WRONG-WAY DRIVING DETECTION

by

Aphinya Chairat

A progress submitted in partial fulfillment of the requirements for the
degree of Master of Engineering in
Computer Science

Examination Committee:	Dr. Matthew N. Dailey
	Dr. YOUR COMMITTEE #1
	Dr. YOUR COMMITTEE #2
Nationality:	Thai
Previous Degree:	Bachelor of Computer Engineering
	Naresuan University, Phitsanulok, Thailand
Scholarship Donor:	Royal Thai Government

Asian Institute of Technology
School of Engineering and Technology
Thailand
May XXXX

Acknowledgments

Write your touching message here..

Abstract

Abstract here..

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Humans use their eyes to capture images and send visual information to their brains to see and visually sense the world around them. Computer vision is the science that attempts to understand the world as humans do.

1.1 Overview

According to the estimates of the World Health Organization (WHO), every year, around 1.25 million people die because of road traffic crashes. Between 20 and 50 million more people are disabled because they got in an accident on the roads. They can not do ordinary work as a result of their injuries. Moreover, oftentimes, family members have to take time from work or school to take care of them.

There are many causes of road traffic crashes. The main cause is human error or bad behaviour of drivers such as over speeding, wrong-way driving, driving under the influence of alcohol or other psychoactive substances, and non-use of motorcycle helmets or seat belts. Wrong-way driving happens when a driver wants to use the shortest path to reach a destination, regardless of traffic rules, increasing the risk of accidents.

To prevent this problem, law enforcement is a good solution, and governments already have traffic officers in place to penalize offenders. But strict and ubiquitous enforcement would require many police officers everywhere. Computer vision can enable advanced technological solutions to help improve the situation through automated wrong-way driver ticketing, as one possibility. Therefore, in this study, I consider wrong-way driving monitoring by computer vision combined with a law enforcement system to help solve the problem of wrong-way driving.

1.2 Problem Statement

According to the WHO, statistics on the number of road traffic crashes is high. Governments try to solve this problem by law enforcement, and in many places, they allow automated ticketing. One behaviour leading to accidents is wrong-way driving.

Recently, computer vision has developed many technological solutions that may help. Background subtraction (Piccardi, 2004) is one method of moving object detection. Furthermore, computer vision and image processing combined with machine learning for detection, tracking, and classification are also possible solutions. Forthoffer et al. (1996) present a new an automatic incident wrong-way vehicle detection system using image processing. There has been a great deal of research about moving object detection. For example, Paragios and Deriche (2000) present geodesic active contours

and level sets for the detection and tracking of moving objects.

Optical flow (Horn & Schunck, 1981) is another useful computer vision technology. It implements the idea of measuring the motion of objects in the scene. Monteiro et al. (2007) propose wrong-way driver detection based on optical flow. The basic idea is good; optical flow is good for estimating the direction of motion of an object in the scene. However, optical flow is not very useful for identifying an object, because it only works with points in the image sequence. So while it might work well for individual vehicles with little occlusion viewed from a high height, it is not easy to group moving points into individual objects, in more crowded scenes, and it may give wrong groupings when objects are close to each other.

One of the most important considerations in building a system for law enforcement for wrong-way driving detection is the identification of the vehicle so that we can send a ticket to the vehicle owner.

1.3 Objectives

As the main objective of my research, I would like to develop a system enabling automated detection and processing of wrong-way driving violations with minimal user set up and calibration under crowded traffic conditions. Towards the main objective, in this special study, I will:

1. Study state of the art methods for estimating motion of objects in a 2D projection of a 3D scene under crowded conditions.
2. Study state of the art methods for detecting wrong-way driving vehicles.
3. Explore existing and possible methods for detecting and ticketing wrong-way driving vehicles with a case study prototype.

1.4 Limitations and Scope

The study is limited to the following:

- Study of the possible ways to detect wrong-way driving. I only focus on motion estimation for rigid 3D objects moving parallel to a ground plane.
- The experiment is preliminary and not expected to be a production-ready solution.

1.5 Thesis Outline

I organize the rest of this study as follows.

In Chapter 2, I provide a literature review.

In Chapter 3, I document a preliminary prototype wrong-way driving detection system.

In Chapter 4, I provide a plan for my thesis study.

Chapter 2

Literature Review

Many efficient techniques based on computer vision are available for classification, object detection, and tracking. Here I review existing work that will be useful for the construction of a wrong-way driving system.

2.1 Background Subtraction

Background subtraction is a commonly-used technique for generating a foreground mask. In background subtraction, a foreground mask is calculated by subtracting the current frame and a background model obtained with a fixed camera. Every object in the background model is considered as part of the background and is to be ignored.

Background modelling has two main steps:

1. Background initialization.
2. Background update, due to possible changes in the scene.

An example is shown in Figure 2.1.

Background subtraction is easy to implement, but it is slow, because background subtraction has to perform its calculation over the whole frame on every frame. Large variance in the background can lead to false negatives, and long-term scene change with a slowly-changing background may not work well.

2.2 Classification

Classification is a general technique for determining what something is (when we do not know its type) or determining where something is (when the type of object we are looking for is known).

2.2.1 Nearest Neighbor Classification

The nearest neighbor decision rule assigns a label to an unclassified sample point. The classification model simply contains a set of previously-classified points. The nearest neighbor algorithm will find the nearest item in the training set and return the label of that item. This method is simple to implement and powerful, requiring no training time. This classification method usually has the best accuracy, if the training set is large enough, but it requires too much memory and compute time with a large training set. Also, as shown in Figure 2.2, if a new data item of class 2 is closer to an item of class 1 than any item of class 2, we would classify the input incorrectly. This can be partly addressed

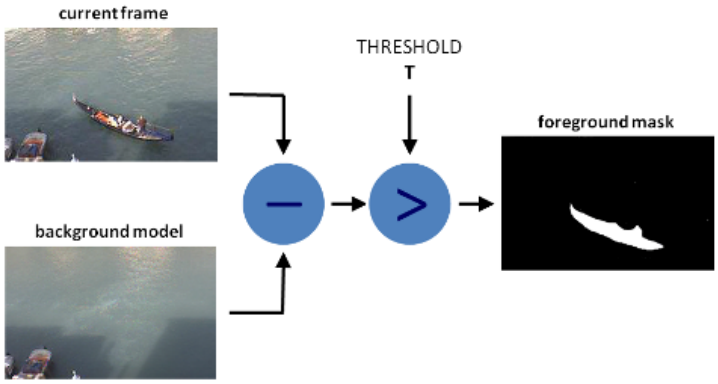

Figure 2.1: Background subtraction example from OpenCV Development Team (2017).

Figure 2.2: Example of nearest neighbor classification.

by the K-nearest neighbor method. K-nearest neighbor classification allows the user to specify K, the number of stored items to compare against. The method finds the nearest K items in the training set and returns the label of the majority within that group of items. Thus, in Figure 2.2, if K = 3, The method will classify the new item as a member of class 2.

2.2.2 Support Vector Machines

A support vector machine (SVM) is a feature-based classifier. Suppose we have a set of given data points, each belonging to one of two classes, and the goal is to decide which class a new data point is in. SVMs use linear hyperplanes for classification. There are many possible hyperplanes that might classify (separate) the training data correctly. Support vector machines choose the unique hyperplane that provides the largest separation of the training data from the hyperplane, as shown for example in Figure 2.3. When the training data are not linearly separable, the optimization of the hyperplane is more complicated, but still can be modeled as a constrained quadratic optimization. Support vector machines can induce non-linear classification boundaries due to the “kernel trick” and do not require a large amount of training data.

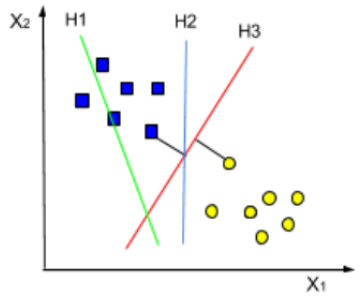

Figure 2.3: Example of large margin classification with a linear hyperplane. H1 does not separate the classes. H2 does separate the classes, but small margin. H3 separates the classes with the maximum margin.

2.2.3 Classical Backpropagation Neural Networks

Backpropagation is a method for training multi-layer neural networks. Usually, a classical neural network would be feature based. Backpropagation over few layers usually has similar performance to SVMs. The goal of backpropagation is to optimize a set of weights so that the neural network correctly maps arbitrary inputs to a correct output. The optimization algorithm repeats a two-phase cycle, forward propagation and backward weight update. The input is propagated forward through the network, layer by layer, until it reaches the output layer. The output layer is then compared to the desired output using some loss function. The loss is propagated back from the output layer toward the input layer. This method requires experimentation to tweak hidden layer size and learning time.

2.2.4 Convolutional Neural Networks

The methods described thus far are normally feature based, requiring the data scientist to mathematically specify the set of features to be extracted from the input data. This is necessary when the training set size is small. For images, CNNs offer the benefit of automatically generated features, offering better performance than feature-based methods so long as a sufficient amount of training data is available.

A CNN usually consists of one or more convolutional layers with subsampling steps followed by one or more fully connected layers. The input to a convolutional layer is a $m \times n \times r$ image, where m is the height, n is the width, and r is the number of channels in the input. The convolutional layer contains k filters or kernels of size $p \times p \times q$, where p is smaller than the dimension of the image and q can be the same as the number of channels (r) or smaller and may be different for each kernel. The convolutional units share weights across the entire input image patch. Many open source frameworks such as Caffe allow building of CNNs. An example convolution neural network is shown in Figure 2.4.

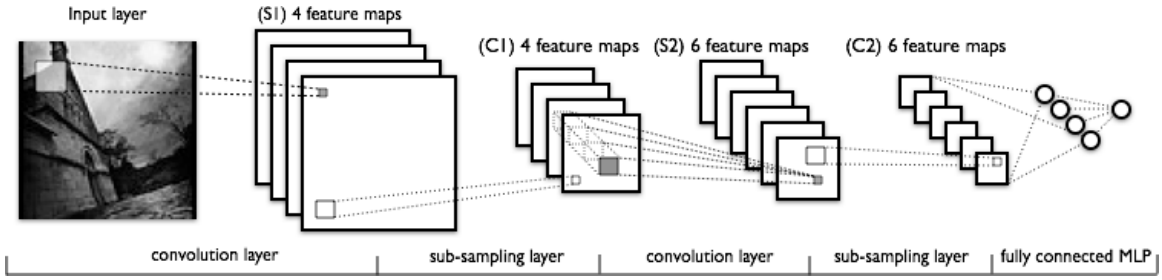

Figure 2.4: Example convolutional neural network (CNN). Reprinted from Deep Learning Development Team (2017).

2.3 Object Detection

Object detection is a very important part of a computer vision system. Detection helps us find objects in an input image. Most detection methods run a scan window over the image and classify each region as a member or not a member of the object class of interest.

2.3.1 Haar Cascades

Object detection using Haar feature-based cascades is one of the most efficient methods to detect an object in the image. Invented by Viola and Jones (2001), this method requires a medium number of input images (positive images and negative images) for training. After training, we use the resulting model to detect objects in other images.

Positive images are images of an object, and negative images are images without that object. In training, we extract the same set of features from both positive and negative images. Each feature is a single value extracted using a fixed binary kernel as shown in Figure 2.5.

A large set of possible sizes and locations of each kernel are considered, giving a huge number of possible kernels, making the training process quite heavy. The trained model might contain a large number of features, which would make it correspondingly inefficient. To increase the efficiency of feature computation, integral images are used as a memoization method enabling calculation of sums of intensities across rectangular region in constant time regardless of block size. The new problem is then to find which is a good feature, because a good feature should focus on the essential region and distinctive aspects of the object only. To select features, we use the Adaboost committee-building algorithm. The method applies each and every feature to all training images. For each feature, we find the best threshold for classifying the object as positive or negative. Each individual feature in the model is a "weak classifier" that has some level of error over its weighted training set. Different weightings of the training set give different optimal weak classifiers. The best classifier is then a weighted sum of the weak classifiers. Applying a complete committee of weak classifier to every window in an image is not a good idea. To accelerate the computation, we use a cascade of classifiers to group the features into different stages of classifiers and apply each stage one-by-one. If a window fails in first stage, we do not need to consider the remaining stages. Windows that pass though all stages are classified as the object of interest.

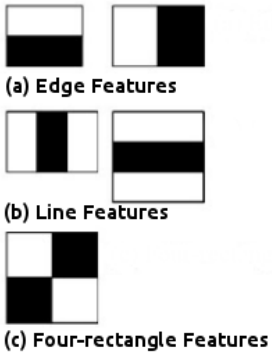

Figure 2.5: Example Haar-like kernels. The sum of the intensities of the pixels in the dark regions is subtracted from those of the light regions. Reprinted from Viola and Jones (2001).

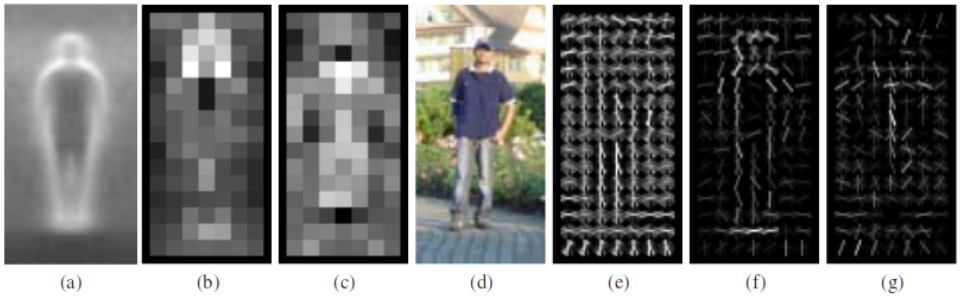

Figure 2.6: HOG classifier. a) The average gradient image over the training examples. (b) Each pixel shows the maximum positive SVM weight in the block centred on the pixel. (c) Likewise for the negative SVM weights. (d) A test image. (e) The test image R-HOG descriptor. (f,g) The R-HOG descriptor weighted by respectively the positive and the negative SVM weights. Reprinted from Dalal and Triggs (2005).

2.3.2 Histogram of Oriented Gradients

Histogram of Oriented Gradients (HOG) (Dalal & Triggs, 2005) is an object detection technique relying on a holistic descriptor of an image patch. HOG generates a representation of the object is contours. Examples of the same object should produce a descriptor as close as possible to the same descriptor calculated when the object is viewed under different conditions. A support vector machine (SVM) is used to recognize HOG descriptors as representing or not representing the object of interest. To recognize objects at different scales, the image is sub-sampled at multiple sizes. Each of these sub-sampled images is then searched to matches. A sample HOG classifier is shown in Figure 2.6.

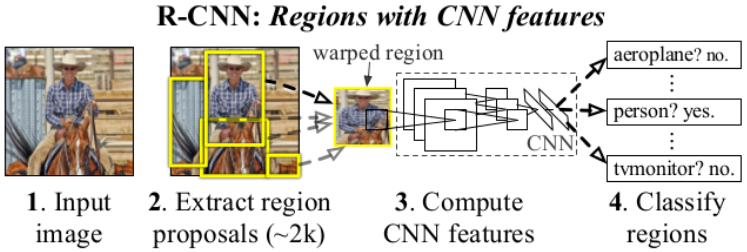

Figure 2.7: Object detection system overview. Reprinted from Girshick et al. (2014).

2.3.3 Region-based CNNs

Region-based CNNs (R-CNNs) (Girshick, Donahue, Darrell, & Malik, 2014) combine region proposals with convolutional neural networks (CNNs). When an input image is presented, the system extracts bottom-up region proposals to localize and segment objects. Then it computes features for each proposal and classifier each proposal by passing each proposal to a convolution neural network. Finally, each region is classified using class-specific linear SVMs. This method yields good results, but it requires a large amount of training data. An overview of the system is shown in Figure 2.7.

For training data, we form a set of images and boxes that include all selective search and ground-truth boxes from validation together with up to N ground-truth boxes per class from train. Training data are required for three procedures in R-CNN: (1) CNN fine-tuning, (2) detector SVM training, and (3) bounding-box regressor training. Linear SVMs are used to review the definitions briefly, for fine-tuning.

2.3.4 FAST-RCNNs

The Fast Region-based Convolutional Network method (Fast R-CNN) by Girshick (2015) builds on R-CNNs to improve training and testing speed while also increasing detection accuracy. Training models in multi-stage pipelines are slow and inelegant. R-CNNs are slow because they perform a ConvNet forward pass for each object proposal, without sharing computation. Fast R-CNN takes an input image and multiple regions of interest (RoIs) as input to a fully convolutional network. Each ROI is pooled into a fixed-size feature map and then mapped to a feature vector by fully connected layers. The network has two output vectors per RoI: a vector softmax probabilities and a vector of per-class bounding-box regression offsets. The architecture of the fast R-CNN is shown in Figure 2.8.

2.4 Optical Flow

Optical flow (Horn and Schunck (1981)) is a technique used for estimating the motion of moving objects in a scene by keeping track of features between consecutive frames. “Sparse” optical flow methods start by finding strong corners in the image and then calculating the optical flow for the

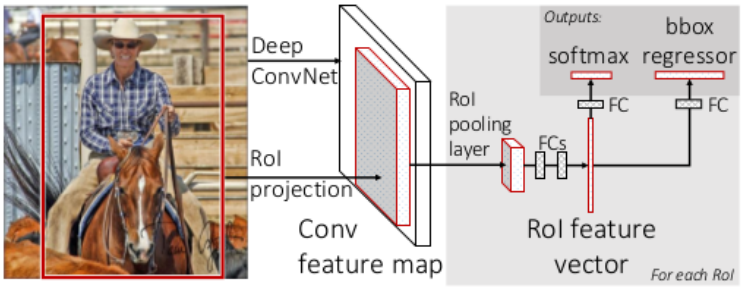

Figure 2.8: Fast R-CNN architecture. Reprinted from Girshick (2015).

sparse feature set. The most popular and efficient method is the iterative Lucas-Kanade pyramid method.

Optical flow works on two assumptions:

1. The pixel intensities of an object do not change between consecutive frames.
2. Neighbouring pixels have similar motion.

Consider the motion of a pixel $I(x, y, t)$ (where t is time) that is observed in several consecutive frames. If the object moves by a distance (dx, dy) between two frames taking dt time, under the invariant intensity assumption, we have

$$I(x, y, t) = I(x + dx, y + dy, t + dt). \tag{2.1}$$

We perform a Taylor series approximation

$$I(x + dx, y + dy, t + dt) \approx I(x, y, t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt.$$

From equation (2.1), since $I(x, y, t) - I(x + dx, y + dy, t + dt) = 0$, dividing by dt, we obtain

$$\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t}\frac{dt}{dt} = 0,$$

which we rewrite as

$$f_x u + f_y v + f_t = 0,$$

where

$$f_x = \frac{\partial I}{\partial x} \; ; \; f_y = \frac{\partial I}{\partial y}; \; f_t = \frac{\partial I}{\partial t}$$

$$u = \frac{dx}{dt} \; ; \; v = \frac{dy}{dt}.$$

The image gradients are f_x and f_y. f_t is a gradient over time. u and v (the motion of the pixel) are unknown. One method to solve this problem is Lucas-Kanade.

According to the assumption that neighbouring pixels must have similar motion, Lucas-Kanade takes a 3×3 patch around each point. These nine pixels are all assumed to have the same motion, so we calculate $(f_x, f_y, and \; f_t)$ for these nine points. We obtain nine equations in the two unknown variables. The least-squares solution for this over constrained system of linear equations is shows

below.

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_i {f_{x_i}}^2 & \sum_i f_{x_i} f_{y_i} \\ \sum_i f_{x_i} f_{y_i} & \sum_i {f_{y_i}}^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i f_{x_i} f_{t_i} \\ -\sum_i f_{y_i} f_{t_i} \end{bmatrix}.$$

Large motions will be difficult to identify with this method, as the assumption is no longer met. Hence, Lucas-Kanade is applied in a pyramid. At higher levels of the pyramid, small motions are ignored, and large motions of large regions become small motions of small regions that can be identified using the same method as above.

2.5 Traffic Flow Direction Learning

Monteiro et al. (2007) present work on a system whose basic idea is to get the correct direction of motion of vehicles in different lanes is during a learning period when it is assumed that traffic is flowing in the correct direction. Hence, the system learns the correct direction for each point in the scene. The authors propose a learning method that analyzes many frames. A Gaussian mixture model is learned for each block in the image by analysis of vehicles' movement over time. This method is good; it is able to detect vehicles circulating on the wrong side of the road, and it runs in real-time. But this method may have errors in identifying objects and may not work well on more crowded scenes. Learning process has shown in Figure 2.9.

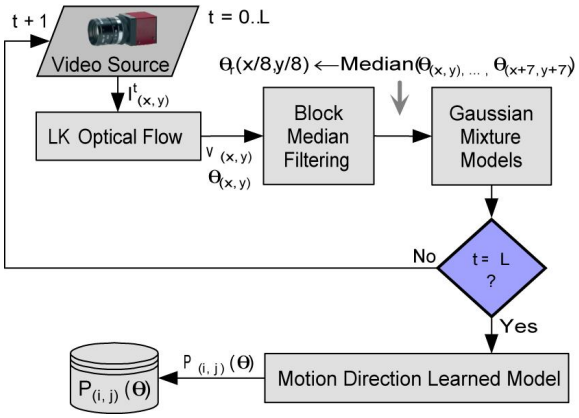

Figure 2.9: Flow chart of the traffic flow direction learning process. Reprinted from Monteiro et al. (2007).

2.6 Object Tracking

Object tracking has been a problem for research for many years. It is still not a solved problem, but there are many object trackers. Object trackers usually need some initialization step, which can be provided manually or automatically using an object detector. Tracking systems must address two basic problem: motion and matching.

- Motion problem: predict the location of an image element using previous positions.

- Matching problem: identify the image element within the designated search region.

2.6.1 CAMSHIFT

CAMSHIFT is a practical application of the meanshift method to tracking of an object using a color histrogram. Meanshift iteratively finds the mode of an arbitrary probability density using samples from the density and a kernel. Consider the situation in which we have a set of points. We are given a small window. The basic idea of meanshift is to identify the set of points that fit with the small window, then move the window should move to the area containing the maximum number of points. But a problem for the standard meanshift method is that when an object moves closer to camera, the object get bigger, making the fixed size window inapplicable. Camshift tries to solve this problem by adapting the window size and rotation of the target. It updates the size of the window as $s = 2\sqrt{\frac{M_{00}}{256}}$ where M_{00} is the zero-order moment and s is the new window size. Camshift also calculates the orientation of the best-fitting ellipse for the object. It applies meanshift with the new scaled search window and a previous window location. The method keeps executing this process until a threshold is reached.

2.6.2 Foreground Blob Tracking

Foreground blob tracking analyzes the extracted foreground blobs from detection. Usually, foreground blob detection is not very accurate, requiring smoothing using a filter such as the Kalman filter.

2.6.2.1 Kalman Filter

A Kalman filter is an optimal estimator for the state of a system evolving stochastically over time. It is an algorithm that uses a series of measurements observed over time and is recursive, so that new measurements can be processed as they arrive. Suppose we are tracking a single object. After we detect the object, the detector will give us a candidate location of the object. To predict the next position of the player, we will need an object motion model. The detector may not be perfect, so we assume there is noise in the object is location, called measurement noise. The motion model is also not perfect, thus we also have noise in the motion model, called process noise. To estimate the next position of an object, we need three parameters: the object motion model, the measurement noise, and the process noise.

Chapter 3

Methodology

Some intro..

3.1 System Overview

Some text .. Algorithm 1 just a pseudocode.

3.2 System Design

3.2.1 Design A

Some text ..

Algorithm 1 Lame Algorithm

Input: B: set of all current blobs

Input: T: set of all current tracks

Input: M: merged track association matrix

Output: $\widetilde{T}$: set of all revised tracks

Output: $\widetilde{M}$: revised merged track association matrix

$\widetilde{T} \leftarrow \emptyset$; $\widetilde{M} \leftarrow \emptyset$; $L \leftarrow \emptyset$

$A \leftarrow$ GET-OVERLAP-AREA-MATRIX(B, T)

for each $t \in T$ **do**

if t is marked as processed **then** continue

$B' \leftarrow \{b' \mid A(b', t) > 0\}$ {B' contains candidate blobs for track t.}

$T' \leftarrow \{t\} \cup \{t' \mid M(t, t') = 1\}$ {T' contains all tracks currently merged with t.}

if $|B'| \geq 1$ **then**

for each $t' \in T'$ **do**

Let $b = \underset{b' \in B'}{\operatorname{argmax}} S(b', t')$

$L \leftarrow L \cup \{(t', b)\}$

MARK-TRACK-AS-PROCESSED(t')

end for

end if

end for

for each $(t_i, t_j) \in T \times T$ **do**

If $\exists b$ s.t. $(t_i, b) \in L \wedge (t_j, b) \in L$, $\widetilde{M}_{ij} \leftarrow 1$, otherwise $\widetilde{M}_{ij} \leftarrow 0$

end for

$T^* \leftarrow \{t^* \mid \neg\exists b \in B$ s.t. $(t^*, b) \in L\}$ {T^* contains tracks for which "stale count" will be increased.}

$\widetilde{T} \leftarrow$ UPDATE-OR-DELETE-STALE-TRACKS(T, T^*)

$B^* \leftarrow \{b^* \mid \neg\exists t \in T$ s.t. $(t, b^*) \in L\}$ {B^* contains blobs with no tracks assigned.}

$\widetilde{T} \leftarrow$ ADD-NEW-TRACKS-FOR-NOT-LINKED-BLOBS$(\widetilde{T}, B^*)$

Chapter 4

Experimental Results

Some intro..

4.1 Section Name in Experimental Results

Table 4.1 shows a table.

Table 4.1: Some table.

Batch method	TP	FP	TN	FN	TPR	FPR
Local (z-scoring)	24	42	444	0	1	0.086
Local (LRT)	24	486	0	0	1	1
Global (z-scoring)	24	217	10	0	1	0.956
Global (LRT)	24	223	4	0	1	0.982

Chapter 5

Conclusion and Recommendations

Some text..

5.1 Conclusion

Text..

5.2 Recommendations

Text..

References

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE computer society conference on computer vision and pattern recognition (CVPR)* (pp. 886–893).

Deep Learning Development Team. (2017). *Deep learning documentation.*

Forthoffer, M., Bouzar, S., Lenoir, F., Blosseville, J., & Aubert, D. (1996). Automatic incident detection: Wrong-way vehicle detection using image processing. In *World congress on intelligent transport systems: Realizing the future.*

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440–1448).

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580–587).

Horn, B. K., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, *17*(1-3), 185–203.

Monteiro, G., Ribeiro, M., Marcos, J., & Batista, J. (2007). Wrongway drivers detection based on optical flow. In *IEEE international conference on image processing (ICIP)* (pp. V–141).

OpenCV Development Team. (2017). *The OpenCV reference manual.*

Paragios, N., & Deriche, R. (2000). Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*(3), 266–280.

Piccardi, M. (2004). Background subtraction techniques: A review. In *IEEE international conference on systems, man and cybernetics* (pp. 3099–3104).

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *IEEE computer society conference on computer vision and pattern recognition (CVPR).*

Appendix A

.. TITLE HERE ..

Section Name

Figure A.1 shows something.

Some text ..

Figure A.1: CCTV monitoring room. Reprinted from the Twenty First Security Web site (`http://www.twentyfirstsecurity.com.au/`).