Video Detection and Classification of Pedestrian Events at Roundabouts and Crosswalks

Final Report

Prepared by:

Ted Morris
Xinyan Li
Vassilios Morellas
Nikos Papanikolopoulos

Department of Computer Science and Engineering
University of Minnesota

CTS 13-27

Technical Report Documentation Page

1. Report No.	2.	3. Recipients Accession No.
13-27		

4. Title and Subtitle	5. Report Date
Video Detection and Classification of Pedestrian Events at Roundabouts and Crosswalks	August 2013
	6.

7. Author(s)	8. Performing Organization Report No.
Ted Morris, Xinyan Li, Vassilios Morellas, Nikos Papanikolopoulos	

9. Performing Organization Name and Address	10. Project/Task/Work Unit No.
Department of Computer Science University of Minnesota 200 Union St SE Minneapolis, MN 55455	CTS Project #2012049
	11. Contract (C) or Grant (G) No.

12. Sponsoring Organization Name and Address	13. Type of Report and Period Covered
Intelligent Transportation Systems Institute Center for Transportation Studies University of Minnesota 200 Transportation and Safety Building 511 Washington Ave. SE Minneapolis, MN 55455	Final Report
	14. Sponsoring Agency Code

15. Supplementary Notes
http://www.its.umn.edu/Publications/ResearchReports/

16. Abstract (Limit: 250 words)

A well-established technique for studying pedestrian safety is based on reducing data from video-based in-situ observation. The extraction and cataloging from recorded video of pedestrian crossing events has largely been achieved manually. Although the manual methods are generally reliable, they are extremely time-consuming. As a result, more detailed, encompassing site studies are not practical unless the mining for these events can be automated. The study investigated such a tool based on utilizing a novel image processing algorithm recently developed for the extraction of human activities in complex scenes. No human intervention other than defining regions of interest for approaching vehicles and the pedestrian crossing areas was required. The output quantified general event indicators—such as pedestrian wait time, and crossing time and vehicle-pedestrian yield behaviors. Such data can then be used to guide more detailed analyses of the events to study potential vehicle-pedestrian conflicts and their causal effects. The evaluation was done using an extensive set of multi-camera video recordings collected at roundabouts. The tool can be used to support other pedestrian safety research where extracting potential pedestrian-vehicle conflicts from video are required, for example at crosswalks at urban signalized and uncontrolled intersections.

17. Document Analysis/Descriptors	18. Availability Statement
Computer vision, Learning (Artificial intelligence), Pedestrian flow, Pedestrian safety, Crosswalks, Roundabouts, Pedestrian accessibility	No restrictions. Document available from: National Technical Information Services, Alexandria, Virginia 22312

19. Security Class (this report)	20. Security Class (this page)	21. No. of Pages	22. Price
		37	

Video Detection and Classification of Pedestrian Events at Roundabouts and Crosswalks

Final Report

Prepared by:

Ted Morris
Xinyan Li
Vassilios Morellas
Nikos Papanikolopoulos

Department of Computer Science and Engineering
University of Minnesota

August 2013

Published by:

Intelligent Transportation Systems Institute
Center for Transportation Studies
University of Minnesota
200 Transportation and Safety Building
511 Washington Ave. S.E.
Minneapolis, Minnesota 55455

Acknowledgment

The authors would like to acknowledge individuals and organizations that made this research possible. They study was funded by the Intelligent Transportation Systems Institute, affiliated with the University of Minnesota's Center for Transportation Studies (CTS). Financial support was provided by the United States Department of Transportation's Research and Innovative Technologies Administration (RITA) University Transportation Centers (UTC) program. The authors would also like to give special acknowledgement to the Minnesota Traffic Observatory within the Department of Civil Engineering for access to their roundabout video database of pedestrian and vehicle activities.

Contents

Contents

1 Introduction 1
2 Background 3
3 Pedestrian Event Detection Methodology 5
3.1 Vehicle and pedestrian foreground extraction 5
3.2 Test Phase 9
3.3 Online Adaptation 10
3.4 Pedestrian Event Characterization 12
4 Experimental Testing And Data Reduction 15
5 Results 17
5.1 Minnehaha Intersection 17
5.2 Portland Intersection 20
6 Conclusions 23
References 25

List of Figures

Figure 1 The pedestrian event detection tool algorithm and workflow 5
Figure 2 Batch training algorithm for learning the background model. 8
Figure 3 Foreground detection algorithm for online test phase. 10
Figure 4 Pedestrian event detection processing, Minnehaha Ave. 16
Figure 5 Pedestrian event detection processing, Portland Ave. 16
Figure 6 Natural scene crosswalk occlusions at Minnehaha roundabout 18
Figure 7 Forground detection errors, Portland roundabout 21

List of Tables

Table 1 Video data selected for analysis 15
Table 2 Event Detection Accuracy, Minnehaha 17
Table 3 Truth table for Minnehaha traffic detection 18
Table 4 Truth table for Minnehaha yielding detection 19
Table 5 Minnehaha pedestrian-yield wait and crossing times 20
Table 6 Event detection accuracy, Portland Ave. 20
Table 7 Truth table for Portland Ave. traffic detection 21
Table 8 Truth table for Portland yielding detection 21
Table 9 Portland pedestrian-yield wait and crossing times 22

Executive Summary

A well-established technique for studying pedestrian safety is based on reducing data from video-based in-situ observation. Such video data was recently collected to evaluate pedestrian safety at roundabouts. Video camera sensors are an obvious choice over in-situ manual observational studies, which require tedious, intensive manual labor that is error-prone and cannot be rectified after the fact. The amount of data and events to be categorized in the video data can quickly overwhelm manual measurement and extraction techniques. As a result, more detailed, encompassing site studies are not practical unless the mining for these events can be automated.

The objective of this study was to develop and test the feasibility of a novel computer vision tool to extract and categorize pedestrian events at roundabout intersections. Specifically, two different urban Roundabout intersections were utilized to categorize yielding behaviors, pedestrian wait-at-curb times, and crossing times. The tool integrated a novel and computationally efficient image-processing algorithm that was recently developed for extraction of human activities in very complex scenes. The concept is based on adaptive dictionary learning where the changing background scene is learned with assumptions that foreground changes are relatively random and sparse. Automated tracking was then facilitated to aid crossing event categorization. The evaluation was done by comparing a subset of videos of multi-camera video recordings that were collected at roundabouts from a previous study done by other investigators (Minnesota Traffic Observatory). Note that their video data was not initially collected for the purpose of automated computer vision pedestrian event detection. The subset of data consisted of video segments that provided the largest subsets of different a priori cataloged pedestrian events. At present, there is no widely available, portable,'turn-key' tool to automate extraction of these events from video recordings. There have been many proposed systems aimed at pedestrian tracking but their application has been targeted for real-time traffic control and they do not have a flexible, open, architecture to extract and catalog the events.

The output categorized several different pedestrian events that could be compared to manually tabulated ground-truthed data. We examined and characterized pedestrian events based on pedestrian activity times (waiting to cross and crossing times), and vehicle/pedestrian yield events. The results indicate an event detection retrieval rate above 90%, with similar accuracy for categorizing vehicle-pedestrian presence and yielding behavior. Pedestrian event activity times could not be directly compared with the independently ground-truthed data but the findings corroborate reasonably with previous studies.

Small background movements and image noise and compression artifact could be handled by the algorithm, but detection errors were noted due to excessive camera movement and foliage foreground occlusions for several crosswalk scenes. Future studies should consider a different camera perspective to avoid the latter problems specifically.

The objective of the tool is to automate cataloging of pedestrian events to significantly streamline the process for more detailed analyses of vehicle-pedestrian conflicts and their causal effects. The tool can be used to support other pedestrian safety research where extracting potential pedestrian-vehicle conflicts from video is needed. No human intervention was required, other than defining regions of interest for approaching vehicles and the pedestrian crossing areas. Once

such information is extracted, user interfaces could then be built on top of the tool for rapid retrieval and visual inspection of the event to study potential vehicle-pedestrian conflicts and their causal effects. Robust and computationally efficient foreground detection of moving pedestrians, fed into a tracking algorithm, could be adapted to other applications, such as real-time detection of pedestrian events at unsignalized or signalized intersections to improve traffic safety, or signal timing.

The project scope entailed automated identification and categorization of pedestrian events at roundabouts but the extracted measures and set up procedures are applicable to other intersections as well. Many jurisdictions place camera sensors (fixed or Pan,Tilt,Zoom systems) located on the signal cross arm or light pole overhang, giving similar perspectives of the intersection crosswalks and oncoming traffic to the portable system camera locations used for this study at the roundabouts.

1 Introduction

Pedestrian safety has become a growing concern recently as more people choose this mode (perhaps combined with other public and bicycle transit modes) to complete their trips. A 2003 NHTSA report estimated 70,000 pedestrians were injured and 4,749 killed from vehicle related crashes [1]. More specifically, pedestrian fatalities in single vehicle crashes accounted for over 90 percent of the pedestrian fatalities from all fatal motor vehicle crashes [2]. In 2007, 24.3 percent of all pedestrian fatalities in the United States (1,143 out of 4,699) occurred at intersections. The concern among transportation professionals is also supported by a recent rise of 4% in pedestrian fatalities in 2010, after steadily falling numbers up to 2009 [3]. Nationally some states– including Minnesota– are adopting laws that require planners, MPOs, and engineers to adequately address safety and accessibility needs of pedestrians, bicyclists, and motorists [4]. Consequently, engineers and planners have started to implement new road treatment and intersection designs with the aim of reducing the risk of vehicle-pedestrian crashes. Accordingly, there is a growing need to conduct observational studies to understand underlying factors of roadway treatments that may reduce risk of pedestrian crashes. Roundabouts are one such treatment that has gained considerable attention among planners and engineers. The objective of this study was to test a computer vision-based tool to extract and quantify pedestrian events that occur at such intersections, to aid planners and engineers who wish to monitor and characterize their effectiveness and safety for pedestrians. An overview of their operation and current issues regarding field assessments is provided below.

Roundabout intersections have been prevalent in Europe and Australia for many years. In France and the United Kingdom alone they number about 55,000 of them. Within the last 10 years, popularity of roundabouts has increased dramatically within the United States, after the advent of modern designs originating in England in the early 1960's. Since the first modern roundabout was built in the US in 1990 in the state of Navada, the estimate of roundabouts has grown to about 3,500 (as of the end of 2012) [5]. Modern roundabout designs contain elements to channelize and slow down drivers, as well as offset designated pedestrian crossings with mid-road refuge islands. The assertion that they improved traffic safety and operations over the signalized intersections they replace has been generally supported by various studies over the last decade since their introduction, in addition to costing less to maintain and install.

The traffic speeds have been noted to be less than the main line, and the nature of crashes that do occur have tended to be less severe because high speed right-angle-left-turn, and head-on crashes are mitigated. A 2001 Institute study of 23 intersections in the United States reported that converting intersections from traffic signals or stop signs to roundabouts reduced injury crashes by 80 percent and all crashes by 40 percent [6]. Similar results were reported by [7]: a 72 percent decrease in injury crashes and a 47 percent decrease in total crashes at 33 intersections that were converted from traffic signals or stop signs to roundabouts. A study of 17 higher-speed rural intersections (speed limits of 40 mph or higher) found that the average injury crash rate per million entering vehicles was reduced by 84 percent and fatal crashes were eliminated when the intersections were converted to roundabouts. Studies of intersections in Europe and Australia that were converted to roundabouts have reported 25-87 percent reductions in injury crashes and 36-61 percent reductions in severe injury crashes.

At the onset of this project, despite the noted vehicular traffic safety benefits, observation intensive studies were lacking that analyzed more specifically pedestrian safety and accessibility at roundabout intersections. For example there have been questions raised that legally blind individuals cannot discern the traffic. In [8] investigators attempted to resolve blind pedestrian-vehicle yeild conflicts by adding plastic 'rumble-strips' on the interior roundabout lanes to make 'click-clack' sound cues. Yields were detected more readily from the blind subjects than the same single-lane roundabout. However, there was a significant increase in the occurrence of false-positives. Two-lane roundabouts were investigated with the same type of pavement treatment but was not effective since the majority of vehicles did not reach the rumble-strips before stopping. Other more recent studies introduced probabilistic models to predict gap crossing selection and wait times for sighted vs. bind populations [9], as well as quantified 'high risk' gap selections for different pedestrian crossing locations [10]. Note that the data for the aforementioned studies were collected using a limited number of subjects by in-situ observations.

Recently Hourdos et. al [11] completed a much more thorough video observation-based study amid perceptions from surrounding communities that pedestrian safety was being compromised at these intersections. In particular, annectodal evidence suggested that drivers do not yield to pedestrians at the crosswalks, and in any case there was little understanding of underlying factors of such designs and ensuing traffic conditions that would indeed impact pedestrian safety or accessibility. Over 1,900 hours were manually ground-truthed to categorize over 12,000 pedestrian crossing events representative of the several weeks of data collection that was completed. In their results they noted low yielding rates for pedestrians waiting to cross, particularly for a two-lane roundabout, but otherwise 'close calls' were extremely rare, and they reported on-average crossing times (including the wait-time at the curbs) were better than what is typically observed at high volume urban intersections. Originally their intent, in addition to understanding potential factors that affect pedestrian safety, was to understand aforementioned blind population pedestrian crossing behaviors within naturalistic contexts but unfortunately no such observations were found during their study.

The objective of our study was to develop a computer vision tool to automated the extraction of pedestrian events from the two different urban Roundabout intersections utilizing the aforementioned collected video data from Hourdos et. al [11] to evaluate the feasibility of the approach. The tool integrates a novel computer vision foreground detection algorithm, used for extraction of human activities in very complex scenes [12], to extract general yielding behaviors between pedestrians and vehicles, pedestrian crossing delays, and level of pedestrian utilization. Note that their video data was not initially collected for the purpose of automated computer vision pedestrian event detection. The tool can be used to automate cataloging of pedestrian events in order to significantly stream-line the process for more detailed analyses of vehicle-pedestrian conflicts and their causal effects.

The remaining report is organized as follows. Chapter 2 will review the literature on general methodologies to pedestrian detection. Chapter 3 will provide an overview of our methodology of pedestrian detection based on identifying foreground saliency from dictionary learning techniques. The data and experiments for evaluating the tool will be discussed in Chapter 4, with the results from the experiments presented in Chapter 5. Conclusions of the work will be provided in Chapter 6.

2 Background

Pedestrian and vehicle video surveillance and analysis from infrastructure-mounted video cameras have been a topic of research for many years [13, 14, 15]. Pedestrian safety studies have frequently relied on archived video-based field observations to examine pedestrian-vehicle conflicts and interactions. This is because assessing the risk from available crash records alone is inconclusive since occurrences of these events are relatively scarce. Reliable automated extraction of pedestrian crossing events has proven to be non-trivial, and consequently many investigations are done manually. A review of various machine-vision automated techniques and their application to pedestrian safety is provided below.

In general, video processing consists of (at least) three steps: segmentation of moving objects, object tracking and object motion recognition or object analysis. Reference [16] developed and tested an automated real-time pedestrian tracking algorithm for pedestrian control at signalized intersections using an active deformable model to recognize pedestrians. A two-day field test indicated a pedestrian detection accuracy of 90%. The study purported that shadows, contrast, and camera placement affect accuracy. [17] proposed pedestrian crowd tracking through optical flow rather than tracking pedestrians individually; the authors discuss its utility for controlling busy signalized intersections and crowd traffic control. [18] developed a multilevel tracking approach for tracking pedestrians and vehicles at crowded intersections. They presented modules for foreground separation of the pedestrians and vehicles from the static scene, blob tracking, Kalman-based predictive filtering for improving tracking when the pedestrian or vehicle is occluded, and finally an incident detection module based on frame-by-frame overlap of principal axis rectilinear boundary representations of the vehicles or pedestrians. The segmentation step in these formulations was based on background suppression techniques which compare the actual pixel values with the corresponding values in the (statistically) learned model of the static scene [19, 20].

Once objects (e.g. pedestrians) remain static in the image for long periods, they become invisible with the motion detection because they begin to blend into the background. Gibson et. al [21], utilized horizontally aligned stereo high-resolution digital cameras to track pedestrians within the crosswalk in real-time (5 fps) to overcome this problem. Their approach first separates pedestrians from the environment by thresholding disparity images —a gray-scale image whose pixel values represent the distance between corresponding pixels in the L/R (Left/Right) images, projected along the epipolar line. The distance of pedestrian from the stereo camera and loss of textural information from lighting and pedestrian appearance confounded foreground/background separation and the authors proposed a disparity layering technique to overcome these issues. Commercial vision-based sensors specifically designed for pedestrian presence detection at crosswalks are now available (www.traficon.com). A stereo-camera sensor configuration has the capability to track pedestrians waiting near the curb-line to cross. These cannot be used as a generalized tool for pedestrian data reduction, because they are closed systems embedding the camera sensors as part of the package, and the proprietary hardware cannot be adapted as a general post-analysis portable tool.

Lately, various groups have proposed and demonstrated machine vision tracking techniques to quantify surrogate pedestrian safety measures from recorded videos. In [22], on-coming vehicle trajectories and pedestrian cross-walk presence counting was extracted and characterized using

a blob tracking and calibration algorithms developed by [23, 24]. They recommended further studying the feasibility of their pedestrian detection algorithms to automate crossing warnings rather than utilizing pedestrian-actuated push buttons to initiate the crossings.

Rather than investigating vehicle speed profiles alone as a surrogate risk measure for pedestrian cross-walk safety, Ismail et. al [25, 26] quantified pedestrian and vehicle trajectories to estimate several traffic conflict indicators at signalized intersection crosswalks, using edge feature a tracking algorithms to track the pedestrians and vehicles [27, 28]. Specifically, the extracted pedestrian and vehicle trajectories were used to compute Time to Collision (TTC), Post-Encroachment Time (PET), Deceleration-to-Safety Time (DST), and Gap Time (GT). Time to collision (TTC) is the estimate of a time of collision if neither the vehicle or pedestrian changed course. Post Encroachment Time (PET) is the observed temporal proximity between the pedestrian and the vehicle (the point in time where they were occupying the same location). They defined GT as the PET calculated at each instant before a potential conflict while DST is the required deceleration time in order for PET > 0 (avoid a crash). The latter calculation of DST assumes the "movements of the conflicting road users remain unchanged". Extensive literature demonstrates the utility of such spatial and temporal proximity measures as surrogates to evaluate severity but are also very challenging to estimate from in-situ observations. Contrast this with longitudinal crash studies that require collecting enough crash events for statistically valid analyses (for example, in Lee and Abdel-Aty [29]). The investigators state that such an approach can be viewed as more 'reactive' and less 'proactive' due to the fact that the very crashes needed to conduct the study may have otherwise been avoided if another method to quantify severity without waiting for the crashes were instead used. As an example, their technique was used to conduct a before/after study on the efficacy of pedestrian scramble phase signal treatments using two hours video for each condition and found that there were fewer potential conflicts between pedestrians and vehicles with the scramble-phase control plan [30]. Note that operator observation and correction is necessary in some cases to correct tracking inaccuracies, and incorrect grouping and identification of pedestrians.

A limited study by [31] developed a human-assisted trajectory extraction tool to analyze the crossing behavior of 154 pedestrian crosswalk events extracted from 77 pedestrians at a single-lane roundabout. The image processing consisted of creating several regions of interest (ROI); pedestrian waiting zones, conflict zone (intersection of cross walk area and roadway), and approaching and passed vehicle zones. Pedestrian crossing speed, gap acceptance, critical gap value, and level of driver yield behavior were computed and analyzed in order to develop a traffic microsimulation model of a roundabout that can consider pedestrian interactions. Effects of changing background and lighting from environmental factors were not considered for these studies.

To conclude, many computer vision-based approaches have been proposed, although there is no widely available, field-proven, turnkey tool that can be used to extract and categorize pedestrian crossing events over long periods where changes in lighting, camera, and background motion occur. The aim of this project was to test the feasibility of such a tool that leverages a recent approach for robust foreground segmentation of moving objects in static scenes. The output from the tool compliment the accessibility and general safety Measures Of Effectiveness (MOEs) - for example wait times, and yielding order – as presented above in [31], [9], [11] and [10]. The next chapter will summarize the methodology and initial results from previous work and its formulation and adaptation for outdoor pedestrian surveillance.

3 Pedestrian Event Detection Methodology

The tool pedestrian event detection framework which integrates our approach is shown in figure 1. We analyzed the intersections in the context of how they are intended to operate. Several user-defined Region Of Interests (ROI) are constructed to monitor pedestrian or vehicle presence. The pedestrian crossing is broken into sidewalk/island regions and the crosswalk. Region Of Interests (ROI) are constructed which encompassed the pedestrian crossing area and adjoining sidewalk/curb regions. A second ROI defines vehicular traffic movement within, and outside the roundabout: a) traffic inflows leading into the roundabout, and b) traffic outflows coming from the roundabout (figures 4 and 5). Dictionary background learning was used for foreground object detection, followed by a blob tracking algorithm to discriminate between vehicles and pedestrians and track their locations (differentiation between pedestrians and bicycles was not part of the detection tool). Then activity time in the defined ROIs are used to output pedestrian detection event characterizations: a) vehicle presence, b) vehicle/pedestrian yield, and c) pedestrian delay and crossing times. The background and foreground detection methodology are described next, followed by the subsequent steps to characterize the pedestrian events.

3.1 Vehicle and pedestrian foreground extraction

The basic assumption behind the vehicle and pedestrian foreground segmentation is that their presence is transient and occurs infrequently relative to the remaining background, which has a much lower temporal variance [32, 33]. In particular the foreground objects are modeled as a sparse error, and a model of the background can be learned from a sequence of training images

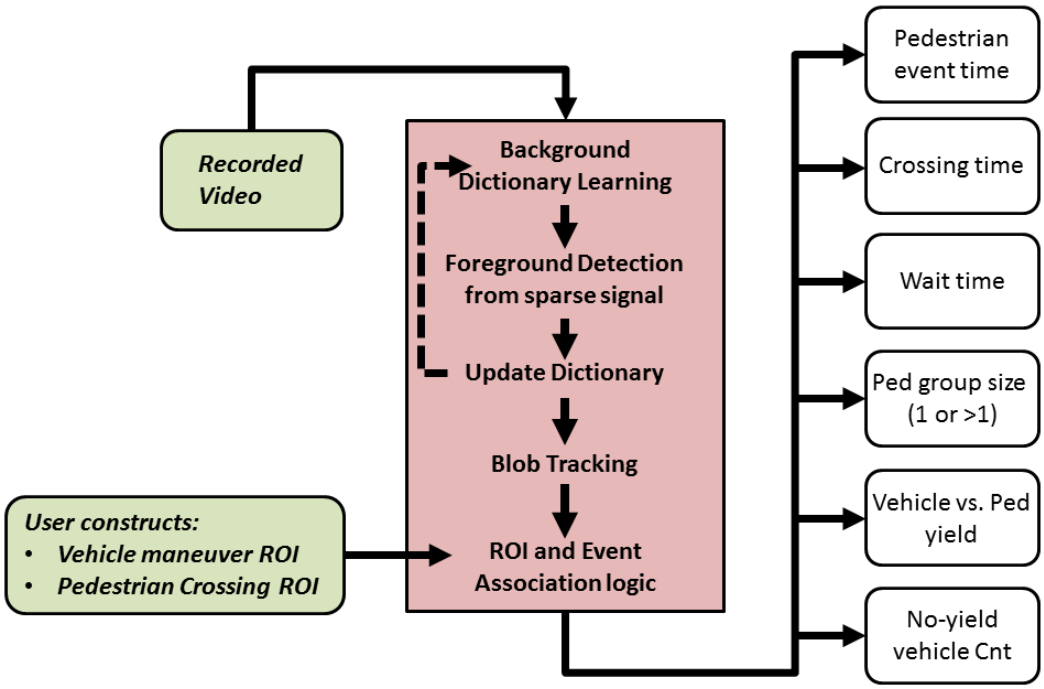


Figure 1: The pedestrian event detection tool algorithm and workflow

from the camera. More concretely, given a static view of the roundabout approach leg and crosswalk, we modeled the background as a linear combination of dictionary atoms (e.g., columns of a dictionary matrix, W), and separate out pedestrians and vehicles as the sparse error produced when a new image is linearly approximated using a linear combination of the dictionary atoms. The first step is a batch training phase to learn the background model dictionary, and then the online test (or performance) phase where we use the learned model to perform foreground detection of any pedestrians or vehicles within incoming frames. The derivation and algorithm steps are discussed below.

Let $A \in \mathbf{R}^{m \times n}$ be the data matrix consisting of n training images as columns, with $m = m_1 \times m_2$, the number of pixels in each image. The model is represented by decomposing the data matrix A as

$$A = WH + Y, \tag{1}$$

where $W \in \mathbf{R}^{m \times k}$, $H \in \mathbf{R}^{k \times n}$, $Y \in \mathbf{R}^{m \times n}$, and k is the number of atoms in the background model, or dictionary, W. Each incoming image is represented as a linear combination of atoms in the background model, plus a sparse error Y which represents the foreground. A requirement of the model is that the image signal dimension m is very large, while $k < n \ll m$.

Given the above model representation of the scene, the training phase then involves estimating the matrices W, H and Y from A. Only the background model W is retained from the training phase to perform foreground detection in incoming frames during the online test phase. The optimization problem is posed as:

$$\min_{W,H,Y} \quad \sum_{j=1}^{n} \|Y_j\|_1 + \lambda \|H\|_F^2, \tag{2}$$

or substituting for the error term, Y in equation (1),

$$\min_{W,H} \quad \sum_{j=1}^{n} \|A_j - WH_j\|_1 + \lambda \|H\|_F^2, \tag{3}$$

where the subscript j denotes the j^{th} column of the matrix, and the subscript F refers to the Frobenius (matrix) norm and $\|\cdot\|_1$ is the ℓ_1 norm (absolute value).

A smoothness constraint is imposed on the coefficient matrix H, so that the different frames use the atoms similarly. In other words, the frames should not vary drastically in their usage of a particular atom, which translates into minimizing the variability of pixel intensities across the training frames. In this study λ = 1.

The optimization problem in (3) is not convex in both W and H, but is convex for one of the variables as long the other one is fixed. Thus, by alternating the minimization of W and H, we are guaranteed to arrive at a fixed point (W^*, H^*) which is either a local minimum or a saddle point [34]. Empirical evidence shows that the local optima obtained are sufficient for good background modeling and foreground detection. Each iteration of the training phase thus consists of two steps, as follows:

GIVEN W, FIND H

Given a fixed dictionary W, we optimize (3) over H:

$$\min_{H} \quad \sum_{j=1}^{n} \|A_j - WH_j\|_1 + \lambda \|H\|_F^2 \tag{4}$$

$$\text{or,} \quad \min_{H,Y} \quad \sum_{j=1}^{n} \|Y_j\|_1 + \lambda \|H\|_F^2. \tag{5}$$

Using (1), we get

$$\min_{Y} \quad \sum_{j=1}^{n} \|Y_j\|_1 + \lambda \|\tilde{A} - W^{\dagger}Y\|_F^2, \tag{6}$$

where $\tilde{A} = W^{\dagger}A$, and $W^{\dagger} = \left(W^T W\right)^{-1} W^T$ is the pseudo-inverse of W. Note that since $W^T W$ is only a $k \times k$ matrix, where k is small, the matrix inversion is not too computationally expensive. We replaced H in the above minimization using the transformation

$$WH = A - Y \quad \Rightarrow \quad H = W^{\dagger}(A - Y) \quad \Rightarrow \quad H = \tilde{A} - W^{\dagger}Y, \tag{7}$$

Substituting back into (5), and rewriting this, we get

$$\min_{Y} \quad \sum_{j=1}^{n} \left\{ \frac{1}{2} \|\tilde{A}_j - W^{\dagger}Y_j\|_F^2 + \tilde{\lambda} \|Y_j\|_1 \right\}, \tag{8}$$

where $\tilde{\lambda} = 1/2\lambda$. This is a standard Lasso problem in $\tilde{A}$, $W^{\dagger}$ and Y, and thus can be solved quickly and efficiently.

GIVEN H, FIND W

Given the coefficient matrix H, we now optimize (3) over W, as follows:

$$\min_{W} \quad \sum_{j=1}^{n} \|A_j - WH_j\|_1 = \sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij} - \sum_{l=1}^{k} W_{il} H_{lj}|. \tag{9}$$

Our method to learn the optimal dictionary W is motivated by the K-SVD algorithm of Aharon et al. [35]. We learn the optimal W by optimizing each atom W_p at a time. The new W_p is computed such that the approximation error induced by omitting the atom is minimized in the ℓ_1-sense:

$$\min_{W} \quad \sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij} - \sum_{l=1}^{k} W_{il} H_{lj}| \tag{10}$$

$$= \min_{W} \quad \sum_{i=1}^{m} \sum_{j=1}^{n} \left| \left(A_{ij} - \sum_{l \neq p}^{k} W_{il} H_{lj} \right) - W_{ip} H_{pj} \right|. \tag{11}$$

Given : n training frames in $A \in \mathbf{R}^{m\times n}$, parameters k and λ
Initialization : Start with a random dictionary $W \in \mathbf{R}^{m\times k}$.
Repeat until convergence,

Given W, find H:

- Compute $W^{\dagger} = pinv(W)$, $\tilde{A} = W^{\dagger}A$ and $\tilde{\lambda} = 1/2\lambda$.
- Use Lasso, or any other sparse coding algorithm, to compute $Y_j \in \mathbf{R}^m$ for $j = 1,2,\ldots,n$

$$\min_{Y_j} \ \frac{1}{2}\|\tilde{A}_j - W^{\dagger}Y_j\|_F^2 + \tilde{\lambda}\|Y_j\|_1.$$

- Compute $H = \tilde{A} - W^{\dagger}Y$.

Given H, find W: For $p = 1,2,\ldots,k$,

- Compute $Y^{(p)} = A - W^{(p)}H$.
- Obtain the new atom W_p^*, as

$$W_p^* = \arg\min_{w} \ \sum_{j=1}^{n} \|Y_j^{(p)} - H_{pj}w\|_1.$$

Upon convergence, compute the sparse error $Y = A - WH$. The foreground in the training frames is given by $E = |Y|$, where $|\cdot|$ represents the element-wise absolute value.

Figure 2: Batch training algorithm for learning the background model.

For $p = \{1,2,...,k\}$, we thus compute the matrix $Y^{(p)} = A - W^{(p)}H$, where $W^{(p)} = [W_1, W_2, ..., W_{p-1}, \mathbf{0}, W_{p+1}, ..., W_k]$, and then find the atom W_p^*, where

$$W_p^* = \arg\min_{w} \ \sum_{j=1}^{n} \|Y_j^{(p)} - H_{pj}w\|_1. \tag{12}$$

Since the cost function is strictly non-negative and the elements w_i are independent of each other, we can separate the optimization into single variables as

$$W_{ip}^* = \arg\min_{w_i} \ \sum_{j=1}^{n} |Y_{ij}^{(p)} - H_{pj}w_i| = \sum_{j=1}^{n} |H_{pj}| \left| \frac{Y_{ij}^{(p)}}{H_{pj}} - w_i \right|.$$

$$W_{ip}^* = \arg\min_{w_i} \sum_{j=1}^{n} |Y_{ij}^{(p)} - H_{pj} w_i| \tag{13}$$

$$= \arg\min_{w_i} \sum_{j=1}^{n} |H_{pj}| \left| \frac{Y_{ij}^{(p)}}{H_{pj}} - w_i \right|. \tag{14}$$

W_{ip}^* is the weighted ℓ_1-norm minimizer of the sequence $\left\{ \frac{Y_{ij}^{(p)}}{H_{pj}} \right\}_{j=1}^{n}$ (formed from the rows of $Y^{(p)}$ and the row H_p), with corresponding weights $\left\{ |H_{pj}| \right\}_{j=1}^{n}$, i.e. the weighted median. Due to this weighting, data points that do not use the particular atom (i.e., $H_{pj} = 0$) are not included in the optimization. In this fashion each element of W_p^* is computed.

The first step in each iteration involves solving n Lasso problems. The second step involves solving mk 1-dimensional weighted median problems, each of which can be solved in $O(n)$ time, resulting in a time complexity of $O(mnk)$. The dictionary update step is therefore relatively inexpensive compared to the coefficient update. Thus, by alternating minimization of W and$H$, we obtain a solution (W^*, H^*) to the problem (3). The background model W^* obtained from the training is retained for the test phase, where we perform foreground detection on the incoming video images.

3.2 Test Phase

Once the background model W is learned, it can be used to perform foreground detection in the online test phase, where new images come in. The foreground detection step is the same as step (1) in the batch training phase (Refer to figure 3 for further details).

Given a new image $a \in \mathbf{R}^m$, the sparse error y is estimated, with a regularization on the coefficient vector $h \in \mathbf{R}^k$. The model is, as before, given by

$$a = Wh + y. \tag{15}$$

The optimization problem to perform online foreground detection is given by

$$\min_h \quad \|a - Wh\|_1 + \lambda \|h\|_2^2, \tag{16}$$

which is again convex and is easily solved to get the optimal solution. By a similar change of variables as before, $h = W^\dagger(a - y) = \tilde{a} - W^\dagger y$, $\tilde{a} = W^\dagger a$, we get

$$\min_y \quad \|y\|_1 + \lambda \|a - W^\dagger y\|_2^2. \tag{17}$$

Rewriting this, we get

$$\min_y \quad \frac{1}{2}\|a - W^\dagger y\|_2^2 + \tilde{\lambda} \|y\|_1, \tag{18}$$

where $\tilde{\lambda} = 1/2\lambda$. This is also a simple Lasso problem, which makes the online foreground detection procedure relatively fast.

Given : A new image frame $a \in \mathbf{R}^m$, regularization parameter λ, learned background model W ($W^\dagger$)

- Compute $\tilde{a} = W^\dagger a$ and $\tilde{\lambda} = 1/2\lambda$.
- Use Lasso, or any other sparse coding algorithm, to compute $y \in \mathbf{R}^m$

$$\min_y \ \frac{1}{2}\|\tilde{a} - W^\dagger y\|_2^2 + \tilde{\lambda}\|y\|_1.$$

- Compute $h = \tilde{a} - W^\dagger y$.

From the sparse error $y = a - Wh$, obtain the foreground as $e = |y|$.

Figure 3: Foreground detection algorithm for online test phase.

The foreground detection step can operate at a speed of $\sim$ 20–25 frames per second on a 120×160 pixel grayscale image, with an optimized LARS implementation [36]. Further, there is an extensive body of research on adapting the Lasso algorithm for directly or approximately solving large-scale problems efficiently, which our method can benefit from. The Sparse Modeling Software C++ toolbox, using the MatLAB mex-file interface was used to carry out the training process as summarized in figure 2 using a Lasso ℓ_1 minimization solver [37]. The same library was then used for estimating the foreground as summarized in figure 3. A selected foreground error threshold value $|y| \geq 0.5$ (the intensity range between 0 and 1), provided consistent results for our experiments. For our tests we utilized $n = 50$ training images randomly selected from the first minute of video, with a dictionary size of $k = 10$ atoms with row size m equal to $\frac{1}{2}$ the video image size (360×240). Since the error term, Y is expressed during the dictionary learning phase, the foreground pedestrian detection is also accomplished during this step (if such an event occurs).

3.3 Online Adaptation

An important characteristic for any background subtraction algorithm is the ability to adapt to gradually changing scenes due to for example, changes in sunlight. It should also recover quickly in cases of more sudden changes in lighting. We outline below the process of adapting the learned background model during the test phase. Since our method fits a linear model to the incoming frame, the coefficients h would appropriately scale up or down the brightness of the model to fit the image illumination, in order to handle gradual global illumination changes. At the same time, we wish to ensure that pedestrian events remain as salient foreground errors.

For every incoming frame a, we obtain the corresponding coefficient vector h and the sparse error y as part of the foreground detection step. We treat the training set A as a first-in-first-out queue and push in a new frame after t test images from the video. The training data matrix and the corresponding coefficient matrix become

$$A_{new} = [\, A_2 \ldots A_n \; a \,] \quad H_{new} = [\, H_2 \ldots H_n \; h \,]. \tag{19}$$

We then run one iteration of step (2) of the training phase (given H update model W), which is a relatively inexpensive computation and in our case, done infrequently.

As discussed in [12], the algorithm benefits from an arrival rate $1/t$ for updating the dictionary W to randomly vary throughout time, rather than be held constant. A Poisson random variable with the mean arrival rate $1/t$ was implemented to model the random variation gradual background changes of the scene occurring at random periods, $t \geq 0$ (if $t = 0$ another random period selection is generated). The shortest time between updates can be estimated if the distribution of pedestrian wait times were known *a priori*. If pedestrian is waiting at the curb line for an extended period of time, t should be selected such that incoming frames would rarely capture this state, in order to avoid 'learning' the pedestrian as background information. The study in [9] provides one way to estimate such wait delays. Without going into great detail, a probabilistic log-regression model was derived from empirical observations to predict average pedestrian crossing delay, d_p at roundabouts expressed as the following:

$$d_p = -0.78 - 14.99 \cdot \ln(P_{cross}) \tag{20}$$

where, P_{cross}, the probability of crossing is computed by:

$$P_{cross} = P(Y_{ENC}) \cdot P(GO|Yield) + P(CG_{ENC}) \cdot P(GO|CG) \tag{21}$$

where,
$P(Y_{ENC})$ =probability of encountering a yielding vehicle
$P(GO|Yield)$ =probability of utilizing a yielding vehicle
$P(CG_{ENC})$ =probability of encountering a sufficient crossing critical gap,$\geq T_{CG}$
$P(GO|CG)$ =probability of of utilizing a sufficient crossing gap $\geq T_{CG}$

As noted in [9], an estimate of $P(CG_{ENC})$ can be computed assuming hourly traffic volumes are Poisson distributed. Furthermore, the critical crossing gap time, T_{CG} can be estimated from Highway Capacity Manual using the crosswalk length, L_{cw}, pedestrian walking speed, S_{walk}, and a pedestrian start-up delay, T_d [38]. Neither the hourly traffic volume counts or Origin/Destination turning ratios were available for the intersections. From the average daily traffic volumes we spread out between 6AM and 6PM (12 hours) associated with the road crossings tested herein. In addition we hypothesize utilization; for blind subjects utilization was significantly lower than sighted individuals (a good fit was obtained for $P(GO|CG) = P(GO|Yield) \approx 50\%$. Sighted individuals will be considerable higher; suppose we choose utilization levels of 80% for $P(GO|Yield)$ and $P(GO|CG)$, with drivers yielding 30% of the time. Then, for Minnehaha following average pedestrian wait delay time can then be estimated as:

$$T_{HW} = \frac{3600 sec./hr}{333 vph} = 10.8 sec./veh$$

$$T_{CG} = \frac{L_{cw}}{S_{walk}} + T_d = \frac{14 ft.}{3.5 ft./sec.} + 2 sec. \approx 6.0 \text{ sec.}$$

$$P(CG_{ENC}) = P(headway \geq T_{CG}) = e^{-\frac{T_{CG}}{T_{HW}}} = 0.573$$

then based on the latter utilization assumptions , using equation 20 results in:

$$d_p = -0.78 - 14.99 \cdot \ln(0.8 \cdot 0.573 + 0.8 \cdot 0.3) = 4.6 \text{ sec.}$$

A similar calculation (assuming similar utilization characteristics) for the Portland intersection roundabout will yield:

$$T_{HW} = \frac{3600 sec./hr}{670 vph} = 5.37 \text{sec./veh}$$
$$T_{CG} = \frac{L_{cw}}{S_{walk}} + T_d = \frac{23 ft.}{3.5 ft./sec.} + 2 sec. \approx 8.57 \text{ sec.}$$
$$P(CG_{ENC}) = P(headway \geq T_{CG}) = e^{-\frac{T_{CG}}{T_{HW}}} = 0.20$$

with an average delay of:

$$d_p = -0.78 - 14.99 \cdot \ln(0.8 \cdot 0.2 + 0.8 \cdot 0.3) = 13 \text{ sec.}$$

The analysis gives some general insights on a minimum t frame interval to ensure reliable foreground detection. A reasonable arrival rate value which reflects twice the amount of times in the above calculations could serve as a minimum boundary (10 seconds to 30 seconds). For our experiments the Poisson arrival parameters was set to 5 minutes with pedestrian detection rates above 90%. To test algorithm sensitivity, future experiments could diminish the arrival rates, based on the pedestrian event rate detection to set the lower boundaries as done above. This implies an adaptive feedback loop for setting minimum t inter-frame arrival times for updating the dictionary, W.

3.4 Pedestrian Event Characterization

Once foreground objects were extracted, tracking is performed to track the progression of vehicles within the oncoming traffic region-of-interests, and the pedestrian interactions within the crosswalk region of interest. After the foreground binary image is obtained, a sequence morphological erosion and dilation operations are performed to filter the noise and fill holes in the blobs. Very small regions are removed by erosion and dilation using a $3 \times 3 pixel$ structured rectangular element, while aggregating blobs are completed by dilation and then erosion using a $15 \times 15 pixel$ structured rectangular element. The structured elements sizes were determined empirically. A discrete Kalman filter-based predictive tracking algorithm was implemented using the Matlab vision toolbox. The association of detection to the same object is based completely on motion state. The motion of each foreground object track is estimated by the Kalman filter, which then predicts the track's location in each frame and is used to determine the likelihood of any given detection being assigned to an existing track. The measurement state vector X_j, used for the predicted centroid of blob j, is based on pixel position and velocity, e.g.

$$X_j = [x, y, \dot{x}, \dot{y}]_j^t \tag{22}$$

with the state transition prediction, X_j^*, with the state matrix, A, at frame n defined by:

$$
\begin{aligned}
X^*_{j,n} &= A \cdot X_{j,n-1} \\
&= \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot X_{j,n-1}
\end{aligned} \tag{23}
$$

The observation measurement update, Z computed directly from the observed blob centroid, X_j:

$$
\begin{aligned}
Z &= H \cdot X_j \\
&= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \cdot X_j
\end{aligned} \tag{24}
$$

and the initial constant process noise covariance, Q, error covariance P_0, and constant observation measurement noise covariance, R, given as:

$$
\begin{aligned}
P_0 &= \begin{bmatrix} \sigma^2_{xx} & 0 & \cdots & 0 \\ 0 & \sigma^2_{yy} & & \vdots \\ \vdots & & \sigma^2_{\dot{x}\dot{x}} & \vdots \\ 0 & \cdots & \cdots & \sigma^2_{\dot{y}\dot{y}} \end{bmatrix} \\
Q &= \begin{bmatrix} q^2_{xx} & 0 & \cdots & 0 \\ 0 & q^2_{yy} & & \vdots \\ \vdots & & q^2_{\dot{x}\dot{x}} & \vdots \\ 0 & \cdots & \cdots & q^2_{\dot{y}\dot{y}} \end{bmatrix} \\
R &= \begin{bmatrix} \rho^2_{xx} & 0 & \cdots & 0 \\ 0 & \rho^2_{yy} & & \vdots \\ \vdots & & 0 & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix}
\end{aligned} \tag{25}
$$

where $\sigma^2_{xx} = \sigma^2_{yy} = 200\ pixels^2$, $\sigma^2_{\dot{x}\dot{x}} = \sigma^2_{\dot{y}\dot{y}} = 50\ (pixel/frame)^2$,
$q^2_{xx} = q^2_{yy} = 100\ pixels^2$, $q^2_{\dot{x}\dot{x}} = q^2_{\dot{y}\dot{y}} = 25\ s(pixel/frame)^2$, and
$\rho^2 xx = \rho^2_{yy} = 100\ pixels^2$.

Equations (23), (24), and (25) are then used to obtain predictions to the error covariance and state described in equation (23) for each blob centroid state X_j:

$$
P^*_n = AP^{n-1}A^T + Q \tag{26}
$$

and measurement state and covariance updates for each blob centroid X_j, to the Kalman gain, K_n, *a posteriori* error, $y_{j,n}$, using a standard discrete Kalman filter:

$$y_{j,n} = Z_{j,n} - HX^*_{j,n} \tag{27a}$$
$$K_n = P^*_n H^T (HP^* H^T + R)^{-1} \tag{27b}$$
$$X_{j,n} = X^*_{j,n} + K_n y_{j,n} \tag{27c}$$
$$P_n = (I - K_n H) P^*_k \tag{27d}$$

When tracks become lost (for example if a pedestrian walks behind another pedestrian) the Kalman filter prediction steps in (26) and (23) are used to maintain the trajectory of the object. For every time step a cost matrix of size $Mtracks \times Ndetections$ is computed based on the Euclidean distance between the Kalman predicted track centroid and the detection. A minimum total cost assignment calculation between the M tracks and N detections was done using Munkres' Assignment Algorithm. The cost for not assigning a track (maximum cutoff) is set experimentally based on expected deviations from motion and pixel noise. A very small value caused the tracks to quickly get lost, while large values tended to force blobs which touch each other (such as multiple pedestrians) to merge; however casual observations of the data indicated the propensity for groups of people to walk close together, and were hard to differentiate with blobs at such a long distance from the camera over different maximum cost cutoff values; 12 pixels produced consistent results across the videos we tested.

When a foreground object (such as a pedestrian reappears), the cost estimation assigns it to a new or existing track based on the minimum cost calculation previously discussed. Pedestrians are identified by the region of interest and tracking the blob area. If a tracked object remains in prediction mode to update position and velocity states with the (e.g., occluded or missing) for 10 frames (about 1.25 seconds), then the track is lost. In the same an object must be visible for at least 1 second before it is considered to be tracked.

The final step uses the region occupancy logic to decipher a.) vehicle presence, b.) vehicle passed through the crosswalk while the pedestrian was in a waiting zone, and c.) pedestrian crossing and waiting times.

4 Experimental Testing And Data Reduction

Eight cameras were housed in a single dome enclosure, suspended approximately 35' above the roadway surface using a trailer crank-up mast, to simultaneously monitor the entire circumference of the roundabout. The video data sets were collected on two roundabouts with different lane geometries and different urban settings. The camera locations for the 66th and Portland Ave. Richfield site, the two lane roundabout intersection adjoining two urban arterials, were approximately 100 ft. (30.5m) from the crosswalks. For the single-lane roundabout intersection, located at Minnehaha Parkway and Minnehaha Avenue, in Minneapolis, the cameras were actually located further from the crosswalk at approximately 120 ft. (36.7m). The trailer was deployed at the Richfield site from August 7, 2010 until September 4, 2010, totaling 29 days. Recordings were made between 7 AM and 9 PM daily. A commercial DVR recorded compressed, proprietary H.264 mpeg video at 7 fps (variable bit rate at approximately 1 mbps), full 720x480 NTSC video, which was transcoded into Microsoft AVI files in open standard Xvid (www.xvid.org) MPEG4 format.The data separated into 1 hour chunks, represented within the files.

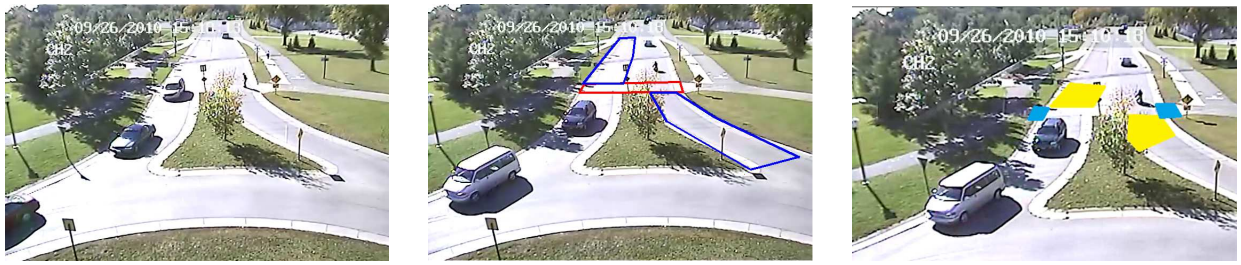
The same video collection set up was followed for the data collection at the Minneapolis roundabout. The trailer was deployed September 18, 2010 and remained in this location until October 11, 2010. Recording was scheduled between 5 AM to 8 PM daily. The authors noted that not all video could be manually analyzed due to time and cost constraints but nevertheless 4,730 pedestrian crossing and 7,302 bicycle crossing events were tabulated from $4\,cameras \times 15\,hours \times 16\,days \times 2\,sites = 1,920\,hours$. Further details on the experimental apparatus and sites can be found in [11].

Table 1 summarizes a subset of video data containing pedestrian events were used from [11]. In accordance to the original aim of the study, our objective was to find a reasonable subsample of the data which balanced at-large total pedestrian events with pedestrian-vehicle interactions, and different environmental conditions. This was achieved by building a database using all the ground-truthed data, to associate every event back to the original recording files, and times relative to sunset (if past noon) or sunrise (if before noon). We then searched the database to sort the top 12 video sets to meet aforementioned objective. Due to time limitations, 6 of these were used for the study.

Table 1: Video data selected for analysis

Intersection	Video File	Total events	Ped.-veh. events
Minnehaha	100926-ch04-161018-171018	205	14
	100926-ch02-151018-161018	126	14
	100926-ch04-151018-161018	162	14
	100919-ch02-150646-160646	137	27
Portland	100820-ch01-110415-120415	11	8
	100817-ch01-140400-150400	12	9

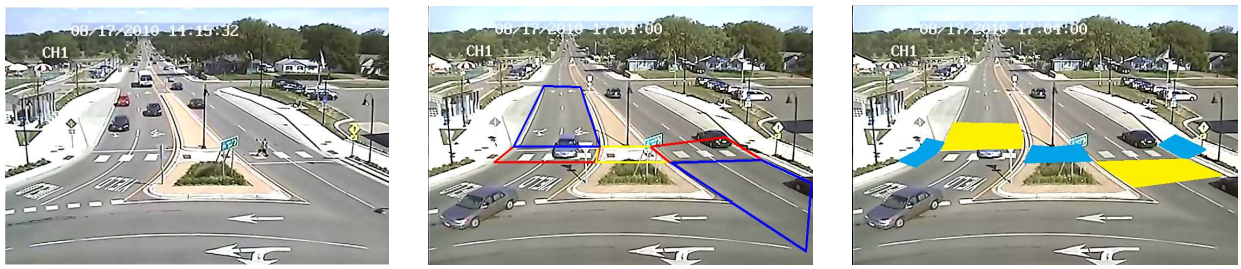
As mentioned previously, Matlab was utilized as the primary interface to implement the approach. For this study, the Dictionary on-line learning and training phases were done separately, with the resulting video stored for subsequent processing. Figures 4 and 5 illustrates the steps and resulting processed object tracks used for the event measurements.



(a) (b) (c)

(d) (e) (f)

Figure 4: The pedestrian event detection at the Minnehaha Roundabout, inputs: (a) raw video stream, (b) pedestrian crossing and vehicle detection ROI, (c) pedestrian yield ROI; image processing steps: (d) Foreground dictionary result, (e) thresholding=0.5, and (f) blob-based object tracking



(a) (b) (c)

(d) (e) (f)

Figure 5: The pedestrian event detection at the Portland Roundabout, inputs: (a) raw video stream, (b) pedestrian crossing and vehicle detection ROI, (c) pedestrian yield ROI; image processing steps: (d) Foreground dictionary result, (e) thresholding=0.5, and (f) blob-based object tracking

5 Results

The results will be presented separately for the two roundabouts. The ground-truthed data as described in chapter 4 were reviewed by manual observation to compare our results. We would not expect the onset (arrival) times between the human observed ground-truthed events and computer vision detection events to align exactly. The first level of analysis is pedestrian detection sensitivity and accuracy. Then, we drill down logically to determine if traffic was present during the crossing event, and if the vehicle yielded. Pedestrian crossing delays were then analyzed. As discussed earlier in the report, the latter two characterizations are useful parameters to aid in understanding accessibility and utilization of the intersection. Initially we attempted to quantify the number of pedestrians crossing with blob tracking but they frequently blended together. We also attempted to utilize a HOG appearance (Histogram of Oriented Gradients) tracker by [39] but were unsuccessful due to insufficient pixel areas to obtain a robust set of HOG features.

5.1 Minnehaha Intersection

The general Pedestrian detection rate from the Minnehaha intersection is provided in table 2. Four videos with a highest number of pedestrian events were used, on two separate days in early and late September from mid to late afternoon, from two different crosswalk locations. The data indicates an overall detection accuracy of 92.6% , with a tendency to be over-sensitive to the pedestrian detection (false positives). Channel 04 performed better than channel 02 because in channel 04 a tree occluded one side of a crossing, while in channel 02, a tree obstructs most of the middle of the crossing (figure 6). Visual observation of the events revealed that the over counting (false positives detections) were due to 'double-counting' a correctly detected pedestrian event. In this case, 100919-ch02-150646-160646 revealed 23 (17%) cases that were 'double-counted' with a correctly detected event, while 10 (7%) others were not detected (false negatives), due to the partial occlusions and blob fragmentation causing the pedestrian tracks to become lost. The remaining videos in table 2 exhibited similar trends; in segment 100926-ch02-151018-161018 the algorithm over-counted 17 (13%) false positives, with 6 (5%) missed detections. Within the second crosswalk the algorithm over-counted segment 19 (8%) false positives, and missed 7 events (3%), and 18 (11%) false positives and 9 (5%) missed events, for segments 100926-ch04-161018-171018 and 100926-ch04-151018-161018 respectively.

Table 2: Event Detection Accuracy, Minnehaha

Video data	Ground Truth	Comp.Vision	Difference	% Accuracy	Cumulative% Accuracy
100919-ch02-150646-160646	137	150	13	90.51	
100926-ch02-151018-161018	126	137	11	91.27	92.59 ± 1.99
100926-ch04-161018-171018	205	217	12	94.15	
100926-ch04-151018-161018	162	171	9	94.44	

For examining yield and traffic interactions for the detected pedestrian events, video 100919-ch02-150646-160646 was examined. Of the 137 ground-truthed event observations, 19 of them had no traffic. Of the aforementioned 10 missed pedestrian crossing event detections, 2 were without any traffic. The algorithm detected 127 events correctly, with 15 of them having no traffic interactions, and the remaining 112 with traffic interactions. For the 23 aforementioned 'over-detected' events, 5 of them were observed with traffic but were classified as containing no traffic, and another 9 were observed with traffic and classified as such. Table 4 summarizes only the comparison with the correctly detected pedestrian events, without considering the 'double-counts'. Table 3 indicates robust event classification for these cases with a miss classification rate of less than 2%.

Table 3: Truth table for Minnehaha traffic detection

		Observed Ground Truth	
		traffic	*no traffic*
100919-ch02-150646-160646			
Comp. Vision	*traffic*	110	2
	no traffic	0	15
100926-ch02-151018-161018			
Comp. Vision	*traffic*	105	1
	no traffic	0	14
100926-ch04-161018-171018			
Comp. Vision	*traffic*	146	2
	no traffic	1	49
100926-ch04-151018-161018			
Comp. Vision	*traffic*	85	0
	no traffic	0	68

(a) Channel 04 Crosswalk (b) Channel 02 Crosswalk

Figure 6: The trees created problems breaking up blobs associated with pedestrians utilizing the Minnehaha Roundabout crosswalks.

With the properly detected data, the algorithm is used to characterize the order which vehicles yielded using the aforementioned tracks and ROI to determine the arrival time and presence of the pedestrian and vehicles to assess event detection accuracy as indicated in the truth table, 4. For example, with the first video segment in table (2) there were 27 pedestrian-yielded events and 91 vehicled-yielded events within the video. Within 8 missed detected events with the traffic interaction, 6 of them were pedestrian-yielded events with a single pedestrian, with the remaining 2 classified as vehicle-yielded events. For the remaining of the 21 pedestrian-yielded events, 19 were classified correctly, setting a 3 second (21 frames) threshold. Note that the other 2 events were not considered as vehicle-yielded events by the detection algorithm since the pedestrian waited less than 3 seconds during the pedestrian-yield cases. The results from the other three video segments in table 2 indicated similar trends with the aforementioned yield wait time threshold. To conclude there appears to be subjective variability in the ground truth data which cannot be duplicated by the algorithm which by definition must determine a fixed criteria for this event categorization.

Table 4: Truth table for Minnehaha yielding detection

		Not classified	Observed Ground Truth	
			Vehicle yielded	*Ped. yielded*
100919-ch02-150646-160646				
Comp. Vision	*Vehicle yielded*	3	86	0
	Ped. yielded	2	0	19
100926-ch02-151018-161018				
Comp. Vision	*Vehicle yielded*	3	91	0
	Ped. yielded	1	0	10
100926-ch04-161018-171018				
Comp. Vision	*Vehicle yielded*	5	129	0
	Ped. yielded	2	0	11
100926-ch04-151018-161018				
Comp. Vision	*Vehicle yielded*	4	69	0
	Ped. yielded	4	0	8

Lastly for the aforementioned pedestrian-yield events an approximate waiting time delay is computed. The activity ROIs were used to parse wait time at the curb (or island) and the crossing times for Minnehaha (table 5 and Portland 9). Unfortunately, the ground-truthed data were not done to do a valid comparison (a smaller subset of the videos were manually analyzed regarding these data). The somewhat broad range in crossing times was due to the fact that some pedestrians were on bicycles and thus the crossing time is reduced compared to walkers. Although we cannot generalize too much from these data, it is interesting to note that even the wait times for pedestrian-yeilds would be less than a signal phase cycle, which could be at least 30 seconds (for example 1/2 of 60 second cycle, and so on) if this intersection were signalized.

Table 5: Minnehaha pedestrian-yield wait and crossing times

	Wait time(at curbside)	Crossing Time	Total Time
Min	3.0	2.6	6.6
Max	26.1	9.7	35.7
Mean	9.4	5.1	14.5
Std. dev	7.5	2.0	8.4

5.2 Portland Intersection

The average daily vehicle traffic levels, as discussed in Chapter 4 is considerably higher than the Minnehaha intersection, but the number of pedestrians utilizing this intersection was found to be substantially lower. Two videos containing the highest pedestrian traffic levels were analyzed and are summarized in table 6. The data for this intersection indicated an overall detection accuracy of 94.3%. Once again there is a slight over sensitivity to pedestrian events. Note that we observed a maximum +2 frame difference ($1/7\ sec. \times 2 = 0.28$) ahead of what is observed, with a maximum of +5 frames delay.

Table 6: Event detection accuracy, Portland Ave.

Video data	Ground Truth	GT (stable cam)	Comp. Vision	Diff.	% Accuracy	Cum% Accuracy
100817-ch01-140400-150400	12	11	10	1	90.91	94.7 ± 6.4
100820-ch01-110415-120415	11	8	8	0	100	

Unlike the Minnehaha intersection the were no significant foreground occlusions resulting in more consistent, complete tracks. However, here was was considerable camera movement in these videos, which significantly impacted the performance of the on-line dictionary learning foreground detection (figure 7). During these conditions, the total foreground regions was significantly larger than surrounding frames. Therefore, a tolerance value based on the ratio of forground vs. background pixels on the non-ROI areas of the image was used to remove the frames. This resulted in loss of some pedestrian detection events. Even though the pedestrian usage at this intersection was substantially less than Minnehaha, and resulted in only a few missing detections, in general this problem will need to be addressed in future work.

For this intersection video 100817-ch01-140400-150400_xvid, the ground truth indicated a total of 12 events, with 5 of them no-traffic events. The algorithm miss-detected one event, containing traffic. Eleven events were detected correctly with four of them no-traffic conditions, and seven, with-traffic interactions (table 7).

There are were 9 pedestrian-yielded events and 2 vehicle-yielded events that were indicated from the ground-truthed data. For the missed pedestrian detection event with traffic interaction, it was a pedestrian-yielded event. For the remaining 8 pedestrian-yielded events, the algorithm successfully detected 7 of them with approximately 3 second threshold. But the remaining event

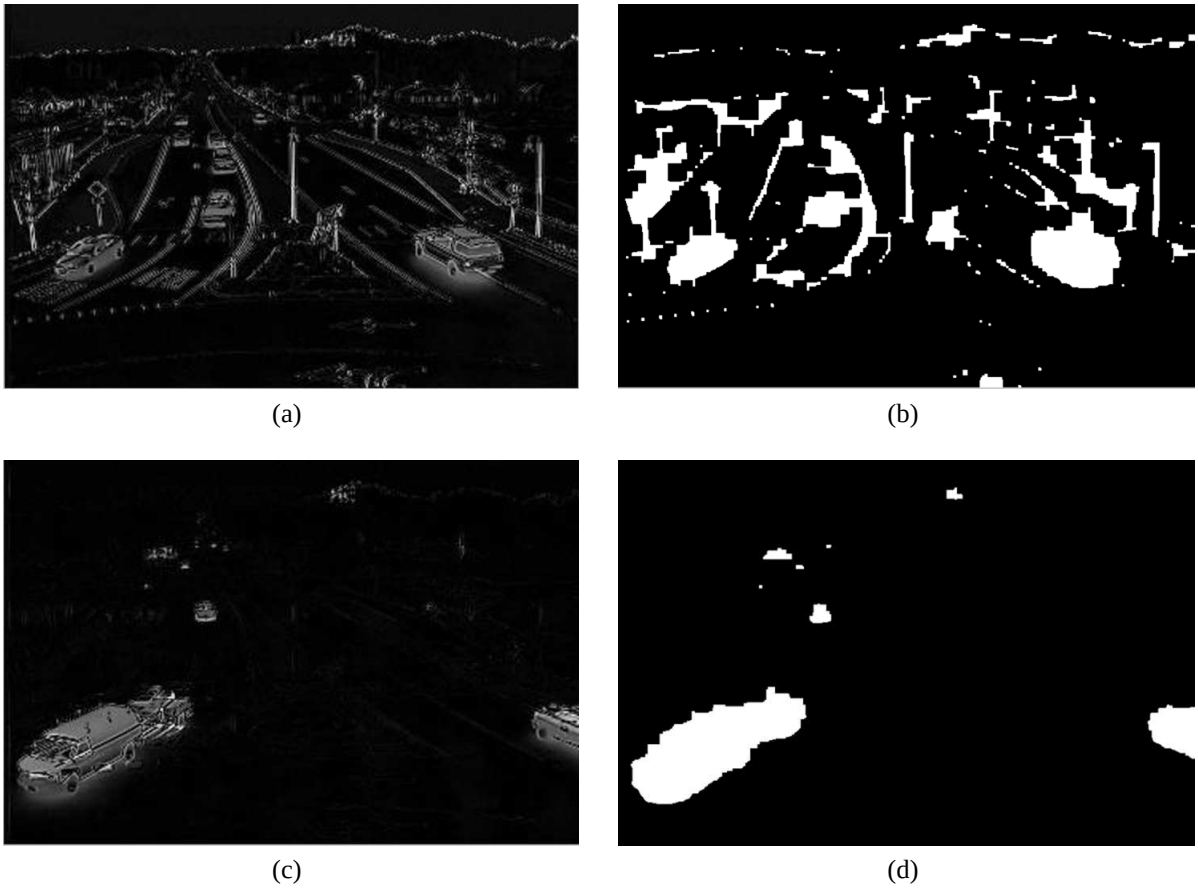
Table 7: Truth table for Portland Ave. traffic detection

000817-ch01-140400-150400		Observed Ground Truth	
		traffic	*no traffic*
Comp. Vision	*traffic*	6	1
	no traffic	0	4

was not considered in the ground-truth data as a vehicle-yielded event when the pedestrian waited less than 3 seconds and a vehicle passed through. The 3 vehicle yielded events were detected correctly (table 8).

Table 8: Truth table for Portland yielding detection

000817-ch01-140400-150400		Not classified	Observed Ground Truth	
			Vehicle yielded	*Ped. yielded*
Comp. Vision	*Vehicle yielded*	0	2	0
	Ped. yielded	1	0	7



(a) (b)

(c) (d)

Figure 7: Failed background subtraction. (7a and 7b the raw and thresholded result); Normal background subtraction (7c and 7d the raw and thresholded result)

The waiting and pedestrian crossing times for pedestrians shown in 9) generated longer wait times for yielding pedestrians than Minnehaha. With two lanes, and a larger refuge island that was not occluded, it was possible to extract both crossing legs times. As in the case of the Minnehaha intersection, there were no independently ground-truthed data to use for comparison. Once again, with so few samples analyzed one cannot substantiate any claims to pedestrian utility at this intersection. It is interesting to note however, longer yielding wait times for pedestrians than the Minnehaha intersection. This generally corroborates with [11]. Also, the crossing times are generally longer which is to be expected due to the fact that this is a two lane roundabout.

Table 9: Portland pedestrian-yield wait and crossing times

	Wait time(at curbside)	Wait time (at Island)	Crossing Time	Total Time
Min (sec.)	6.1	3.9	6.6	15.6
Max (sec.)	28.4	33.9	12.0	64.0
Mean (sec.)	28.0	12.7	10.3	31.6
Std. dev. (Sec.)	10.5	11.5	2.1	16.7

6 Conclusions

We tested the feasibility of integrating a novel foreground detection scheme into a tool for extracting pedestrian crossing events. The variability from ground-truthed data are partially due to subjective variability in manual observations and criteria set forth in our methodology. The approach was generally robust to small changes in the background and noise in the video. Large camera shifts due to the cameras being hoisted on a crank-up mast and windy conditions could not be mitigated with the technique. The investigation focused on such events at urban roundabout intersections, but the same methodology can be applied at other intersections as well. We posit that the methodology could be extended, as part of a roadside sensing system, for real-time pedestrian event detection for driver warning systems at critical intersections, for example. The basis of detecting pedestrian activity relies on building sparse dictionaries that are robust to natural changes in lighting and other natural environmental affects associated with outdoor sensing. For studies or systems that require long-term deployments and continual, round-the-clock detection, such dictionaries could first be learned for example over a 24-hour period, hashed with respect to time-of-day, and normalized for sunrise and sunset times, to accommodate general trends in lighting. For the pedestrian events detected, it was not necessary to conduct homography calibration to map image coordinates to the road and crosswalk plane.

The study indicated that the approach performed satisfactorily. The classification rates are suitable for data mining purposes for video-based observational studies (above 90% for the detection measurements and characterizations we considered). Many of the detection error problems arose from foreground occlusions (tree foliage at one intersection), which could be avoided by placing the camera in a different location to avoid it). As noted at the beginning of the report, the original video data collection was designed for manual observation specifically without considering optimal placement for computer vision tracking tasks.

A future direction to this work is to extend the tool to different intersection designs. A Graphic User Interface could be designed to allow the user to steer pedestrian event classifications for the intersection under study. For example, [40] broke down right-turn permitted and protected vehicle-pedestrian yielding behaviors, on mainline crosswalks; appropriate regions of interests could be entered by the users, as well as categories to define the pedestrian events (vehicle vs. pedestrian yielding, pedestrian wait-time and crossing time, alone vs. groups, and so on). The tool can be extended to extract pedestrian crossing gap acceptance as done by in-situ manual observation as in [9, 10, 31], to assess probabilistic crossing delays and potentially risky pedestrian crossing behaviors.

Finally, computational efficiency can be improved with this approach. The learning and update process involve calculations that were $O[mkn]$. For a robust real-time monitoring application, the dictionary updates can be performed as an occasional background process as opposed to running in sequence with the processing steps.

The output characterization and identification of pedestrian events from the tool compliment the accessibility and general safety Measures Of Effectiveness (MOEs) - crossing gap acceptance, wait times, and yeilding order – as presented above in [31], [9], [11] and [10]– none of which require accurate extrinsic and intrinsic camera calibration.

References

[1] *Traffic Safety Facts 2003: A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System*. Technical report, National Highway Traffic Safety Administration, U.S. Department of Transportation, Washington D.C., 2004.

[2] Umesh Shankar. *Pedestrian Roadway Fatalities*. Technical report, National Center for Statistics and Analysis, FHWA, U.S. Department of Transportation, Wash. D.C., 2003.

[3] National Highway Transportation Safety Administration. NSCA data resource website: Fatality Analysis Reporting System (FARS) Encyclopedia. http://www-fars.nhtsa.dot.gov, 1999. (*accessed Jan 2013*).

[4] Minnesota DOT. MN Complete Streets, sec 52,174.75. http://www.dot.state.mn.us/planning/completestreets/legislation.html, 2010. *(Jan 2013)*.

[5] Bill Baranowski. Roundabouts USA. http://www.roundaboutsusa.com/history.html, 1999. (*accessed Dec 2012*).

[6] B. N. Persaud, R. A. Retting, P. E. Garder, and D. Lord. "safety effect of roundabout conversions in the united states: Empirical bayes observational before-after study". *Transportation Research Record*, (1751):1– 8, 2001.

[7] S. Eisenman, J. Josselyn, G. List, B. Persaud, B. Robinson C. Lyon, E. Waltman M. Blogg, and R. Troutbeck. *Operational and Safety Performance of Modern Roundabouts and Other Intersection Types*. Technical report, New York State Department of Transportation, 2004.

[8] Vaughan W. Inman, Gregory W. Davis, and Dona Sauerburger. "Roundabout Access for Visually Impaired Pedestrians: Evaluation of a Yielding Vehicle Alerting System for Double-Lane Roundabouts". In *National Roundabout Conference 2005 Proceedings*, Vail, Colorado, May 22- 25 2005.

[9] Bastian J. Schroeder and Nagui M. Rouphail. "Mixed-Priority Pedestrian Delay Models at Single-Lane Roundabouts". *Transportation Research Record*, (2182):129–138, 2010.

[10] David A. Guth, Richard G. Long, Robert S. Wall Emerson, Paul E. Ponchillia, and Daniel H. Ashmead. "Blind and Sighted Pedestrians' Road-Crossing Judgments at a Single-Lane Roundabout". *Human Factors*, Sept 2012.

[11] John Hourdos, Veronica Richfield, and Melissa Shauer. *Investigation of Pedestrian/Bicyclist Risk in Minnesota Roundabout Crossings*. Technical report, Minnesota Department of Transportation, St. Paul, MN, Sept. 2012.

[12] R. Sivalingam, A. D'Souza, M. Bazakos, R. Miezianko, V. Morellas, and N. Papanikolopoulos. "Dictionary Learning For Robust Background Modeling". In *Robotics and Automation (ICRA), 2011 IEEE International Conference*, Shanghai, China, 2011.

[13] Tarak Gandhi and Mohan Manubhai Trivedi. "Pedestrian Protection Systems: Issues, Survey, and Challenges". *IEEE Transactions On Intelligent Transportation Systems*, 8(3):413– 430, 2007.

[14] R. Hosie, S. Venkatesh, and G. West. "Detecting Deviations From Known Paths And Speeds In A Surveillance Situation". In *Proc. of the Fourth International Conference on Control, Automation, Robotics and Vision*, pages 3–6, Singapore, Dec. 1996.

[15] B. Heisele and C. Wohler. "Motion-Based Recognition of Pedestrians". In *Proc. of the Fourteenth International Conference on Pattern Recognition*, pages 1325–1330, Brisbane, Australia, August 1998.

[16] Osama Masoud and N. Papanikolopoulos. *Pedestrian Control At Intersections: Phase IV, Final Report*. Technical report, Minnesota Department of Transportation, Saint Paul, MN, 2000.

[17] B. Maurin, O. Masoud, S. Rogers, and N. Papanikolopoulos. *Pedestrian Control Issues At Busy Intersections And Monitoring Large Crowds*. Technical report, Minnesota Department of Transportation, St. Paul, MN, March 2002.

[18] Harini Veeraraghavan, Osama Masoud, and Nikolaos P. Papanikolopoulos. "Computer Vision Algorithms for Intersection Monitoring". *IEEE Transactions On Intelligent Transportation Systems*, 4(2):78–89, 2003.

[19] C. Stauffer and W. Grimson. "Adaptive Background Mixture Models For Real-Time Tracking". In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, page 252, Fort Collins, Colorado, June 23 - 25 1999.

[20] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. "Pfinder: Realtime Tracking Of The Human Body". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[21] David R. P. Gibson, Bo Ling, Michael Zeifman, Shaoqiang Dong, and Uma Venkataraman. "Multipedestrian Tracking". *Public Roads*, 69(5), Mar-Apr 2006.

[22] Thomas J. Smith, Curtis Hammond, Guruprasad Somasundaram, and Nikolaos Papanikolopoulos. *Warning Efficacy of Active Versus Passive Warnings for Unsignalized Intersection and Mid-Block Pedestrian Crosswalks*. Technical report, Minnesota Department of Transportation, St. Paul, MN, 2009.

[23] O. Masoud and N. Papanikolopoulos. "Using Geometric Primitives to Calibrate Traffic Scenes". In *Proc. of the 2004 IEEE/RJS International Conference on Intelligent Robots and Systems*, pages 1878– 1883, Sendai, Japan, 2004.

[24] G. Somasundaram, C. Hammond, T.J. Smith, and N. Papanikolopoulos. "A Vision-Based System for Studying the Efficacy of Pedestrian Crosswalk Warnings". Poster Presentation, University of Minnesota Center for Transportation Studies Nineteenth Annual Transportation Research Conference, Saint Paul, MN, May 2008.

[25] K. Ismail, T. Sayed, and N. Saunier. "Automated Analysis Of Pedestrian-Vehicle Conflicts Using Video Data". *Transportation Research Record*, pages 44–54, 2009.

[26] K. Ismail, T. Sayed, and N. Saunier. "automated Analysis of Pedestrian-Vehicle Conflicts". *Transportation Research Record: J. of the Transportation Research Board*, (2198):52–64, 2010.

[27] S. T. Birchfield. KLT: An Implementation of the Kanade-Lucas-Tomasi Feature Tracker. http://www.ces.clemson.edu/~stb/klt/, 2002. (*accessed Jan 2013.*)

[28] J. Shi and C. Tomasi. Good Features to Track. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 593– 600, Seattle, Washington, June 21-23 1994.

[29] Chris Lee and M. Abdel-Aty. "Comprehensive Analysis Of Vehicle-Pedestrian Crashes At Intersection In Florida". *Accident Analysis & Prevention*, 37(4):775–786, 2005.

[30] K. Ismail, T. Sayed, N. Saunier, and C. Lim. "Automated Analyisis Of Pedestrian-Vehicle Conflicts: Context For Before-And-After Studies". *Transportation Research Record*, (2198):52 – 64, 2010.

[31] KoSok Chae and N. Rouphail. "Emperical Study of Pedestrian-Vehicle Interactions in the Vicinity of Single-Lane Roundabouts". In *Proc. Transportation Research Board Annual Meeting*, Washington D.C., January 13-17 2008.

[32] Volkan Cevher, Aswin Sankaranarayanan, Marco F. Duarte, Dikpal Reddy, Richard G. Baraniuk, and Rama Chellappa. "Compressive Sensing for Background Subtraction". In *ECCV '08: Proceedings of the 10th European Conference on ComputerVision*, pages 155–168, Marseille, France, 2008.

[33] M. Dikmen and T.S. Huang. "Robust Estimation of Foreground in Surveillance Videos by Sparse Error Estimation". In *Proc. 19th International Conference on Pattern Recognition, 2008*, pages 1–4, Tampa, FL, 2008.

[34] James C. Bezdek and Richard J. Hathaway. "Some Notes on Alternating Optimization". In *Proc. of the 2002 AFSS International Conference on Fuzzy Systems, Calcutta*, pages 288–300, London, UK, 2002.

[35] M. Aharon, M. Elad, and A. Bruckstein. "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation". *IEEE Transactions on Signal Processing*, 54(11):4311–4322, Nov. 2006.

[36] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. "Least Angle Regression". *Annals of Statistics*, 32(2):407–499, 2004.

[37] Willow Project. SPArse Modeling Software. http://spams-devel.gforge.inria.fr, 2007. (*accessed Dec 2012*).

[38] Naugi M. Raouphail, Joseph E. Hummer, and Joseph S. Milazzo. *Recommanded Procedures Chapter 13, "Pedestrians," of the Highway Capacity Manual*. Technical report, Federal Highway Administration, US Department of Transportation, McLean, Virginia, 1998.

[39] H. Pirsiavash, D. Ramanan, and C. Fowlkes. "Globally-Optimal Greedy Algorithms for Tracking a Variable Number of Objects". In *Computer Vision and Pattern Recognition (CVPR) Colorado Springs*, Colorado Springs, CO., June 2011.

[40] S. M. L. Hubbard, R. J. Awwad, and D. M. Bullock. Assessing the Impact of Turning Vehicles on Pedestrian Level of Service at Signalized Intersections. *Transportation Research Record*, (2027):27 – 36, 2007.