), compared to the ground truth of 89.9 minutes (37.5% of video duration), with 1.7% error between the estimated results and ground truth. The inter-arrival time is the time duration between when the previous dump truck enters and the next dump truck enters. The inter-arrival time for this video sequence is shown in Figure 38. For the first dump truck, the inter-arrival time quantity measures the amount of time from the start of the video to when the dump truck first entered. The estimated average inter-arrival time between the dump trucks was 7.317 minutes excluding the automatically identified outliers, while the ground truth was 7.315 minutes for a 0.02% error.

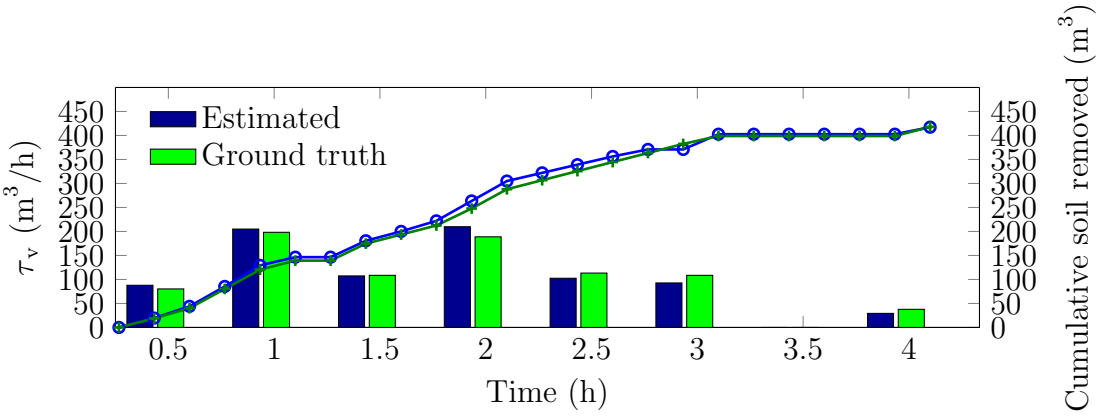


Figure 39: Soil removal statistics for Josephsplatz site

The statistical evaluation of the video analysis results is depicted in Figure 36. It indicates what percentage of total recording time the different activity states of the

dumper trucks were observed. The aggregate statistics indicate that it is possible to improve on the efficiency of the process by incorporating more dump trucks to reduce the idle times of the excavator and increase the amount of soil removed from the site since only 39% of the available filling time was used. Analyzing further, the fairly steady arrival of dump trucks, with the exception of the few outliers, indicates that there were no transportation issues. The average loading time, 3.76 minutes, and the average inter-arrival time, 7.315 minutes, also show that it's possible to incorporate additional dump trucks. The inter-arrival times did increase in the latter half of the day, the source of which should be investigated further by the project manager.

Using the corrected bucket capacity $v_{bucket\_cor} = 2.36m^3$ together with the detected bucket load provides soil removal estimates. Figure 39 shows a plot of the cumulative soil removed over the 240 minutes of recorded video compared to the ground truth. The volume versus time plot provides information on the overall productivity of the excavator and dumper collaboration. Progress was steady for the first three hours, then stalled.

3.3.4 Hospital Building translaTUM

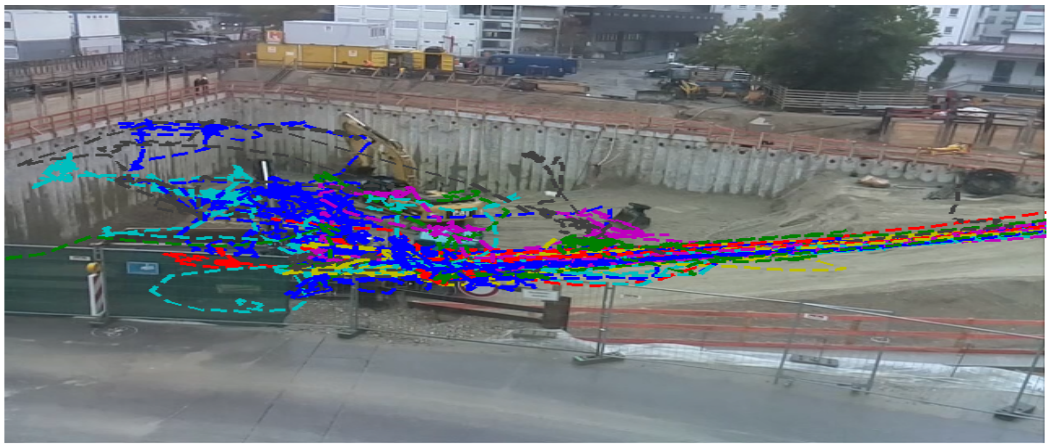


Figure 40: Tracking results for day 8 of the translaTUM construction site.

Video recorded on days 8 and 11 of the earth-moving operation was processed

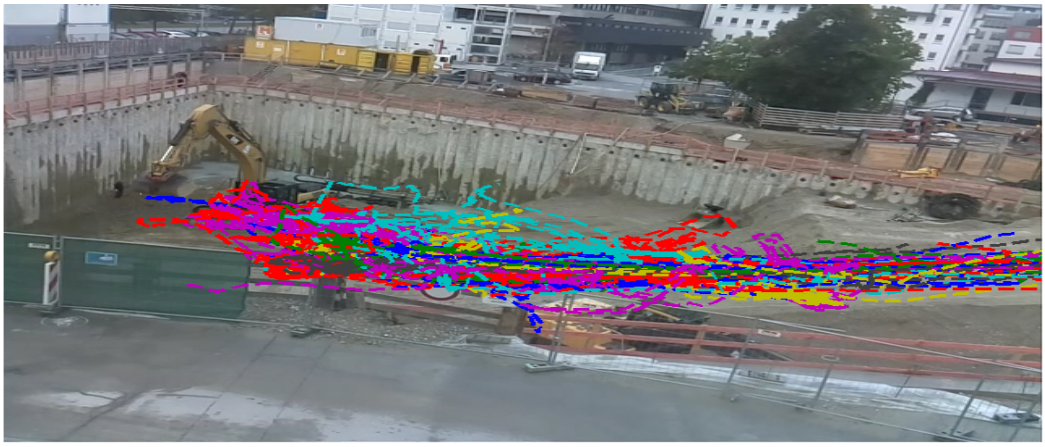
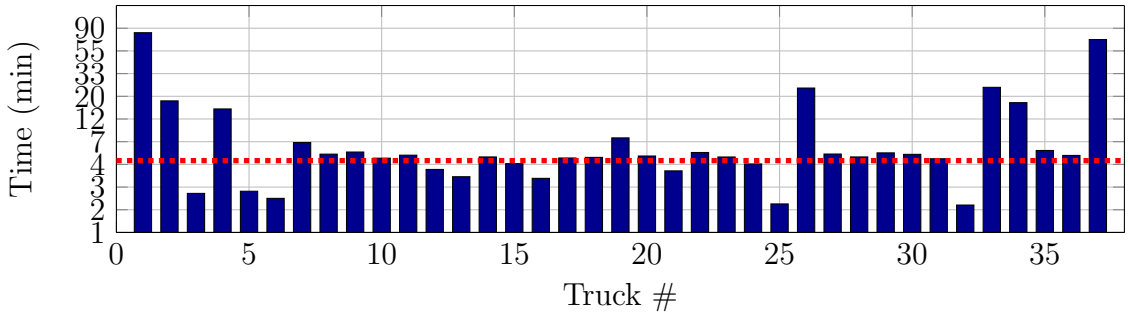


Figure 41: Tracking results for day 11 of the translaTUM construction site.

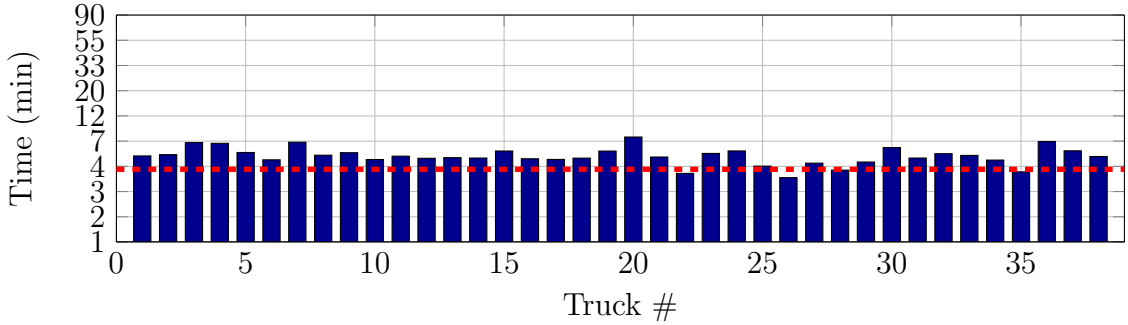
(days 9 and 10 fell on the weekend). The ground truth was again manually annotated for the activity information sought. On day 8, 37 out of 38 trucks were detected by the system. The system detected 305 out of 313 bucket loads (2.56% error). On day 11, the system detected 74 trucks when there were actually 73 trucks, and 334 out of 596 bucket loads (55.04% error). The total volume of the excavator's bucket was $1m^3$. Using the ground truth $n_{buckets} = 909$, results in a performance factor of $\tau_{v,gt} = \frac{1m^3 \cdot 909}{24h} = 37.88\frac{m^3}{h}$. Using the estimated $n_{buckets} = 639$, results in a performance factor of $\tau_v = \frac{1m^3 \cdot 639}{24h} = 26.63\frac{m^3}{h}$. The measurement errors on day 11 lead to a 29.70% error in the performance factor.

Calculating the corrected volume of soil excavated per bucket, using the ground truth, results in $v_{bucket\_corgt} = \frac{774m^3}{909} = 0.85m^3$. Calculating the corrected volume of soil excavated per bucket, using the estimation, results in $v_{bucket\_cor} = \frac{774m^3}{639} = 1.21m^3$. The performance factor error leads to a larger, unrealistic excavation volume for the bucket.

Figures 43 and 42 contains charts of the amount of time each truck spent in the scene as part of the loading process, for the two days analyzed. The average estimated loading time for day 8 is 4.87 minutes, with a ground truth time of 4.23 minutes, for



(a) Day 8 estimate



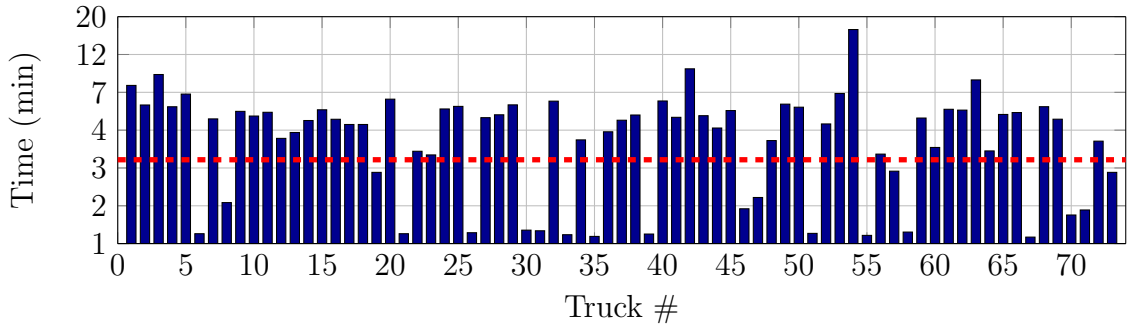
(b) Day 8 ground truth

Figure 42: Time spent per dump truck in the scene while being filled for day 8 (translaTUM) (Log scale).

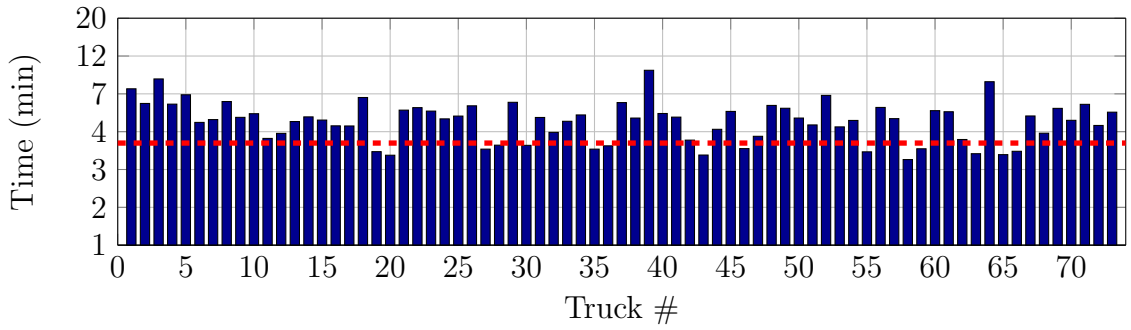
a 15.1% error. The average estimated loading time for day 11 is 3.03 minutes, with a ground truth loading time of 3.85 minutes, for a 21.30% error.

The major outliers over the average red line in Figure 42(a) were due to the tracker losing the targets and drifting to another foreground object in the scene. Track loss was caused by the excavator moving behind the fence. Such errors can be prevented by better camera placement. The major outliers below the average red line in Figure 43(a) were due to occlusion-based track loss. The detector detected trucks it had missed when they first entered the scene, and it also re-detected the trucks the tracker lost as they exited the scene.

Figures 45 and 44 charts the inter-arrival times of the trucks tracked as well as the ground truth (the red line is the averge value). The average estimated inter-arrival time for day 8 is 6.07 minutes, with a ground truth time of 6.24 minutes, for a 2.72%



(a) Day 11 estimate



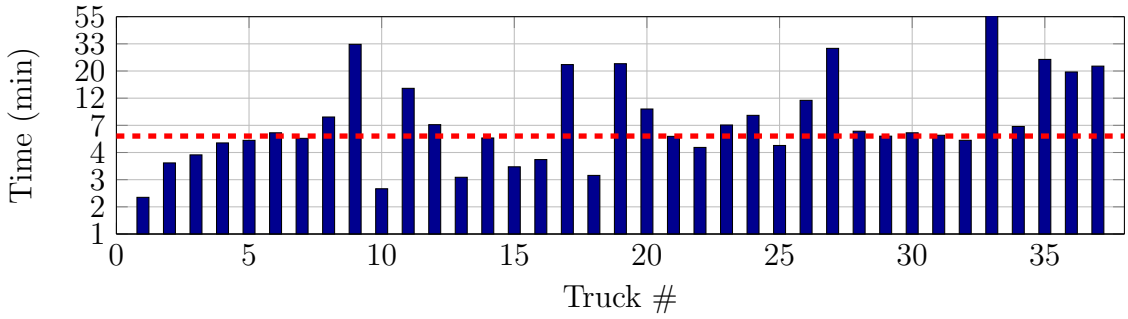
(b) Day 11 ground truth

Figure 43: Time spent per dump truck in the scene while being filled for day 11 (translaTUM) (Log scale).

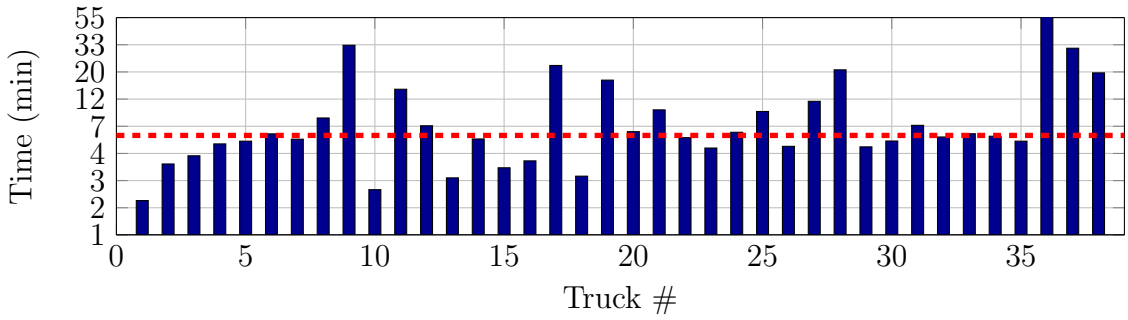
error. The average estimated inter-arrival time for day 11 is 6.07, with a ground truth time of 6.4 minutes, for a 5.44% error.

The event processor generated the pie charts in Figure 46 of the aggregate statistics for the two days. The statistics for the two days are similar when considering the amount of time that dump trucks were absent versus within the scene. Much like the Josephsplatz case, the high percentage of dump truck absence means that additional dump trucks should improve progress.

Using the bucket capacity of $1m^3$ obtained from capping the corrected bucket capacity v_{bucket_cor} to the maximum possible, together with the detected bucket loads over time provides the soil removal statistics as depicted in Figures 47-48. Figure 47 shows the estimated amount of soil removed per hour on both days, and Figure 48 shows the cumulative soil removed on both days. The first day does a good job



(a) Day 8 estimate



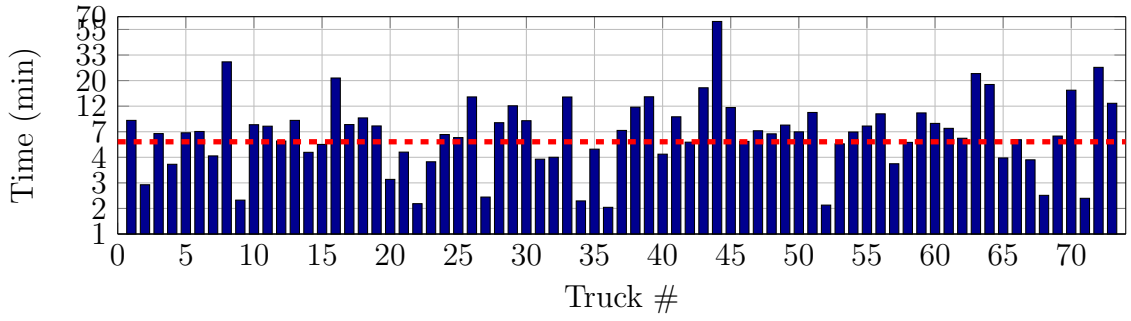
(b) Day 8 ground truth

Figure 44: Inter-arrival times between trucks for day 8 (translaTUM) (Log scale).

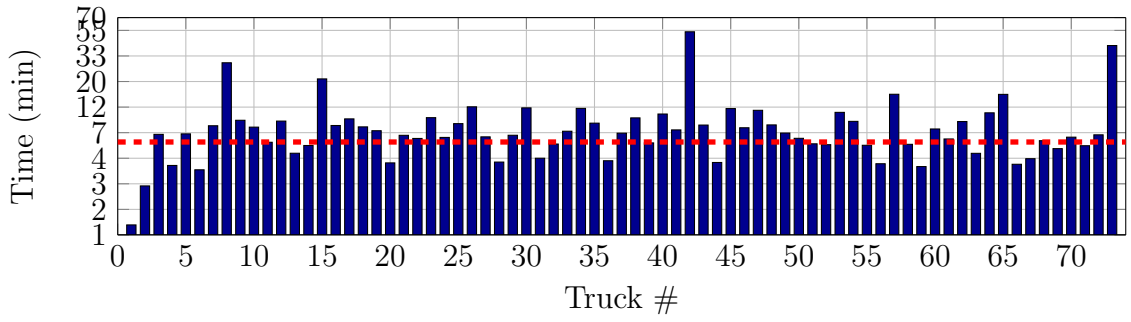
matching the ground truth. Due to near total occlusions of one of the excavators on the second day, many of the shovel loads were missed. The estimated volume excavated is 639 m^3 which is 17.44% off from the photogrammetry measurement. In both cases, consideration of the slopes without regard to the actual values provides a high-level indication of productivity throughout the day. On Day 8, productivity was quite good up until the fourth hour (240 minutes in) at which point it practically halted. Day 11 productivity experienced a lull at about 5 hours in (300 minutes), then resumed for about 3 hours before dying off.

3.3.5 Discussion

The goal of this chapter was to demonstrate that a vision-based approach can be used for automatic estimation of productivity associated to earthwork processes on a construction site. As mentioned, earthwork processes are usually subjected to



(a) Day 11 estimate



(b) Day 11 ground truth

Figure 45: Inter-arrival times between trucks for day 11 (translaTUM) (Log scale).

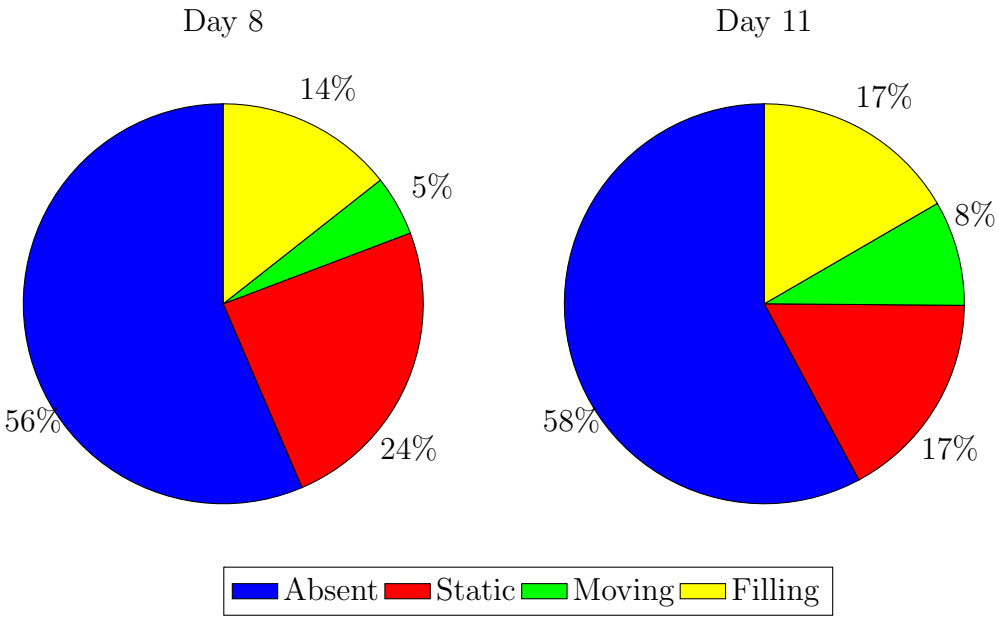
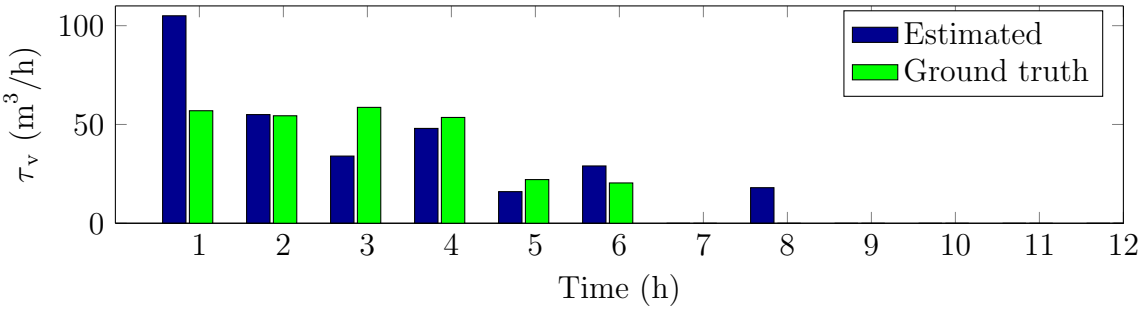
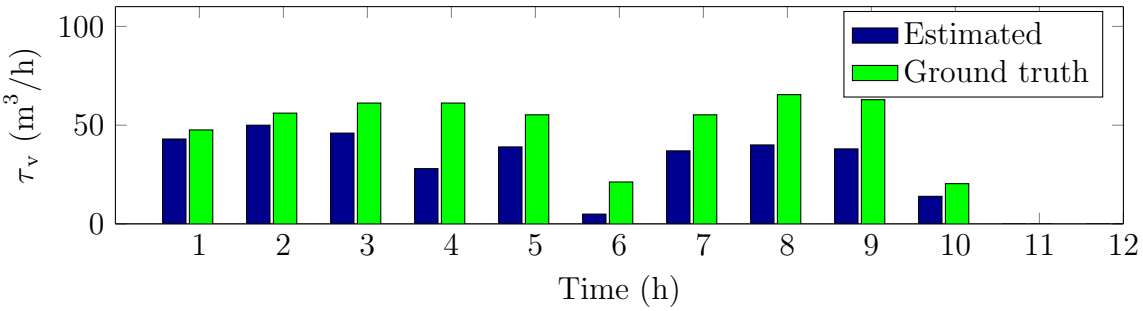


Figure 46: Aggregate statistics for translaTUM.



(a) Day 8 estimate and ground truth



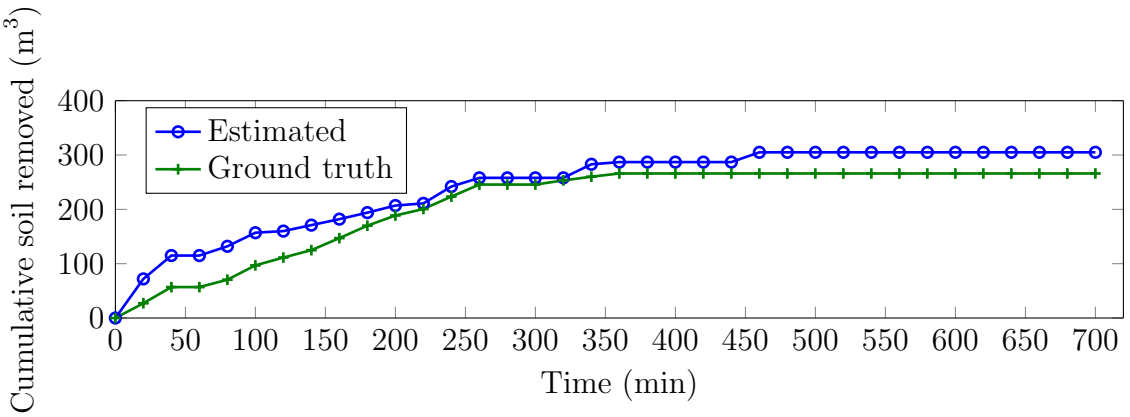
(b) Day 11 estimate and ground truth

Figure 47: Performance factor over time (translaTUM) .

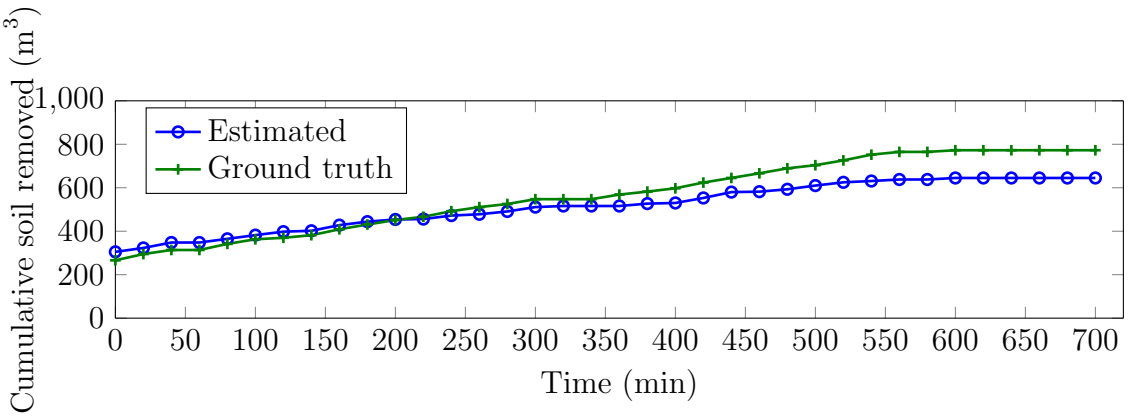
	Detection	Time in scene	Time in ROI	Inter-arrival
CLC	0	2.83	2.10	0
EBB	0	10.60	9.30	0.05
JosephPlatz	0	1.70	2.4	0.02
translaTUM	1.80	3.46	18.07	3.96

Table 3: Error % for the different construction sites

unanticipated delays that can adversely impact progress, therefore, having a way to measure the amount of work performed by the machines is useful to project managers. As shown in Table 3, the system is able to detect whenever machines are present in the scene, which is useful to managers who need to know the exact amount of resources being used. The system was able to provide the statistics on the total amount of time each machine spent on the construction site with less than 11% error. The system was able to give the total time in the ROI with less than 19% error and the inter-arrival times between the trucks with less than 4% error. The errors were mostly due to severe occlusion and illumination changes because of sudden cloud movement. The



(a) Day 8 estimate and ground truth.



(b) Day 11 estimate and ground truth.

Figure 48: Estimated cumulative soil removed per 20 minute interval (translaTUM).

system was able to correctly associate a target lost due to occlusion to its tracker once it was no longer occluded but the amount of time the target is lost however is not accounted for in the statistics since the target's whereabout is unknown during the occlusion.

Although the system was not 100% accurate, a skilled project manager can still identify the productivity levels given the results provided. For instance, given the results from the EBB construction site, they can tell that there is definitely a large amount of time spent waiting by the dump truck drivers. For the JosephPlatz and translaTUM construction site, the opposite is true. The excavator spends a lot of

time waiting for the dump trucks to come in. Given the information provided by this thesis, improvements can be made to the productivity of earth-work processes on these construction sites.

Future work can consider utilizing multiple cameras to handle occlusions. Future work can also consider performing event detection on other machines on the construction work site to create a more comprehensive report of the construction site as a whole. The results can be connected to progress information in order to generate productivity estimates and determine ways to increase productivity.

CHAPTER IV

SENIOR HOUSING SURVEILLANCE

4.1 Introduction

Social interactions in retirement communities' shared spaces are a key component to preventing social isolation and loneliness among older people. Given the under-utilization of these spaces, placing technologies to promote utilization of- and socialization in- shared spaces might improve independence and quality of life among older adults. This chapter describes the use of the proposed semi-automated system to generate statistics regarding the effect of a proposed technological intervention. The long term goal is to achieve quantitative analysis of long-term responses to designed technological interventions by older adults in retirement communities. Specifically to demonstrate positive socialization impacts through these interventions. This chapter communicates the initial efforts towards that goal, focusing on two aspects: (1) observing the resulting effects of a designed technology installed in retirement community shared spaces, and (2) the design and demonstration of a surveillance system for quantitative evaluation of social interactions. The preliminary research showed that the designed system is capable of generating statistics from video data over a short period of time. This chapter demonstrates that developed system can be used to quantify interaction statistics associated to interventions in a retirement community in less time than it takes for a human to manually annotate the same amount of time, thereby demonstrating it as a valid decision support tool.

4.2 Methodology

This section describes the current understanding of the use of public spaces in two local independent retirement communities (with populations of about 200 each). Being an independent retirement community means, the population is considered to be

in good physical and mental health, and each member has a private living space. Each floor has a shared common area furnished much like a living room (see Figure 49, right). They have access to a shared dining hall with predefined meal times, and are free to participate in the communities' activities. The community activities are organized to promote socialization amongst the older adults and vary from being single-time events, to being weekly or monthly events. This section describes a mixed methods study involving both qualitative observations, to understand the state of the existing communities with regards to socialization, as well as a technological intervention study, to analyze the potential impact of introducing designed technology and its impact on socialization.

4.2.1 Qualitative Observations

Unstructured observations were performed by students conducting research in the Industrial Design Department of Georgia tech visiting the two local retirement communities. The observations were conducted at different times of the day and different days over the course of a year (with almost weekly visits, or multiple visits during the week, during some portions of the study). The focus was to document the living spaces and their utilization. Researchers had the opportunity to observe older adults and assess their use of common areas. The qualitative studies involved collecting photographic evidence, such as depicted in Figure 49, and personal notes regarding older adults and their activities in the community and in the shared spaces. They spoke to the older adults informally as the occasion arose, in efforts to preserve the natural feel of the environment. They also spoke to staff at the communities.

Shared spaces were consistently empty (see Figure 49, right) and the spaces that afforded higher rates of use (i.e. main floor gathering living rooms) were utilized by a small percentage of the community population. Much like was seen in [110], the televisions in the shared spaces did not promote social interaction. However here, they

Figure 49: Qualitative studies in retirement communities.

were predominantly powered off and thus merely occupied space. In one instance, the residents indicated that the television was non-functional, had been so for a while, and then asked if the researchers could fix it.

The observations regarding social interaction indicated many members had difficulty interacting or did not at all, with a small percentage of the community being social. It was common to see two, or more, older adults co-located in the same shared spaced but not interacting socially (as noted in other literature [110]), or doing so unsuccessfully. For example, we observed some community members repeatedly using the same topic (typically children) to initiate conversation, even if it had been attempted with the recipient before.

Even though social activities were promoted in the community by staff coordinators and publicized through bulletin boards (Figure 49, middle), attendance was poor. This observation was also confirmed from feedback provided by staff members. Yet, there was a definitive need by the members to socialize. Aside from the scheduled activities, the only other event of the day common to all residents is lunch time, which is at the same time daily. Residents would consistently arrive considerably earlier to sit in chairs lined along the corridors and wait for lunch service to commence (Figure 49, left). The waiting time and available chairs become a mechanism for older adults to interact with one another on a daily basis. This observation was also confirmed by community staff noting that the behavior was mainly performed for socialization,

and that the space had been filled with chairs as a consequence of their behavior.

In summary, it was found that the spaces designated for social activities in retirement communities were underutilized to the extent of being neglected. These spaces are suitable for social interactions but they are not effective in their current conditions, due to underutilization. Further, there was a need for socialization as indicated by a repurposed public area, and potentially also a need for mechanisms to induce diversity in the conversation topics.

4.2.2 Technology Interventions Implications

To better understand the current shared spaces and how modifications, specifically design and technology interventions, to the shared spaces could promote their use and ultimately impact socialization, the qualitative studies were followed up with an experiment implementing a designed technology in the shared spaces of one of the retirement communities. As an initial study on the impact of technologies among older adults in retirement communities, we designed an intervention consisting of an iPad tower, Figure 50. The tower was about waist height and had embedded in it an iPad for viewing. The iPad displayed a looped slideshow consisting of 180 images from two categories: 1) images of landscapes, flowers, monuments, historical events, and of powerful/impactful events from Times magazine and from Pulitzer Prize winning photographs; and 2) images of residents' decorations in the community, taken during the qualitative study.

The iPad tower and its display contents were chosen to address the previous observations, as well as those from the social robotics literature. Given that the television and VCR combination in the shared spaces is too complicated for the residents to interface with when things go wrong, and that it does not in general promote interaction, a simpler design with no controlling interface was preferred. Further, as the social robotics literature indicates that robots are more effective at

Figure 50: Designed technology intervention.

promoting social interaction when a caregiver is present [110], or requires a caregiver to fulfill its purpose [103], it was believed that the choice of a simple technology with a slideshow would provide more diverse stimulation than a functionally limited robot would, while not requiring a caregiver for continual stimulation of the older adults. Currently, social robotics cannot fulfill the expectations of older adults with regards to their capabilities [102, 136, 137, 110, 138], thus motivating a simple technology with easily met power requirements, that did not produce unrealistic expectations, and yet that was also quickly and intuitively understood. The photo choices were meant to elicit memories from the past and the present. Memories were noted to be effective at eliciting discussion in [110].

4.2.3 Experimental Intervention

The experimental intervention was implemented one of the local retirement communities studied; a gated community offering apartment homes for adults aged 62 and older. The community is an 11-story building in which all floors are equally arranged including a common shared space near the elevators and long corridors for accessing

individual apartments. Each floor has approximately 20 units with most of the older adults living alone. This building provides an ideal setting for experimentation to randomly sample floors in order to study older adults' behaviors in the space. Residents' behavior was observed under two conditions: no technology intervention in the shared space and technology intervention in the shared space.

The observations were performed on three floors, one with a pre-existing entertainment activity (a puzzle), Floor 7, and one with only the standard furnishings, Floor 3 (randomly chosen), and one chosen to never receive the intervention, Floor 2 (randomly chosen). Further, the two intervention-receiving floors were initially observed for a period of time in their natural configuration to get a measurement of the naturally occurring utilization, after which the technology intervention was placed in the area. The floors do not vary in structure, thus the floor selections are expected to be representative with regards to the population of the community, which consists of male and female older adults. The population was informed that there would be an experiment, that it would involve a technological intervention, and that the social responses to it were being studied. During the surveillance period, a flyer was posted for all to be aware of the recorded. As shown in Table 4, the floors receiving the technology intervention were monitored for a week to get a measure of their baseline utilization before the intervention was introduced. Since the three floors have identical structures, the same equipment layout shown in Figure 51 was used. The iPad tower was installed in front of the white structure in the middle of the common area, next to an elevator in the retirement home. The cameras were strategically positioned to give the best coverage of the common area and the all its access routes.

For data collection, a script to record video data at 12 frames per second over a period of days was written in PERL. The video data was collected over a period of 4 weeks as indicated in the table below. Video data is approximated to be 21GB a day, so for recording 4 weeks of data, a 1TB hard drive was needed for each computer.

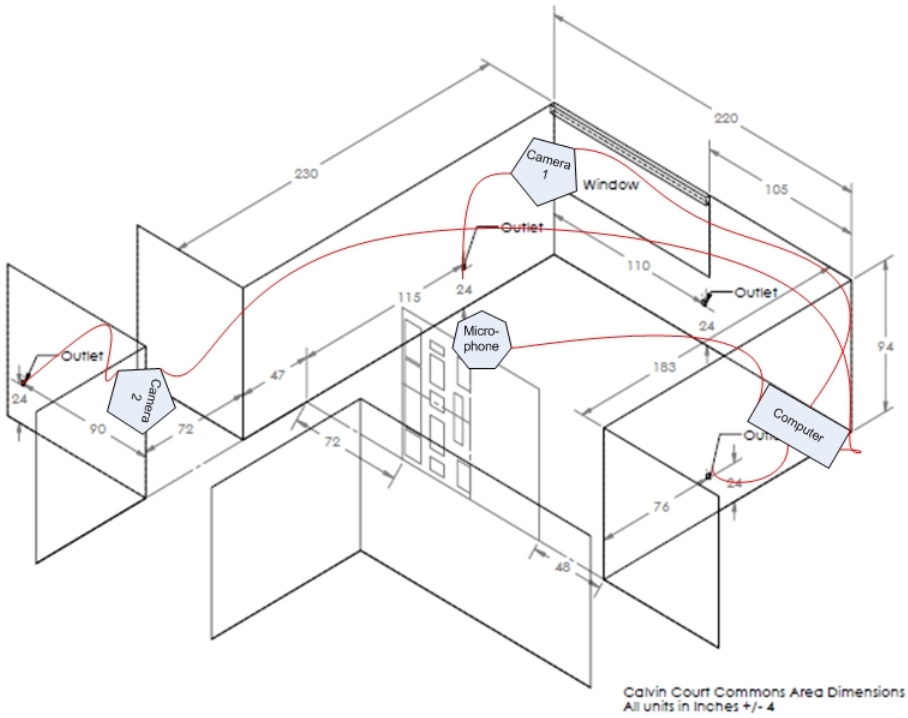


Figure 51: Equipment Layout

The total equipment needed for the set-up consisted of 4 PTZ cameras, 2 computers, and 2 1TB hard drives.

Table 4: Data collection table

Floor	Week 1	Week 2	Week 3	Week 4
2	C	C	C	C
3	C	T		
7			C	T

C - Current behavioral patterns without intervention
T - Technology intervention

4.3 User Interfaces

While the data was being collected, the simple graphical user interface (GUI) shown in Figure 52 was developed to collect ground truth. The GUI was better suited for collecting the necessary ground truth than off-the-shelf data annotation software since it was designed specifically for this application. The user first enters the base path

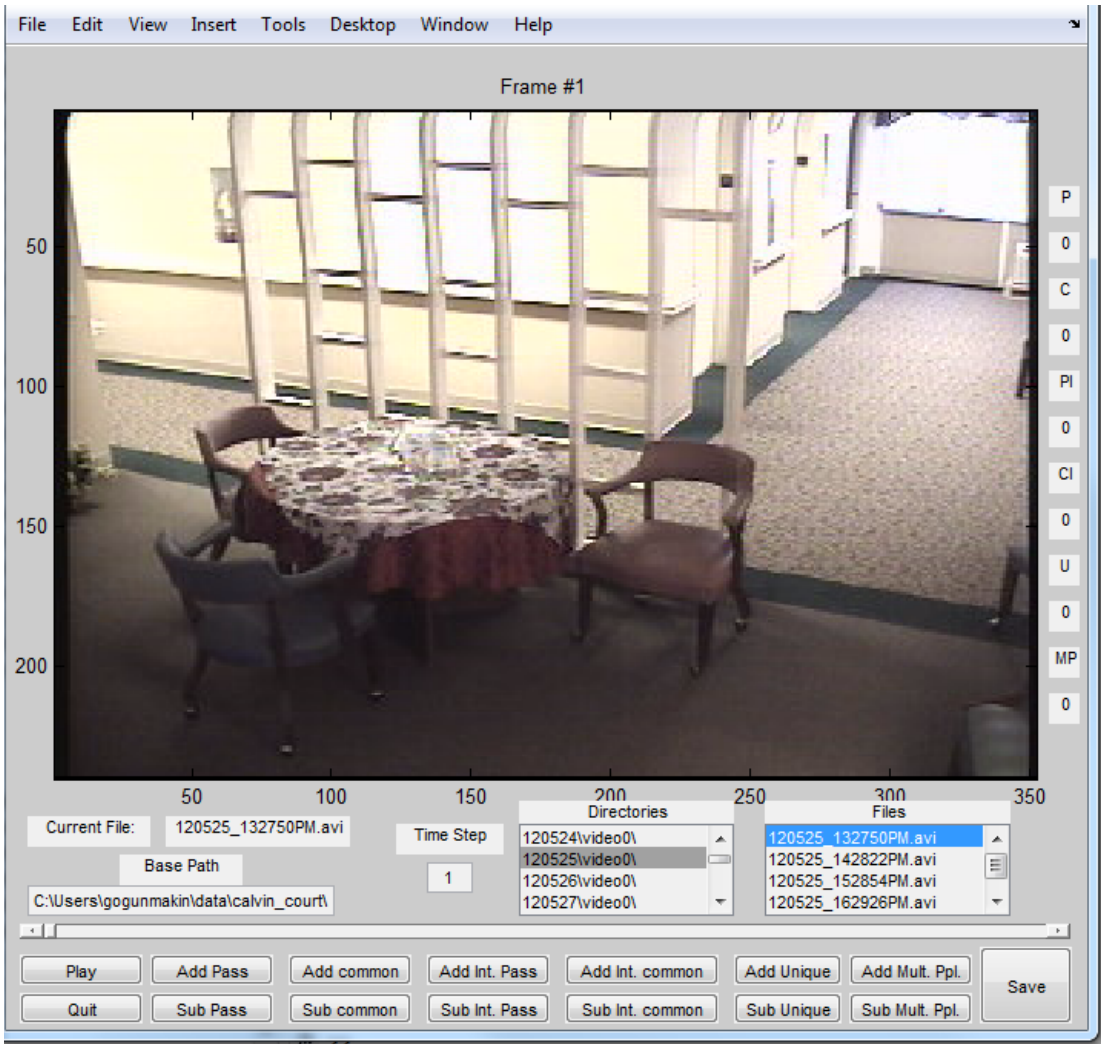


Figure 52: Simple ground truth GUI

where the video data is stored, whereupon the list of directories in the base path and the videos in each directory are displayed in the directories and files list boxes. The user can then select which file to process (each hour of video was placed in a file). The GUI collects 6 types of information: the number of people who pass the common area without entering, the number of people who enter the common area, the number of people who interact inside of the common area, the number of people who interact outside of the common area, the number of people that are in the scene, and the number of unique people who enter the scene. "Add pass" increments the pass without entering the common area counter, "Add common" increments the entered

the common area counter, "Add Int. Pass" increments the interaction while passing the common area counter, "Add Unique" increments the unique people who have entered the scene counter, and "Add Mult. Ppl." increments the counter when there are multiple people in the scene. In addition to incrementing the counters, the frame numbers are also recorded. If the user mistakenly increments the counter, they can adjust it by clicking on the subtract buttons. The time step can be changed to speed up the video when there are periods of inactivity.

A second GUI, shown in Figure 53, was also designed to review and correct the automated system's results. It shows the user all the different visualization options available for the data sequence. Additionally, it allows the user to fix errors due to the re-identification module by merging or splitting the trajectories of a target given its history. Given that contemporary visual processing and surveillance algorithms cannot achieve 100% accuracy, the ability to correct false information is essential. The user is able to enter the base path containing the folders for each day processed. A list of folders in the base path is loaded under the directories listbox, and whenever a folder is clicked, the videos for each hour of the day in the folder are displayed under the files listbox. Once a file is selected, it can be loaded using the "Load Video" button. The current file loaded is displayed right next the "Current File" textbox. The user then enters the name of the file where the automated results are stored and loads it using the "Load Results" button to get started. Once the results are loaded, the "Targets List" listbox is populated with the targets detected throughout the day and the trajectories for all the targets are plotted. The time is also marked in order to know how long it takes to correct each days results. When a target is selected, the video in which it was first detected is loaded, the current file is updated so that the user knows which file is currently being viewed, the slider is updated to the entrance point so that the user can hit the Play button to see how the target traverses the scene, its trajectory is plotted, and its entrance and exit frames into

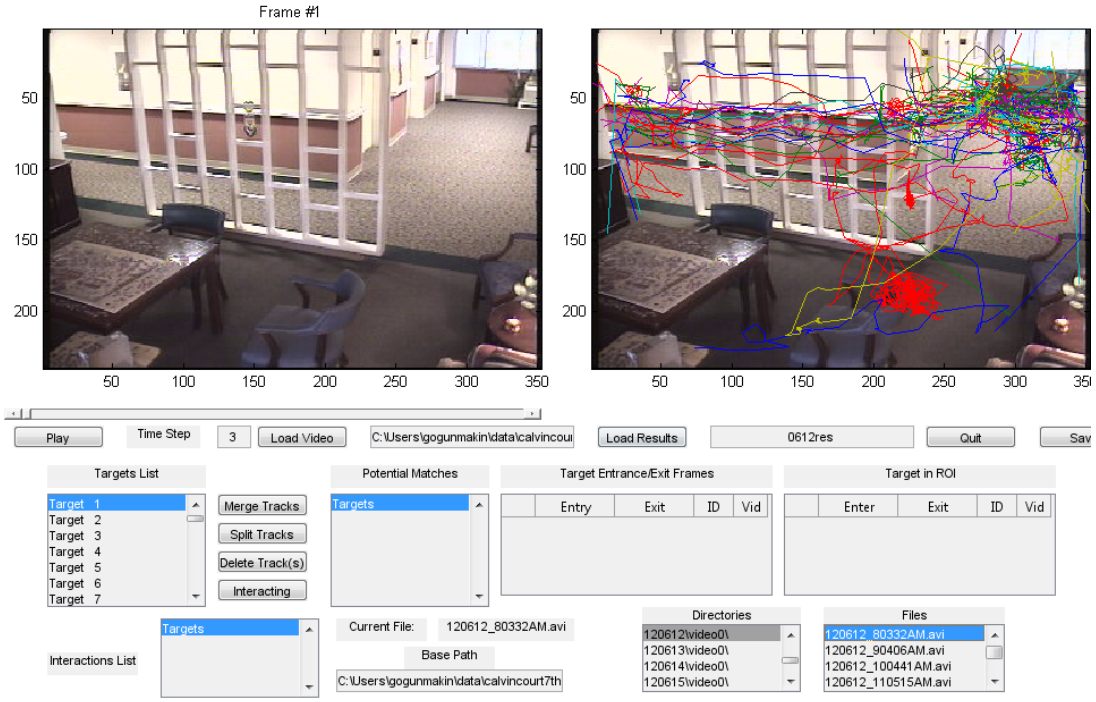


Figure 53: Visualization/Correction GUI

the scene and region of interest, along with its interactions list are displayed. Other targets that potentially match the selected targets are also displayed in the potential matches listbox. This allows the user to merge disjointed tracks of the same target.

The "Merge Tracks" button allows the user to merge two trajectory segments together, the "Split Tracks" button allows the user to remove an inconsistent trajectory segment from the targets history. The "Delete Track(s)" button allows the user to delete a track in case of a false positive. The "Interacting" button allows the user to declare an interaction between different targets that was not identified by the automated system. The "Interactions List" listbox shows the targets that the selected target interacted with, and it allows the user to delete an erroneous interaction. The updated results are saved when the user selects a different days folder, or when the user clicks the "Save" or "Quit" button.

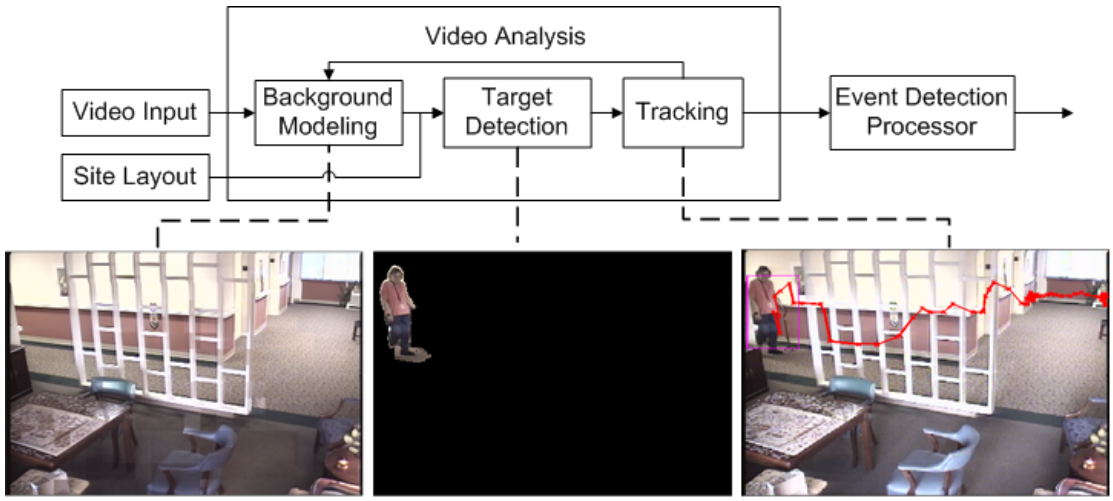


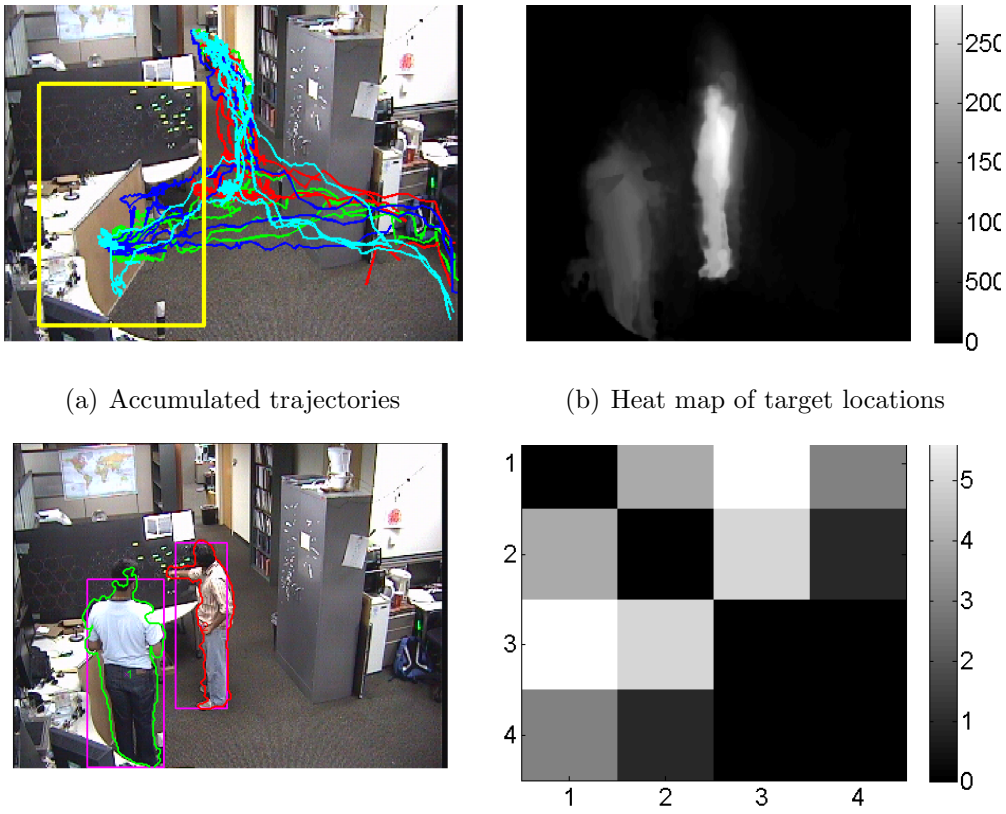
Figure 54: Process flow for the automatic surveillance system.

4.4 Experimental Results

4.4.1 Video Sequence 1

The video test sequence contained 10,800 frames with students walking in and out and observing the technology or interacting with each other. Four targets entered and exited the scene a total of 18 times. The system was able to track these targets throughout the scene. Figure 55(a) shows the trajectories of the targets through the space. The system was able to detect when foreground targets entered the scene but the re-identification module could not always correctly identify a previously seen target. Of the 18 entrances, the system correctly identified 5 as previously seen targets and incorrectly instantiated 9 new trackers. The post-processing step corrected for these errors. For two of the targets, the trajectories simply needed to be merged to result in two extended trajectories from 7 disjoint trajectory segments. For the other two targets, the resemblance required some disambiguation of the target trajectories. The graphical interface simplified the correction step. The interaction processor detected that the technology intervention was being observed in 7,172 frames. It also detected 19 instances of interactions amongst the targets for a total of 2,676 frames. Figure 55(b) shows the heat map for the experiment. The whiter regions show where

the targets spent most of their time. Figure 55(c) shows a sample tracked frame with the targets bounding boxes and segmentations. Figure 55(d) shows the interaction matrix for this test sequence. The interaction matrix shows which targets interacted with each other, and how often. Targets 1 and 3 interacted the most, while targets 3 and 4 did not interact at all. Figure 56 and 57 shows the states versus time of the art piece and a target. The states indicate which type of activity the target was engaging in, as well as the type of attention the art piece/region of interest was receiving.



(a) Accumulated trajectories

(b) Heat map of target locations

(c) Sample frame with tracking results

(d) Interaction matrix

Figure 55: Sample Outputs from sequence 1

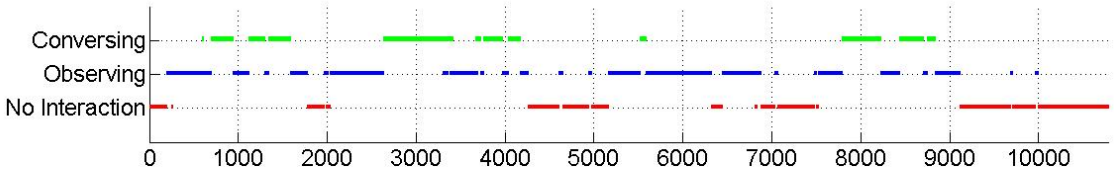


Figure 56: Art Piece States

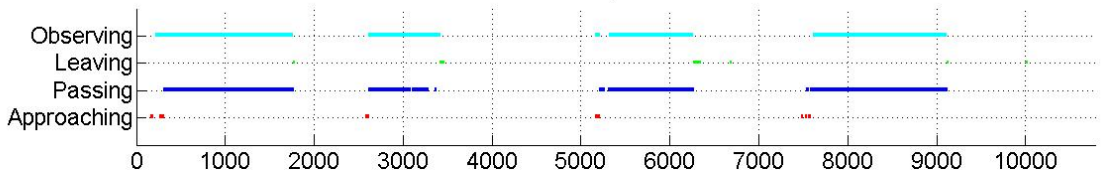


Figure 57: Target 1 States

4.4.2 Video Sequence 2

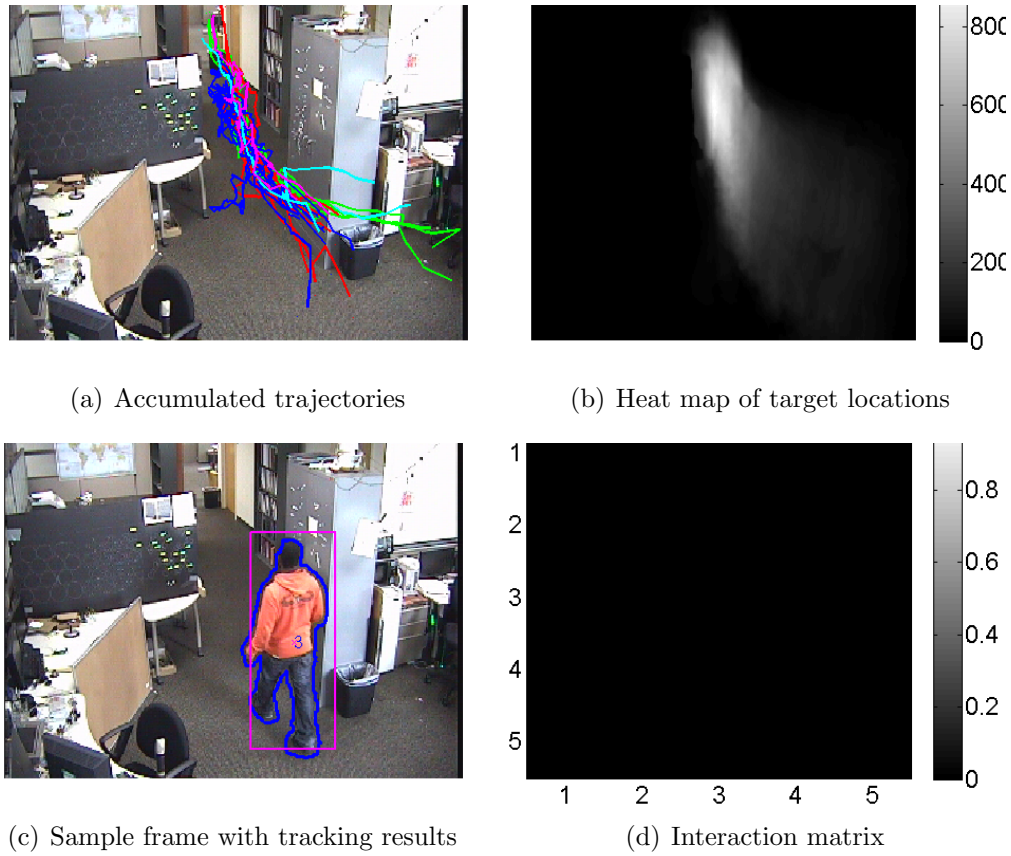
A second test sequence consisting of 23,727 frames with users walking in and out without observing the technology nor interacting with each other. Five targets entered and exited the scene a total of 21 times. The system identified six targets in total, meaning that two trajectories required merging. Figure 58(a) shows the trajectories of the targets through the space. Figure 58(b) shows the resulting heat map which, in contrast to Sequence 1, indicates that little time was spent near the art piece. Figure 58(c) depicts a sample frame and the tracking result, while the interaction matrix in Figure 58(d) correctly indicates that no interactions occurred.

4.4.3 Calvin Court

The semi-automated system was used to process 10 days for each floor. The automated video analysis system is first utilized via a script containing the days and videos to process, and the layout information of the floor. A new background model is initialized at the beginning of each day, and updated as the rest of the day is processed. The targets' state at every frame is stored for use by the event detection processor. When all the videos in a particular day are processed, all the models are cleared, and the system processes the next day. After processing all the days in the script, the results are saved for the user to load and correct using the correction GUI.

4.4.3.1 Floor 2

The results of the experiment for Floor 2 is depicted in Figure 59. The statistics associated to this floor are depicted in Tables 5 and 6. The estimated total number



(a) Accumulated trajectories

(b) Heat map of target locations

(c) Sample frame with tracking results

(d) Interaction matrix

Figure 58: Sample Outputs from sequence 2

of traffic in Week 1 was 425, compared to the ground truth of 372, gives an error of 14.25%. The estimated total number of traffic in Week 2 was 383, compared to the ground truth of 321, gives an error of 19.35%. The estimated total number of interactions in Week 1 was 74, compared to the ground truth of 61, gives an error of 21.31%. The estimated total number of interactions in Week 2 was 45, compared to the ground truth of 35, gives an error of 28.57%. The estimated total number of interactions is 17.41% and 11.75% of the total traffic for week 1 and week 2. Compared to the ground truth of 16.4% and 10.9% of the total traffic for week 1 and week 2 gives an error of 6.18% and 7.76%.

The estimated number of times people passed through the scene without entering the common area for week one was 370 for week 1 and 358 for week 2. Compared to the ground truth of 330 and 307, the errors were 12.12% and 16.61%. The estimated

number of times people entered common area for week one was 55 for week 1 and 25 for week 2. Compared to the ground truth of 42 and 14, the errors were 30.95% and 78.57%. The estimated total number of traffic in the common area is 12.94% and 6.5%of the total traffic for week 1 and week 2. Compared to the ground truth of 11.29% and 4.3% for week 1 and week 2, this gives an error of 14% and 51%.

Table 5: Total daily traffic and interactions on the 2^{nd} floor.

	Total Daily Traffic				Total Daily Interactions			
	Week 1		Week 2		Week 1		Week 2	
	GT	Est.	GT	Est.	GT	Est.	GT	Est.
M	88	99	39	50	14	16	2	4
T	93	101	62	75	18	20	6	7
W	95	107	72	88	23	27	11	14
Th	52	67	74	87	4	7	4	6
F	44	51	74	83	2	4	12	14
Total	372	425	321	383	61	74	35	45
Error	14.25%		19.35%		21.31%		28.57%	

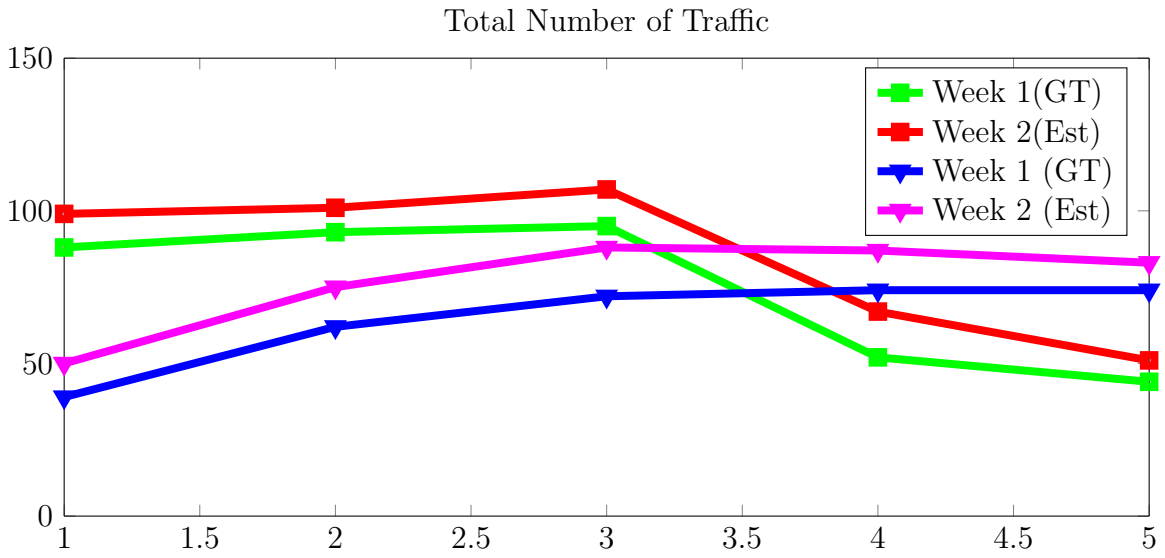


Figure 59: Total traffic on the 2nd Floor

Table 6: Total number of common area entrances and passes on the 2^{nd} floor.

	Passed w/o Entering CA				Entered CA			
	Baseline		Technology		Baseline		Technology	
	GT	Est.	GT	Est.	GT	Est.	GT	Est.
M	67	75	35	44	21	24	4	6
T	82	89	60	70	11	12	2	5
W	93	102	70	83	2	5	2	5
Th	50	61	70	81	2	6	4	6
F	38	43	72	80	6	8	2	3
Total	330	370	307	358	42	55	14	25
Error	12.12%		16.61%		30.95%		78.57%	

4.4.3.2 Floor 3

The results of the experiment for Floor 3 is depicted in Figures 60. The statistics associated to this floor are depicted in Tables 7 and 8. Week 1 was the baseline week and week 2 was the technology intervention week. The estimated total number of traffic in Week 1 was 402, compared to the ground truth of 351, gives an error of 14.53%. The estimated total number of traffic in Week 2 was 313, compared to the ground truth of 294, gives an error of 6.46%. The estimated total number of interactions in Week 1 was 76, compared to the ground truth of 62, gives an error of 22.58%. The estimated total number of interactions in Week 2 was 37, compared to the ground truth of 28, gives an error of 32.14%. The estimated total number of interactions is 18.91% and 11.82% of the total traffic for week 1 and week 2. Compared to the ground truth of 17.66% and 9.52% of the total traffic for week 1 and week 2 gives an error of 7.08% and 24.16%. It should be noted that there was a medical emergency on Tuesday for the baseline condition of floor 3 which increased the number of total traffic and interactions. Removing the anomaly, the estimated total number of interactions becomes 14.74% of the total traffic and the ground truth becomes 13.15%. This still shows that there was a decrease in the number of interactions on the third floor.

The estimated number of times people passed through the scene without entering

the common area for week one was 359 for week 1 and 275 for week 2. Compared to the ground truth of 319 and 261, the errors were 12.54% and 5.36%. The estimated number of times people entered common area for week one was 43 for week 1 and 38 for week 2. Compared to the ground truth of 32 and 33, the errors were 34.38% and 15.15%. The estimated total number of traffic in the common area is 10.7% and 12.14% of the total traffic for week 1 and week 2. Compared to the ground truth of 9.12% and 11.22% for week 1 and week 2, this gives an error of 17.32% and 8.2%. As mentioned earlier, there was a medical emergency which inflated the number of people who entered the common area. If this anomaly is removed, the estimated total number of traffic in the common area is 6.38% of the total traffic and the ground truth becomes 4.38%. This shows that although the percentage of interactions went down during the technology intervention, the number of people who visited the common area increased.

Table 7: Total daily traffic and interactions on the 3^{rd} floor.

	Total Daily Traffic				Total Daily Interactions			
	Baseline		Technology		Baseline		Technology	
	GT	Est.	GT	Est.	GT	Est.	GT	Est.
M	47	55	44	50	13	15	0	2
T	100	117	63	66	29	34	10	12
W	74	79	61	66	6	8	2	5
Th	90	107	55	57	8	12	2	3
F	40	44	71	74	6	7	14	15
	351	402	294	313	62	76	28	37
Error	14.53%		6.46%		22.58%		32.14%	

4.4.3.3 Floor 7

The results of the experiment for Floor 7 is depicted in Figures 61. The statistics associated to this floor are depicted in Tables 9 and 10. As with floor 3, week 1 was the baseline week and week 2 was the technology intervention week. The estimated total number of traffic in Week 1 was 459, compared to the ground truth of 415,

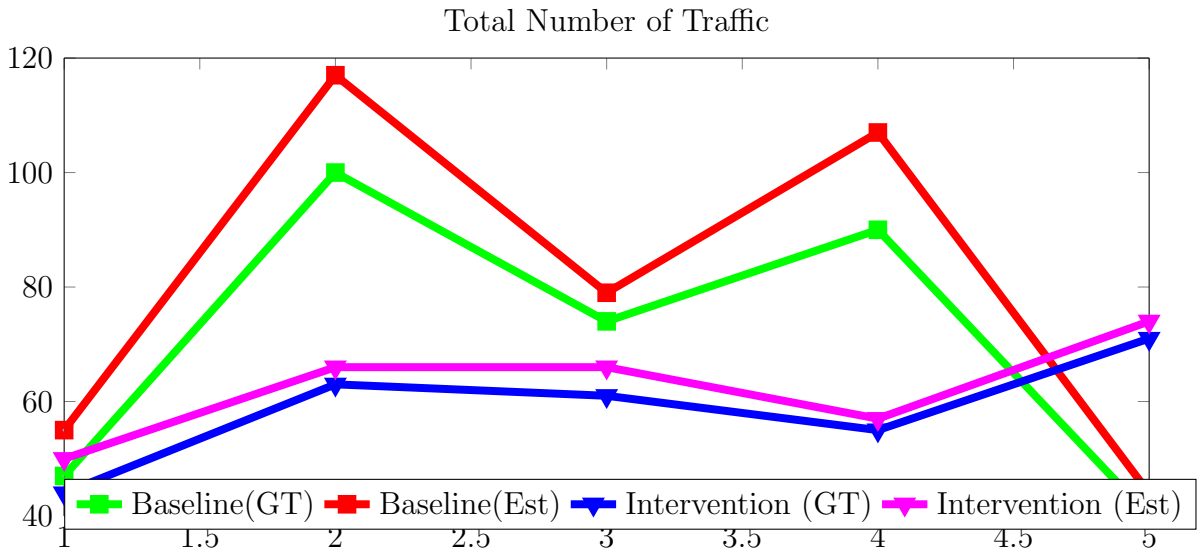


Figure 60: Total traffic on the 3rd Floor

Table 8: Total number of common area entrances and passes on the 3^{rd} floor.

	Passed w/o Entering CA				Entered CA			
	Baseline		Technology		Baseline		Technology	
	GT	Est.	GT	Est.	GT	Est.	GT	Est.
M	42	48	34	37	5	7	10	13
T	79	92	56	59	21	25	7	7
W	71	75	55	59	3	4	6	7
Th	90	104	50	52	0	3	5	5
F	37	40	66	68	3	4	5	6
	319	359	261	275	32	43	33	38
Error	12.54%		5.36%		34.38%		15.15%	

gives an error of 10.6%. The estimated total number of traffic in Week 2 was 541, compared to the ground truth of 505, gives an error of 7.13%. The estimated total number of interactions in Week 1 was 59, compared to the ground truth of 48, gives an error of 22.92%. The estimated total number of interactions in Week 2 was 118, compared to the ground truth of 100, gives an error of 18%. The estimated total number of interactions is 12.85% and 21.81% of the total traffic for week 1 and week 2. Compared to the ground truth of 11.57% and 19.8% of the total traffic for week

1 and week 2 gives an error of 11.06% and 10.1%. This shows that the intervention had the desired effect on this floor.

The estimated number of times people passed through the scene without entering the common area for week one was 407 for week 1 and 469 for week 2. Compared to the ground truth of 371 and 443, the errors were 9.7% and 5.54%. The estimated number of times people entered common area for week one was 52 for week 1 and 72 for week 2. Compared to the ground truth of 44 and 62, the errors were 18.18% and 16.13%. The estimated total number of traffic in the common area is 12.78% and 15.35% of the total traffic for week 1 and week 2. Compared to the ground truth of 11.86% and 14% for week 1 and week 2, this gives an error of 7.76% and 9.6%. This shows that the number of people who visited the common area increased.

Table 9: Total daily traffic and interactions on the 7^{th} floor.

	Total Daily Traffic				Total Daily Interactions			
	Baseline		Technology		Baseline		Technology	
	GT	Est.	GT	Est.	GT	Est.	GT	Est.
M	80	92	114	121	9	12	24	28
T	101	112	120	125	8	10	37	40
W	92	104	103	112	13	16	17	21
Th	62	67	80	88	8	10	8	11
F	80	84	88	95	10	11	14	18
	415	459	505	541	48	59	100	118
Error	10.6%		7.13%		22.92%		18.00%	

Table 10: Total number of common area entrances and passes on the 7^{th} floor.

	Passed w/o Entering CA				Entered CA			
	Baseline		Technology		Baseline		Technology	
	GT	Est.	GT	Est.	GT	Est.	GT	Est.
M	69	78	104	107	11	14	10	14
T	97	106	101	106	4	6	19	19
W	84	94	92	97	8	10	11	15
Th	58	62	70	77	4	5	10	11
F	63	67	76	82	17	17	12	13
	371	407	443	469	44	52	62	72
Error	9.7%		5.54		18.18%		16.13	

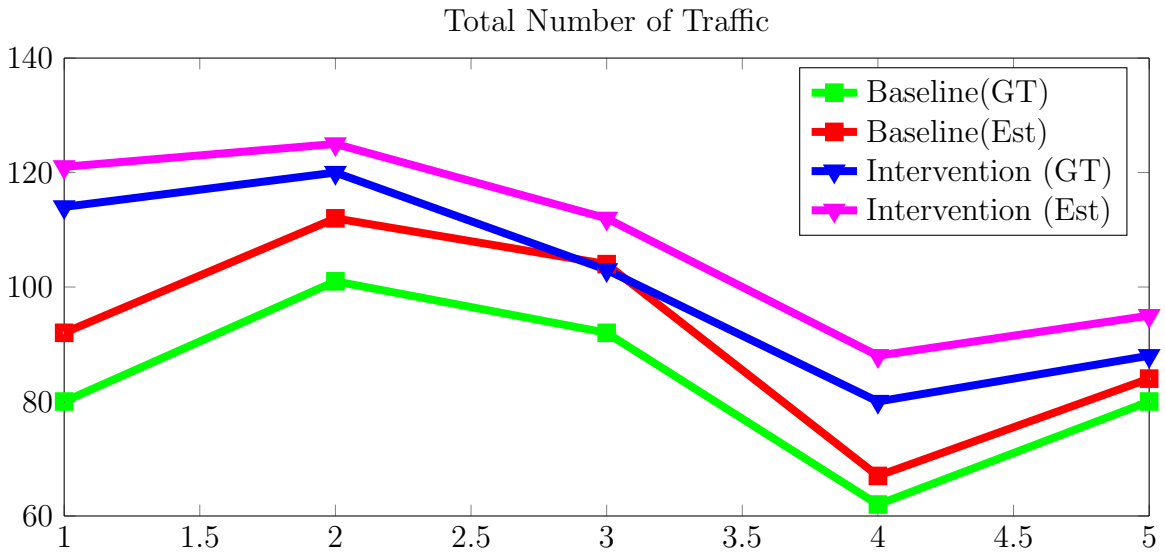


Figure 61: Total traffic on the 7th Floor

4.4.4 User Interface Correction

Table 11 shows the time it took to generate the event statistics using the custom GUI, Noldus Observer XT, and the Semi-Automated tool. The custom GUI time was computed by taking an average of the time it took to manually generate the ground truth results in Tables 5,7, and 9. The Observer time was computed by taking an average of the time it took to process 8 hours of the video using the Observer XT program. The semi-automated tool time shows the average time it takes to process the video automatically, and the time it takes for the user corrections. As hypothesized, the time required for the user to generate statistics using the semi-automated tool is significantly less than manual processing. This shows that the combining the automated system with the correction GUI would save a lot of manpower when used to process weeks worth of information. The semi-automated system also has the additional capabilities of letting the user view each target's trajectory history. Manually selecting the location in each frame to get this information would have been significantly time consuming.

Table 11: Average time it takes to process 60 minutes of video.

Custom GUI	Noldus Observer XT	Semi Automated	
		Surveillance System	User Corrections
22.62 minutes	73.8 minutes	65.7 minutes	5.8 minutes

4.5 Discussion

The automated system detected and tracked 2523 target entries across the whole experiment, which is a 11.8% more than the 2258 entries recorded in the ground truth. The correction tool was used to correct the errors generated by the automated system to generate results with 100% accuracy. The automated system usually generated more false positives as a result of sudden illumination changes. Although the targets have their designated entrance regions, there are windows and light switches located in the common area. Whenever the illumination changes because of sudden cloud movement or someone turning on/off the light switch, the system sometimes erroneously detects and starts tracking a foreground object in the entrance region. This kind of error was corrected using the delete tracks button in the correction GUI. There were also some errors regarding when the targets left the scene. Sometimes, the system is not able to determine that the target has left the scene since it relies on the detection algorithm which sometimes produces false positives to due to the sudden illumination change not being incorporated into the model. This in turns creates false positives when it comes to detecting the number of interactions taking place. To fix this error, the split tracks button was used to split a targets track at the moment it leaves the scene, the excess trajectory in the split track is then deleted.

This chapter presented a tool for semi-automated coding of social interactions from recorded video. The purpose of this chapter was to show that the developed tools were useful in meeting the goals of determining the effect of technology intervention on socialization. To this end, a ground truth collection GUI was developed to show how the space was utilized before and after the technology intervention. An

automated system was developed and validated by demonstrating that it could detect the traffic and interaction taking place. With regards to the desired output, the automated system was 88.26% correct in detecting and tracking targets thereby effectively requiring annotation of a significantly smaller portion of the overall video. With the system, a coder can process 4x more video than with a custom interface, and 13x more video versus with a general purpose software package, given the same amount of labor input. In particular, one 8-hour period of time can be coded in just over 45 minutes. Future work will be to improve the performance of the automated video analysis system so that its processing time is significantly reduced and its accuracy is increased, as well as to evaluate user interface modifications that could reduce the time spent during the correction step.

CHAPTER V

CONCLUSION

The purpose of this thesis was to demonstrate the use of an automatic surveillance system to generate event statistics which can provide help with decision support for interested parties. The system was used to process six different videos from three different construction sites and 6 weeks of data from a retirement community. The main concern for construction sites is the improvement of productivity. This thesis was able to show that although the system was not 100% accurate, it was still able to generate statistics of interest that would allow an operator to make decisions regarding productivity. For the senior housing, this thesis shows that the system was accurate enough to support more efficient video coding to save time when analyzing the data. The system was able to show how much of an effect the intervention had on interactions between people in the retirement community.

Future work can incorporate more layers into the system that will allow for detection of more complex events. For instance, audio or gaze detection could be added to the system to improve the accuracy of interaction detection in the senior housing setting. Adding the ability to track other machines on the construction site performing different tasks will also provide a more comprehensive analysis of the site and allow project managers to make well-informed decisions regarding their next course of action.

REFERENCES

[1] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher, "Appearance-based person reidentification in camera networks: problem overview and current approaches," *Journal of Ambient Intelligence and Humanized Computing*, vol. 2, no. 2, pp. 127–151, 2011.

[2] A. Baumann, M. Boltz, J. Ebling, M. Koenig, H. S. Loos, M. Merkel, W. Niem, J. K. Warzelhan, and J. Yu, "A review and comparison of measures for automatic video surveillance systems," *EURASIP Journal on Image and Video Processing*, 2008.

[3] H. M. Dee and S. A. Velastin, "How close are we to solving the problem of automated visual surveillance?," *Machine Vision and Applications*, vol. 19, no. 5-6, pp. 329–343, 2008.

[4] O. Javed and M. Shah, "Automated video surveillance," in *Automated Multi-Camera Surveillance: Algorithms and Practice*, pp. 1–9, Springer, 2008.

[5] T. Ko, "A survey on behavior analysis in video surveillance for homeland security applications," in *Applied Imagery Pattern Recognition Workshop, 2008. AIPR'08. 37th IEEE*, pp. 1–8, IEEE, 2008.

[6] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, 2013.

[7] A. Taha, H. H Zayed, M. E Khalifa, and E.-S. M El-Horbaty, "Exploring behavior analysis in video surveillance applications," *International Journal of Computer Applications*, vol. 93, no. 14, pp. 22–32, 2014.

[8] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *The Visual Computer*, vol. 29, no. 10, pp. 983–1009, 2013.

[9] C. C. Loy, *Activity understanding and unusual event detection in surveillance videos*. PhD thesis, 2010.

[10] S. A. Assaf and S. Al-Hejji, "Causes of delay in large construction projects," *International Journal of Project Management*, vol. 24, no. 4, pp. 349–357, 2006.

[11] E. Koehn, R. Young, J. Kuchar, and F. Seling, "Cost of delays in construction," *Journal of the Construction Division*, vol. 104, no. 3, pp. 323–331, 1978.

[12] J. Bohn and J. Teizer, "Benefits and barriers of construction project monitoring using high-resolution automated cameras," *ASCE Journal of Construction Engineering and Management*, vol. 136, no. 6, pp. 632–640, 2010.

[13] A. C. Nazare, C. E. dos Santos, R. Ferreira, and W. R. Schwartz, "Smart surveillance framework: A versatile tool for video analysis," in *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pp. 753–760, IEEE, 2014.

[14] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, *et al.*, *A system for video surveillance and monitoring*. Carnegie Mellon University, the Robotics Institute Pittsburg, 2000.

[15] I. Haritaoglu, D. Harwood, and L. Davis, "W4: Real-time surveillance of people and their activities," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 809–830, 2000.

[16] O. Javed, Z. Rasheed, O. Alatas, and M. Shah, "Knight/spl trade: a real time surveillance system for multiple and non-overlapping cameras," in *icme*, pp. 649–652, IEEE, 2003.

[17] Y. Tian, L. Brown, A. Hampapur, M. Lu, A. Senior, and C. Shu, "Ibm smart surveillance system (s3): event based video surveillance system with an open and extensible framework," *Machine Vision and Applications*, vol. 19, no. 5, pp. 315–327, 2008.

[18] T. Bouwmans, "Traditional and recent approaches in background modeling for foreground detection: An overview," *Computer Science Review*, vol. 11, pp. 31–66, 2014.

[19] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Second IEEE Workshop on Visual Surveillance*, pp. 246–252, June 1999.

[20] T. Bouwmans, F. El Baf, and B. Vachon, "Background modeling using mixture of gaussians for foreground detection-a survey," *Recent Patents on Computer Science*, vol. 1, no. 3, pp. 219–237, 2008.

[21] A. Sobral and A. Vacavant, "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos," *Computer Vision and Image Understanding*, vol. 122, pp. 4–21, 2014.

[22] T. Bouwmans, "Subspace learning for background modeling: A survey," *Recent Patents on Computer Science*, vol. 2, no. 3, pp. 223–234, 2009.

[23] N. M. Oliver, B. Rosario, and A. P. Pentland, "A bayesian computer vision system for modeling human interactions," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 22, no. 8, pp. 831–843, 2000.

[24] R. Hsu, M. Abdel-Mottaleb, and A. Jain, "Face detection in color images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 696–706, 2002.

[25] E. Corvee, S. Bak, and F. Bremond, "People detection and re-identification for multi surveillance cameras," in *VISAPP-International Conference on Computer Vision Theory and Applications-2012*, 2012.

[26] M. Nixon, T. Tan, and R. Chellappa, *Human identification based on gait*. Springer-Verlag New York Inc, 2006.

[27] F. Pala, R. Satta, G. Fumera, and F. Roli, "Multi-modal person re-identification using rgb-d cameras,"

[28] M. Ahmed, N. Al-Jawad, and A. Sabir, "Gait recognition based on kinect sensor," in *SPIE Photonics Europe*, pp. 91390B–91390B, International Society for Optics and Photonics, 2014.

[29] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person re-identification*. Springer, 2014.

[30] J. Yang, Z. Shi, and P. Vela, "Person reidentification by kernel pca based appearance learning," in *Computer and Robot Vision (CRV), 2011 Canadian Conference on*, pp. 227–233, IEEE, 2011.

[31] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *Acm computing surveys (CSUR)*, vol. 38, no. 4, p. 13, 2006.

[32] B. Yang and R. Nevatia, "An online learned crf model for multi-target tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2034–2041, IEEE, 2012.

[33] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, pp. 603–619, 2002.

[34] G. HAGER, M. DEWAN, and C. STEWART, "Multiple kernel tracking with ssd," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[35] B. Bascle and R. Deriche, "Region tracking through image sequences," in *Computer Vision, 1995. Proceedings., Fifth International Conference on*, pp. 302–307, IEEE, 1994.

[36] O. Arif and P. Vela, "Robust density comparison for visual tracking," in *British Machine Vision Conference*, 2009.

[37] Y. Cheng, "Mean shift, mode seeking, and clustering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 8, pp. 790–799, 1995.

[38] S. Birchfield and S. Rangarajan, "Spatiograms versus histograms for region-based tracking," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Volume 2-Volume 02*, pp. 1158–1163, IEEE Computer Society, 2005.

[39] C. Zhao, A. Knight, and I. Reid, "Target tracking using mean-shift and affine structure," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pp. 1–5, IEEE, 2008.

[40] P. Chockalingam, N. Pradeep, and S. Birchfield, "Adaptive fragments-based tracking of non-rigid objects using level sets," in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 1530–1537, IEEE, 2009.

[41] H. Uemura, S. Ishikawa, and K. Mikolajczyk, "Feature tracking and motion compensation for action recognition.," in *BMVC*, pp. 1–10, 2008.

[42] B. Yang and R. Nevatia, "Multi-target tracking by online learning a crf model of appearance and motion patterns," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 203–217, 2014.

[43] J. Canny, "A computational approach to edge detection," *Readings in computer vision: issues, problems, principles, and paradigms*, vol. 184, no. 87-116, p. 86, 1987.

[44] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the 4th Alvey Vision Conference*, vol. 15, pp. 147–151, 1988.

[45] S. Birchfield and S. Pundlik, "Joint tracking of features and edges," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–6, IEEE, 2008.

[46] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, 1997.

[47] T. Lindeberg, "Feature detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp. 79–116, 1998.

[48] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *International journal of computer vision*, vol. 60, no. 1, pp. 63–86, 2004.

[49] S. Tran and L. Davis, "Robust object tracking with regional affine invariant features," in *Proc. ICCV*, Citeseer, 2007.

[50] T. F. Chan, B. Y. Sandberg, and L. A. Vese, "Active contours without edges for vector-valued images," *Journal of Visual Communication and Image Representation*, vol. 11, pp. 130–141, 2000.

[51] S.-H. Bae and K.-J. Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 1218–1225, IEEE, 2014.

[52] J. Xing, H. Ai, and S. Lao, "Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1200–1207, IEEE, 2009.

[53] D. Zhai and P. Goodrum, "Relationship between automation and integration of construction information systems and labor productivity," *ASCE Journal of Construction Engineering Management*, vol. 135, no. 8, pp. 746–753, 2009.

[54] P. M. Goodrum, D. Zhai, and M. F. Yasin, "Relationship between Changes in Material Technology and Construction Productivity," *ASCE Journal of Construction Engineering and Management*, vol. 135, no. 4, pp. 278–287, 2009.

[55] M. C. Gouett, C. T. Haas, P. M. Goodrum, and C. H. Caldas, "Activity analysis for direct-work rate improvement in construction," *Journal of Construction Engineering and Management*, vol. 137, no. 12, pp. 1117–1124, 2011.

[56] D. L. Orth, S. Welty, and J. J. Jenkins, "Analyzing labor productivity through work sampling," in *ASC Proceedings of the 42nd Annual conference Colorado State University fort callings Colorado April*, pp. 20–22, 2006.

[57] J. Gong and C. Caldas, "An intelligent video computing method for automated productivity analysis of cyclic construction operations," in *ASCE International Workshop on Computing in Civil Engineering*, pp. 64–73, ASCE, 2009.

[58] C. H. Caldas, D. G. Torrent, and C. T. Haas, "Using global positioning system to improve materials-locating processes on industrial projects," *Journal of Construction Engineering and Management*, vol. 132, no. 7, pp. 741–749, 2006.

[59] N. Pradhananga and J. Teizer, "Automatic spatio-temporal analysis of construction site equipment operations using gps data," *Automation in Construction*, vol. 29, pp. 107–122, 2013.

[60] D. Grau, C. H. Caldas, C. T. Haas, P. M. Goodrum, and J. Gong, "Assessing the impact of materials tracking technologies on construction craft productivity," *Automation in Construction*, vol. 18, no. 7, pp. 903–911, 2009.

[61] A. Vasenev, N. Pradhananga, F. Bijleveld, D. Ionita, T. Hartmann, J. Teizer, and A. Dorée, "An information fusion approach for filtering GNSS data sets collected during construction operations," *Advanced Engineering Informatics*, vol. 28, no. 4, pp. 297–310, 2014.

[62] F. Vahdatikhaki and A. Hammad, "Framework for near real-time simulation of earthmoving projects using location tracking technologies," *Automation in Construction*, vol. 42, pp. 50–67, 2014.

[63] J. Song, C. T. Haas, and C. H. Caldas, "Tracking the location of materials on construction job sites," *Journal of Construction Engineering and Management*, vol. 132, no. 9, pp. 911–918, 2006.

[64] S. El-Omari and O. Moselhi, "Integrating automated data acquisition technologies for progress reporting of construction projects," *Automation in Construction*, vol. 20, no. 6, pp. 699–705, 2011.

[65] A. Costin, N. Pradhananga, and J. Teizer, "Leveraging passive RFID technology for construction resource field mobility and status monitoring in a high-rise renovation project," *Automation in Construction*, vol. 24, pp. 1–15, 2012.

[66] J. Teizer, D. Lao, and M. Sofer, "Rapid automated monitoring of construction site activities using ultra-wideband," in *Proceedings of the 24th International Symposium on Automation and Robotics in Construction, Kochi, Kerala, India*, pp. 19–21, 2007.

[67] T. Cheng, M. Venugopal, J. Teizer, and P. Vela, "Performance evaluation of ultra wideband technology for construction resource location tracking in harsh environments," *Automation in Construction*, vol. 20, no. 8, pp. 1173–1184, 2011.

[68] T. Cheng, J. Teizer, G. C. Migliaccio, and U. C. Gatti, "Automated task-level activity analysis through fusion of real time location sensors and worker's thoracic posture data," *Automation in Construction*, vol. 29, pp. 24–39, 2013.

[69] J. Teizer, C. Kim, C. T. Haas, K. A. Liapi, and C. H. Caldas, "Framework for real-time three-dimensional modeling of infrastructure," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1913, no. 1, pp. 177–186, 2005.

[70] J. Teizer and P. Vela, "Personnel tracking on construction sites using video cameras," *Advanced Engineering Informatics*, vol. 23, pp. 452–462, Oct. 2009.

[71] E. Rezazadeh Azar and B. McCabe, "Automated visual recognition of dump trucks in construction videos," *Journal of Computing in Civil Engineering*, vol. 26, no. 6, pp. 769–781, 2011.

[72] M. Memarzadeh, A. Heydarian, M. Golparvar-Fard, and J. Niebles, "Real-time and automated recognition and 2d tracking of construction workers and equipment from site video streams," in *ASCE International Workshop on Computing in Civil Engineering*, 2012.

[73] M.-W. Park, G. Jog, and I. Brilakis, "Initializing vision based trackers using semantic texton forests," in *International Symposium on Automation and Robotics in Construction*, 2011.

[74] E. Rezazadeh Azar and B. McCabe, "Part based model and spatial–temporal reasoning to recognize hydraulic excavators in construction images and videos," *Automation in construction*, vol. 24, pp. 194–202, 2012.

[75] S. Chi and C. H. Caldas, "Automated object identification using optical video cameras on construction sites," *Computer-Aided Civil and Infrastructure Engineering*, vol. 26, no. 5, pp. 368–380, 2011.

[76] J. Yang, P. Vela, J. Teizer, and Z. Shi, "Vision-based crane tracking for understanding construction activity," *ASCE International Workshop on Computing in Civil Engineering*, pp. 258–265, 2011.

[77] J. Gong and C. H. Caldas, "An object recognition, tracking, and contextual reasoning-based video interpretation method for rapid productivity analysis of construction operations," *Automation in Construction*, vol. 20, no. 8, pp. 1211–1226, 2011.

[78] M.-W. Park, A. Makhmalbaf, and I. Brilakis, "Comparative study of vision tracking methods for tracking of construction site resources," *Automation in Construction*, vol. 20, no. 7, pp. 905–915, 2011.

[79] M.-W. Park and I. Brilakis, "Enhancement of construction equipment detection in video frames by combining with tracking," in *ASCE International Workshop on Computing in Civil Engineering*, pp. 421–428, 2012.

[80] J. Yang, O. Arif, P. Vela, J. Teizer, and Z. Shi, "Tracking multiple workers on construction sites using video cameras," *Advanced Engineering Informatics*, vol. 24, no. 4, pp. 428–434, 2010.

[81] G. Ogunmakin, J. Teizer, and P. Vela, "Quantifying Interactions Amongst Construction Site Machines," in *Proc. of the EG-ICE Workshop on Intelligent Computing in Engineering, Vienna, Austria*, July 2013.

[82] I. Brilakis, M.-W. Park, and G. Jog, "Automated vision tracking of project related entities," *Advanced Engineering Informatics*, vol. 25, no. 4, pp. 713–724, 2011.

[83] D. Ake, "Laser receiver and angle sensor mounted on an excavator, Patent US 6,263,595," 2001.

[84] A. Peddi, L. Huan, Y. Bai, and S. Kim, "Development of human pose analyzing algorithms for the determination of construction productivity in real-time," in *Construction Research Congress*, (Seattle, WA), pp. 11–20, 2009.

[85] T. Weerasinghe and J. Y. Ruwanpura, "Automated multiple objects tracking system (AMOTS)," in *Construction Research Congress*, pp. 11–20, 2010.

[86] A. Khosrowpour, I. Fedorov, A. Holynski, J. C. Niebles, and M. Golparvar-Fard, "Automated Worker Activity Analysis in Indoor Environments for Direct-Work Rate Improvement from Long Sequences of RGB-D Images," in *Construction Research Congress 2014*, pp. 729–738, 2014.

[87] J. Yang, Z. Shib, and Z. Wua, "Automatic Recognition of Construction Worker Activities Using Dense Trajectories," in *International Symposium on Automation and Robotics in Construction and Mining*, 2015.

[88] R. Navon, "Automated project performance control of construction projects," *Automation in Construction*, vol. 14, pp. 467–476, Aug. 2005.

[89] A. D. Fisk, W. A. Rogers, N. Charness, S. J. Czaja, and J. Sharit, *Designing for older adults: Principles and creative human factors approaches*. CRC press, 2009.

[90] T. S. of Certified Senior Advisors, *State of the Senior Housing Industry*. Society of Certified Senior Advisors: Denver, CO, 2011.

[91] A. on Aging, *A Profile of Older Americans: 2011: Living Arrangements*. U.S. Department of Health and Human Services, 2012.

[92] K. Adams, S. Sanders, and E. Auth, "Loneliness and depression in independent living retirement communities: Risk and resilience factors," *Aging & Mental Health*, vol. 8, no. 6, pp. 475–485, 2004.

[93] P. D. St. John and P. R. Montgomery, "Do depressive symptoms predict mortality in older people?," *Aging & mental health*, vol. 13, no. 5, pp. 674–681, 2009.

[94] H. So, H. Kim, and K. Ju, "Prediction model of quality of life in elderly based on icf model," *Journal of Korean Academy of Nursing*, vol. 41, no. 4, pp. 481–490, 2011.

[95] T. O. Obisesan and R. Gillum, "Cognitive function, social integration and mortality in a us national cohort study of older adults," *BMC geriatrics*, vol. 9, no. 1, p. 33, 2009.

[96] D. K. Kiely, S. E. Simon, R. N. Jones, and J. N. Morris, "The protective effect of social engagement on mortality in long-term care," *Journal of the American Geriatrics Society*, vol. 48, no. 11, pp. 1367–1372, 2000.

[97] M. Schwarzbach, M. Luppa, C. Sikorski, A. Fuchs, W. Maier, H. van den Bussche, M. Pentzek, and S. G. Riedel-Heller, "The relationship between social integration and depression in non-demented primary care patients aged 75 years and older," *Journal of affective disorders*, vol. 145, no. 2, pp. 172–178, 2013.

[98] S. Goto, B.-J. Park, Y. Tsunetsugu, K. Herrup, and Y. Miyazaki, "The effect of garden designs on mood and heart output in older adults residing in an assisted living facility," *HERD: Health Environments Research & Design Journal*, vol. 6, no. 2, pp. 27–42, 2013.

[99] L. Van Malderen, T. Mets, and E. Gorus, "Interventions to enhance the quality of life of older people in residential long-term care: a systematic review," *Ageing research reviews*, vol. 12, no. 1, pp. 141–150, 2013.

[100] M. B. Detweiler, T. Sharma, J. G. Detweiler, P. F. Murphy, S. Lane, J. Carman, A. S. Chudhary, M. H. Halling, and K. Y. Kim, "What is the evidence to support the use of therapeutic gardens for the elderly?," *Psychiatry investigation*, vol. 9, no. 2, pp. 100–110, 2012.

[101] E. Wittenberg-Lyles, D. P. Oliver, G. Demiris, and S. Shaunfield, "Benefits and challenges of the passport broadcast intervention in long-term care," *Educational Gerontology*, vol. 38, no. 10, pp. 691–698, 2012.

[102] J. Broekens, M. Heerink, and H. Rosendal, "Assistive social robots in elderly care: a review," *Gerontechnology*, vol. 8, no. 2, pp. 94–103, 2009.

[103] K. Wada and T. Shibata, "Social effects of robot therapy in a care house-change of social network of the residents for two months," in *Robotics and Automation, 2007 IEEE International Conference on*, pp. 1250–1255, IEEE, 2007.

[104] P. Flandorfer, "Population ageing and socially assistive robots for elderly persons: the importance of sociodemographic factors for user acceptance," *International Journal of Population Research*, vol. 2012, 2012.

[105] T. Klamer, S. B. Allouch, and D. Heylen, "adventures of harvey–use, acceptance of and relationship building with a social robot in a domestic environment," in *Human-Robot Personal Relationships*, pp. 74–82, Springer, 2011.

[106] Z. Gabriel and A. Bowling, "Quality of life from the perspectives of older people," *Ageing and Society*, vol. 24, no. 05, pp. 675–691, 2004.

[107] P. Elosua, "Subjective values of quality of life dimensions in elderly people. a sem preference model approach," *Social indicators research*, vol. 104, no. 3, pp. 427–437, 2011.

[108] J. Golden, R. M. Conroy, I. Bruce, A. Denihan, E. Greene, M. Kirby, and B. A. Lawlor, "Loneliness, social support networks, mood and wellbeing in community-dwelling elderly," *International journal of geriatric psychiatry*, vol. 24, no. 7, pp. 694–700, 2009.

[109] S. Zimmerman, C. M. Mitchell, C. K. Chen, L. A. Morgan, A. L. Gruberbaldini, P. D. Sloane, J. K. Eckert, and J. Munn, "An observation of assisted living environments: Space use and behavior," *Journal of gerontological social work*, vol. 49, no. 3, pp. 185–203, 2007.

[110] C. D. Kidd, W. Taggart, and S. Turkle, "A sociable robot to encourage social interaction among the elderly," in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, pp. 3972–3976, IEEE, 2006.

[111] Noldus, "Observer xt.," 2012.

[112] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3273–3280, IEEE, 2011.

[113] D. J. Cook, A. Crandall, G. Singla, and B. Thomas, "Detection of social interaction in smart spaces," *Cybernetics and Systems: An International Journal*, vol. 41, no. 2, pp. 90–104, 2010.

[114] Â. Costa, J. C. Castillo, P. Novais, A. Fernández-Caballero, and R. Simoes, "Sensor-driven agenda for intelligent home care of the elderly," *Expert Systems with Applications*, vol. 39, no. 15, pp. 12192–12204, 2012.

[115] M. A. Hossain and D. T. Ahmed, "Virtual caregiver: an ambient-aware elderly monitoring system," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 16, no. 6, pp. 1024–1031, 2012.

[116] D. Lymberopoulos, A. Bamis, and A. Savvides, "Extracting spatiotemporal human activity patterns in assisted living using a home sensor network," in *Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments*, PETRA '08, (New York, NY, USA), pp. 29:1–29:8, ACM, 2008.

[117] T. Choudhury and A. Pentland, "The sociometer: A wearable device for understanding human networks," in *CSCW'02 Workshop: Ad hoc Communications and Collaboration in Ubiquitous Computing Environments*, 2002.

[118] S. Chernbumroong, S. Cang, A. Atkins, and H. Yu, "Elderly activities recognition and classification for applications in assisted living," *Expert Systems with Applications*, vol. 40, no. 5, pp. 1662–1674, 2013.

[119] C. Wu, A. H. Khalili, and H. Aghajan, "Multiview activity recognition in smart homes with spatio-temporal features," in *Proceedings of the Fourth ACM/IEEE International Conference on Distributed Smart Cameras*, ICDSC '10, (New York, NY, USA), pp. 142–149, ACM, 2010.

[120] S. Park and H. Kautz, "Privacy-preserving recognition of activities in daily living from multi-view silhouettes and rfid-based training," in *AAAI Symposium on AI in Eldercare: New Solutions to Old Problems*, 2008.

[121] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg, "A scalable approach to activity recognition based on object use," in *In Proceedings of the International Conference on Computer Vision (ICCV), Rio de*, 2007.

[122] D. Chen, J. Yang, R. Malkin, and H. D. Wactlar, "Detecting social interactions of the elderly in a nursing home environment," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 3, February 2007.

[123] A. Hauptmann, J. Gao, R. Yan, Y. Qi, J. Yang, and H. Wactlar, "Automated analysis of nursing home observations," *Pervasive Computing, IEEE*, vol. 3, no. 2, pp. 15–21, 2004.

[124] M. Popa, A. K. Koc, L. J. Rothkrantz, C. Shan, and P. Wiggers, "Kinect sensing of shopping related actions," in *Constructing Ambient Intelligence*, pp. 91–100, Springer, 2012.

[125] G. Ballin, M. Munaro, and E. Menegatti, "Human action recognition from rgb-d frames based on real-time 3d optical flow estimation," in *Biologically Inspired Cognitive Architectures 2012*, pp. 65–74, Springer, 2013.

[126] J. Han, E. J. Pauwels, P. M. de Zeeuw, and P. H. de With, "Employing a rgb-d sensor for real-time tracking of humans across multiple re-entries in a smart environment," *Consumer Electronics, IEEE Transactions on*, vol. 58, no. 2, pp. 255–263, 2012.

[127] X. Zhang, J. Yan, S. Feng, Z. Lei, D. Yi, and S. Z. Li, "Water filling: Unsupervised people counting via vertical kinect sensor," in *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pp. 215–220, IEEE, 2012.

[128] V. Reddy, C. Sanderson, and B. Lovell, "An efficient and robust sequential algorithm for background estimation in video surveillance," in *ICIP09*, pp. 1109–1112, 2009.

[129] M.-T. Yang, K.-H. Lo, C.-C. Chiang, and W.-K. Tai, "Moving cast shadow detection by exploiting multiple cues," *Image Processing, IET*, vol. 2, no. 2, pp. 95–104, 2008.

[130] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[131] H. Kingravi, P. A. Vela, and A. Grey, "Reduced set kpca for improving the training and execution speed of kernel machines," in *SIAM Int. Conf. on Data Mining*, SIAM, 2013.

[132] H. A. Kingravi, *Reduced-Set Models for Improving the Training and Execution Speed of Kernel Methods*. PhD thesis, Georgia Institute of Technology, 2014.

[133] C. Williams and M. Seeger, "Using the nyström method to speed up kernel machines," in *Proceedings of the 14th Annual Conference on Neural Information Processing Systems*, no. EPFL-CONF-161322, pp. 682–688, 2001.

[134] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, pp. 1177–1184, 2007.

[135] C. M. Christoudias, B. Georgescu, and P. Meer, "Synergism in low level vision," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 4, pp. 150–155, IEEE, 2002.

[136] S. Hutson, S. L. Lim, P. J. Bentley, N. Bianchi-Berthouze, and A. Bowling, "Investigating the suitability of social robots for the wellbeing of the elderly," in *Affective computing and intelligent interaction*, pp. 578–587, Springer, 2011.

[137] N. Ezer, A. D. Fisk, and W. A. Rogers, "Attitudinal and intentional acceptance of domestic robots by younger and older adults," in *Universal access in human-computer interaction. Intelligent and Ubiquitous Interaction Environments*, pp. 39–48, Springer, 2009.

[138] I. H. Kuo, J. M. Rabindran, E. Broadbent, Y. I. Lee, N. Kerse, R. M. Q. Stafford, and B. A. MacDonald, "Age and gender factors in user acceptance of healthcare robots.," in *RO-MAN*, pp. 214–219, 2009.