Pedestrian Detection Based on Data and Decision Fusion Using Stereo Vision and Thermal Imaging

by

Roy Sun

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Electrical and Computer Engineering

by

April 2016

APPROVED:

Professor Xinming Huang, Major Thesis Advisor

Professor Yehia Massoud, Department Head

Abstract

Pedestrian detection is a canonical instance of object detection that remains a popular topic of research and a key problem in computer vision due to its diverse applications. These applications have the potential to positively improve the quality of life. In recent years, the number of approaches to detecting pedestrians in monocular and binocular images has grown steadily. However, the use of multispectral imaging is still uncommon. This thesis work presents a novel approach to data and feature fusion of a multispectral imaging system for pedestrian detection. It also includes the design and building of a test rig which allows for quick data collection of real-world driving. An application of the mathematical theory of trifocal tensor is used to post process this data. This allows for pixel level data fusion across a multispectral set of data. Performance results based on commonly used SVM classification architectures are evaluated against the collected data set. Lastly, a novel cascaded SVM architecture used in both classification and detection is discussed. Performance improvements through the use of feature fusion is demonstrated.

Acknowledgements

I would like to thank my advisor, Professor Xinming Huang for his teaching, guidance, and especially his generosity throughout my work. This thesis would not have been possible without his consistent encouragement and support.

I would also like to express my deepest gratitude to all the people who assisted me through this work. To everyone in our intelligent transportation group, I couldn't have done this without your ideas, expertise in machine learning and continued efforts.

I would especially like to thank my amazing parents and brother for their nurture, love, support, and constant encouragement that I have gotten over the years. I undoubtedly could not have done this without you.

And finally, to the love of my life, MH, you have always encouraged me to pursue my dreams and have supported me in every way possible; for that I am eternally grateful.

Contents

1 Introduction **1**

1.1 Motivation 2

1.2 Thesis Outline 2

2 Background **4**

2.1 Stereo and Thermal Vision 4

2.1.1 Stereo Vision 4

2.1.2 Thermal Vision 6

2.2 Trifocal Tensor 7

2.3 HOG and SVM 11

2.3.1 HOG 11

2.3.2 SVM 12

3 Related Work **16**

3.1 Background Subtraction Method 16

3.2 Color Feature Extraction Method 17

3.3 Color and Infrared Stereo Vision 18

3.4 Decision Fusion 19

4 Data Acquisition and Processing **20**

4.1 Data Collection Equipment 20
4.1.1 Stereo Vision Camera 21
4.1.2 Thermal Camera 21
4.1.3 Enclosure 22
4.1.4 Mounting System and Cable Management 23
4.1.5 Completed System 24
4.1.6 Software 26
4.2 Application of Trifocal Tensor 27

5 Data Extraction **32**
5.1 Ground Truth Data Labeling 32
5.2 Positive and Negative Sample Extraction 34

6 Pedestrian Classification and Detection **36**
6.1 Classification 36
6.2 Detection 41

7 Conclusion and Future Work **48**
7.1 Conclusion 48
7.2 Future Work 49

List of Figures

2.1 Stereo vision mathematical principle 5

2.2 Infrared spectrum breakdown 7

2.3 HOG calculation process visualized 12

2.4 Examples of many possible separating hyperplanes 13

2.5 Optimized hyperplane based on the maximum margin separating the instances 14

2.6 Schematic of cascaded SVM with feedback architecture 15

3.1 Process flow of background subtraction methodology 17

4.1 StereoLabs ZED camera 21

4.2 FLIR Vue Pro LWIR camera 22

4.3 FLIR Vue Pro cable 24

4.4 Standalone completely assembled system 25

4.5 Completed system mounted on test vehicle 25

4.6 Simple image stacking 27

4.7 Sample images from stereo vision camera 28

4.8 Sample thermal image of LWIR camera 29

4.9 Reconstruction based on trifocal tensor algorithm 31

4.10 Anaglyph overlay of reconstructed thermal image and original left vision image 31

5.1 ViPER tool interface 33
5.2 Positive samples in multispectral images 34
5.3 Negative samples in vision images 35

6.1 SVM feature confidence comparison of visual and thermal 39
6.2 SVM feature confidence comparison of visual, thermal and disparity . 40
6.3 Pedestrian blending into the background 42
6.4 Sliding window detection algorithm[1] 43
6.5 Detection result comparison of different features 45
6.6 Performance plot of different features 46
6.7 Detection result on visual image 47
6.8 Detection result on thermal image 47

List of Tables

6.1 Cross validated SVM classification of raw image data results 37

6.2 Cross validated SVM classification of HOG data results 37

6.3 Detection results based on multilayer & confidence score SVM approach 44

Chapter 1

Introduction

Object detection has received a great deal of attention in recent years. Pedestrian detection is a canonical instance of object detection that remains a popular topic of research due to its diverse applications. Direct applications in car safety, surveillance, and robotics, have continued to drive efforts in creating a more robust solution. More importantly, pedestrian detection is a well defined problem with established benchmarks and evaluation metrics. As such, it serves as a sandbox for exploring different ideas. The main paradigms for object detection are Viola Jones variants, HOG+SVM rigid templates, deformable part detectors (DPM), and convolutional neural networks (ConvNets)[2].

The aim of this thesis work is to explore advancements in presently available techniques and propose new techniques in object detection methodology. By utilizing higher dimensions of data, this work aims to improve pedestrian detection outcomes by ways of data and decision fusion.

1.1 Motivation

According to statistics[3], each year, nearly 1.3 million people die in road crashes worldwide, which is an average of 3,287 deaths a day. In addition to this, another 20-50 million people are injured or permanently disabled due to these accidents. Currently, road traffic crashes rank as the 9th leading cause of death, and is expected to become the fifth leading cause of death by 2030 unless action is taken. By developing a more affordable and robust pedestrian detection scheme, we can stop this trend and establish a safer roadway.

1.2 Thesis Outline

This thesis is organized as follows. Chapter 2 provides the background information and theories required to fully understand this work. Information such as 2D and 3D view geometry as well as commonly used machine learning architectures are provided.

Chapter 3 describes related works in the field of pedestrian detection. This field encompasses a wide array of disciplines and backgrounds, from image processing to theoretical math-based deep neural-net machine learning. Therefore only key contributing pieces of work are present as opposed to a summary of the entire field.

Chapter 4 provides a discussion of data acquisition equipment and data post processing. A custom set of data was collected as opposed to using standard pedestrian detection data sets, captured by cameras, such as INRIA, ETH, TUD-Brussels, Daimler, Caltech-USA, and KITTI as all of these lacked a third set of features: thermal imaging.

Chapter 5 discusses the processes used to extract the necessary data from our raw data set so they could be used for SVM training. Chapter 6 details the chronicle of

pedestrian classification and detection improvements through the use of novel SVM architecture. Finally, Chapter 7, gives the conclusions and an overview of future work.

Chapter 2

Background

This chapter provides important background information related to this work. A discussion of stereo vision camera theory along with infrared camera basics are provided. The mathematical theory of trifocal tensor is provided as a basis for understanding its real world application and its relevance to this work. Lastly a brief overview of Histograms of Oriented Gradients (HOG) along with commonly used Support Vector Machine (SVM) structures and architectures are presented.

2.1 Stereo and Thermal Vision

2.1.1 Stereo Vision

The basis of stereo vision is the human eye, for which we use to perceive depth in a 3D space. A traditional stereo vision system consists of two cameras, displaced horizontally from one another, to obtain two different views on the same scene. By comparing these two images, depth information can be extracted and a disparity map can be generated. A disparity map has values inversely proportional to the scene depth at each corresponding pixel location.

The mathematical principle of stereo vision is as follows. Assuming an object, P, is in the scene, where P_1 and P_2 are the corresponding pixel of P in the left and right image, as shown in Figure 2.1.

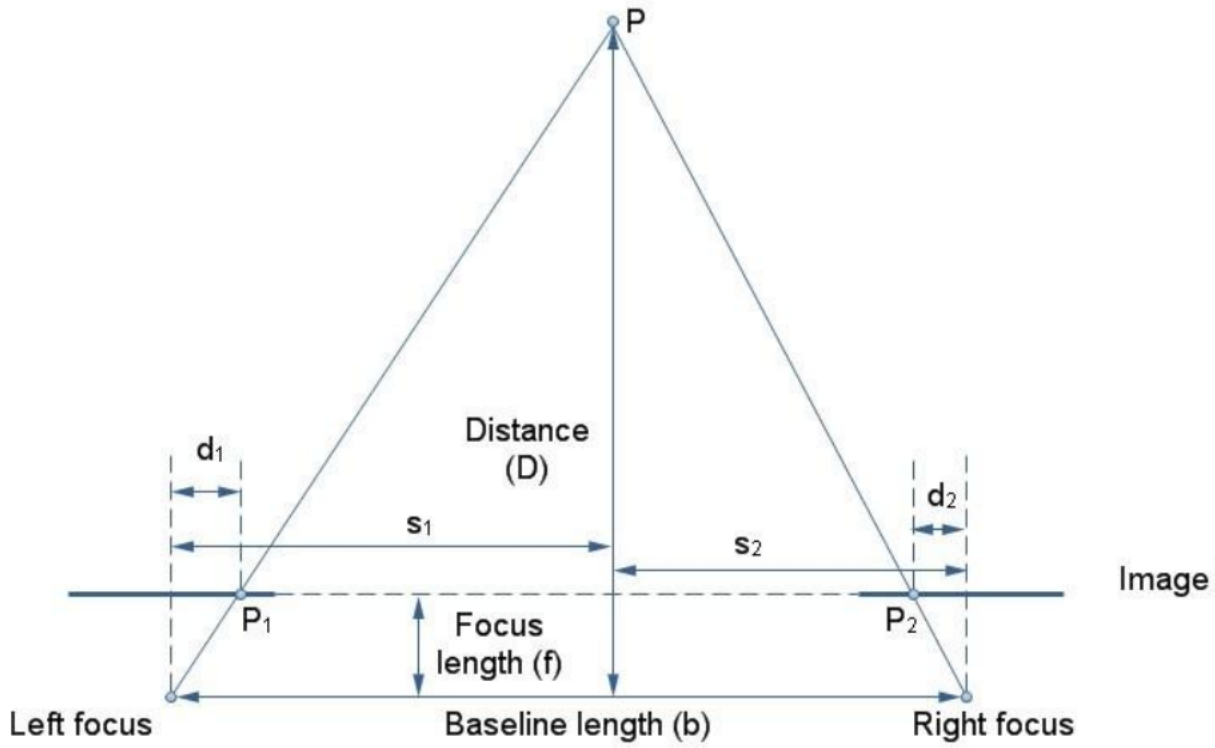

Figure 2.1: Stereo vision mathematical principle

Furthermore, we will also assume that the two cameras are parallel and in the same horizontal plane. Next we will apply the epipolar constraint, which stipulates that since the cameras are in horizontal lines, the epipolar of P_1 and P_2 are also at the same height level. Then based on the geometric principles of similar triangles, the following equation can be derived.

$$\frac{\mathbf{d_1}}{\mathbf{s_1}} = \frac{\mathbf{d_2}}{\mathbf{s_2}} = \frac{\mathbf{f}}{\mathbf{D}} \tag{2.1}$$

By rearranging the equation above, defining disparity d as d_1+d_2, and the baseline

length of the two cameras s as s_1+s_2, the following equations can be derived.

$$\mathbf{s_1} = \frac{\mathbf{d_1 D}}{\mathbf{f}}, \mathbf{s_2} = \frac{\mathbf{d_2 D}}{\mathbf{f}} \tag{2.2}$$

$$\mathbf{D} = \frac{\mathbf{sf}}{\mathbf{d}} : \mathbf{s} = \mathbf{s_1} + \mathbf{s_2}, \mathbf{d} = \mathbf{d_1} + \mathbf{d_2} \tag{2.3}$$

Since the baseline and focal length of our stereo camera system is known and fixed, the distance of the object in the scene can be determined based on the disparity in the two images.

2.1.2 Thermal Vision

Infrared (IR) radiation occurs on a specific wavelength section of the electromagnetic spectrum, more specifically, between 700 nm and 1 mm. This infrared spectrum can then be further subdivided in to five common sections, near-infrared (NIR), short-wavelength infrared (SWIR), mid-wavelength infrared (MWIR), long-wavelength infrared (LWIR), and far-infrared (FIR)[4]. The specific wavelength ranges for these can be observed in Figure 2.2. For the purpose of this paper, LWIR will be the primary focus as it covers the wavelength range of 8-14 µm. This wavelength range corresponds to the typical ambient and human body temperatures. This region is also typically known as thermal infrared[5].

Thermal detectors operate by allowing infrared radiation to heat up the material in which the detector is manufactured. The temperature difference between the material and the background is converted to an electrical signal that would be processed by on board electronics. A micro-bolometer is the specific type of thermal detector used to measure and convert the temperature difference on the material to a difference in electrical resistance to achieve an output signal.

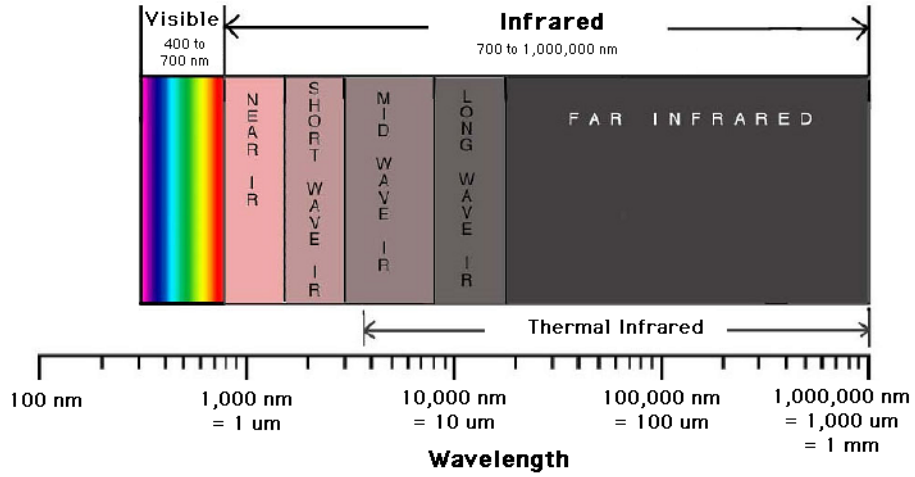

Figure 2.2: Infrared spectrum breakdown

2.2 Trifocal Tensor

The fundamental idea behind trifocal tensor is that when given three views and a pair of matching points in two views along with sufficient information about the placement of the cameras, it is usually possible to determine the location of the point in the third view without reference to image content[6]. To further understand this theoretical technique, the mathematical derivation of the tensor must be looked at in detail.

Trifocal tensor is expressed by a set of three 3×3 matrices, $[\mathbf{T_i}], \mathbf{i} = \mathbf{1}, \mathbf{2}, \mathbf{3}$. It describes the projective geometric relations of image triplets taken from three different cameras. In the case of a view triplet, if the camera matrix of the first view is in the canonical form, $\mathbf{P_1} = [\mathbf{I}|\mathbf{0}]$, then the camera matrices of the other two views can be expressed as $\mathbf{P_2} = [\mathbf{A}|\mathbf{e}']$ and $\mathbf{P_3} = [\mathbf{B}|\mathbf{e}'']$ where $\mathbf{A}$ and $\mathbf{B}$ are 3×3 matrices, and, $\mathbf{e}'$ and $\mathbf{e}''$ are the epipoles corresponding to the image of the first camera on the image plane of the second and third camera respectively. With this, the $3 \times 3 \times 3$ trifocal tensor can be written as:

$$\mathbf{T} = [\mathbf{T_1}, \mathbf{T_2}, \mathbf{T_3}]^T \tag{2.4}$$

$$\mathbf{T_i} = \mathbf{a_i}\mathbf{e''}^T - \mathbf{e'}\mathbf{b_i}^T \tag{2.5}$$

Once the tensor matrix is solved, it can be used in the computation of point and line transfers. If $(\mathbf{l}, \mathbf{l'}, \mathbf{l''})$ is defined as a set of corresponding lines and $(\mathbf{x}, \mathbf{x'}, \mathbf{x''})$ is defined as a set of corresponding points in three images, the transfer function can be represented with the following equations.

$$\mathbf{l^T} = \mathbf{l'^T}[\mathbf{T_1}, \mathbf{T_2}, \mathbf{T_3}]\mathbf{l''} \tag{2.6}$$

$$[\mathbf{x'}]_\times(\sum_{\mathbf{i}} \mathbf{x^i}\mathbf{T_i})[\mathbf{x''}]_\times = \mathbf{0_{3\times 3}} \tag{2.7}$$

where the notation $[\mathbf{x}]_\times$ represents a skew-symmetric matrix of the vector $\mathbf{x} = (\mathbf{x^1}, \mathbf{x^2}, \mathbf{x^3})$ defined as:

$$[\mathbf{x}]_\times = \begin{bmatrix} 0 & -x^3 & x^2 \\ x^3 & 0 & -x^1 \\ -x^2 & x^1 & 0 \end{bmatrix} \tag{2.8}$$

One of the common methods of transfer is by solving the trilinear equations. In tensor notation, $\mathbf{T_i^{ij}}$ denotes the (i,j) entry of the sub-matrix T_i. Thereby an expansion of the point-point-point trilinear equation (2.7) can be written as:

$$\mathbf{x^i}[-\mathbf{x'^3}(\mathbf{x''^3}\mathbf{T_i^{21}} - \mathbf{x''^1}\mathbf{T_i^{23}}) + \mathbf{x'^2}(\mathbf{x''^3}\mathbf{T_i^{31}} - \mathbf{x''^1}\mathbf{T_i^{33}})] = \mathbf{0_{12}} \tag{2.9}$$

This will then lead to a set of nine equations, of which only four are linearly inde-

pendent.

$$\begin{aligned}
\mathbf{x^i T_i^{11}} - \mathbf{x^i x''^1 T_i^{13}} - \mathbf{x^i x'^1 T_i^{31}} + \mathbf{x^i x'^1 x''^1 T_i^{33}} &= \mathbf{0} \\
\mathbf{x^i T_i^{21}} - \mathbf{x^i x''^1 T_i^{23}} - \mathbf{x^i x'^2 T_i^{31}} + \mathbf{x^i x'^2 x''^1 T_i^{33}} &= \mathbf{0} \\
\mathbf{x^i T_i^{12}} - \mathbf{x^i x''^2 T_i^{13}} - \mathbf{x^i x'^1 T_i^{32}} + \mathbf{x^i x'^1 x''^2 T_i^{33}} &= \mathbf{0} \\
\mathbf{x^i T_i^{22}} - \mathbf{x^i x''^2 T_i^{23}} - \mathbf{x^i x'^2 T_i^{32}} + \mathbf{x^i x'^2 x''^2 T_i^{33}} &= \mathbf{0}
\end{aligned} \tag{2.10}$$

This can now be written in a matrix form as $\mathbf{m x^i = 0}$ where $\mathbf{m}$ is a vector of four elements. The vector $\mathbf{m}$ is the most compact representation of the trilinear relationship as only 12 of the 27 tensor entries appear in each of the four equations. When $\mathbf{i}$ has a range of 1 to 3, the linear equations above create a linear system of homogenous coordinates for the point $\mathbf{x}$ in the image.

$$\mathbf{M \times x = 0} \tag{2.11}$$

In the equation above, $\mathbf{M}$ stands for a 4×3 matrix with entries in terms of the tensor $\mathbf{T}$ and a pair of image points (x', x'') between two views. The corresponding point in the first view can be found by solving this linear system as $(\frac{\mathbf{x^1}}{\mathbf{x^3}}, \frac{\mathbf{x^2}}{\mathbf{x^3}})$.

The linear system for the coordinates of point $\mathbf{x'}$ is:

$$\begin{bmatrix} a_{21}^1 & 0 & a_{21}^3 \\ a_{22}^1 & 0 & a_{22}^3 \\ 0 & a_{12}^2 & a_{12}^3 \\ 0 & a_{11}^2 & a_{11}^3 \end{bmatrix} \mathbf{x'} = \mathbf{0} \tag{2.12}$$

where

$$
\begin{aligned}
a_{21}^1 &= -x^1 x''^2 T_1^{33} - x^2 x''^2 T_2^{33} - x''^2 T_3^{33} + x^1 T_1^{32} + x^2 T_2^{32} + T_3^{32} \\
a_{21}^3 &= x^1 x''^2 T_1^{13} - x^2 x''^2 T_2^{13} - x''^2 T_3^{13} - x^1 T_1^{12} - x^2 T_2^{12} - T_3^{12} \\
a_{22}^1 &= x^1 x''^1 T_1^{33} - x^2 x''^1 T_2^{33} - x''^1 T_3^{33} - x^1 T_1^{31} - x^2 T_2^{31} - T_3^{31} \\
a_{22}^3 &= -x^1 x''^1 T_1^{13} - x^2 x''^1 T_2^{13} - x''^1 T_3^{13} + x^1 T_1^{11} + x^2 T_2^{11} - T_3^{11} \\
a_{12}^2 &= -x^1 x''^1 T_1^{33} - x^2 x''^1 T_2^{33} - x''^1 T_3^{33} + x^1 T_1^{31} + x^2 T_2^{31} + T_3^{31} \\
a_{12}^3 &= x^1 x''^1 T_1^{23} - x^2 x''^1 T_2^{23} - x''^1 T_3^{23} - x^1 T_1^{21} - x^2 T_2^{21} - T_3^{21} \\
a_{11}^2 &= x^1 x''^2 T_1^{33} + x^2 x''^2 T_2^{33} + x''^2 T_3^{33} - x^1 T_1^{32} - x^2 T_2^{32} - T_3^{32} \\
a_{21}^1 &= -x^1 x''^2 T_1^{23} - x^2 x''^2 T_2^{23} - x''^2 T_3^{23} + x^1 T_1^{22} + x^2 T_2^{22} + T_3^{22}
\end{aligned}
\tag{2.13}
$$

Finally, the linear system for the coordinates of point $\mathbf{x}''$ is:

$$
\begin{bmatrix}
0 & a_{21}^2 & a_{21}^3 \\
a_{22}^1 & 0 & a_{22}^3 \\
a_{12}^1 & 0 & a_{12}^3 \\
0 & a_{11}^2 & a_{11}^3
\end{bmatrix} \mathbf{x}'' = \mathbf{0}
\tag{2.14}
$$

where

$$
\begin{aligned}
a_{21}^{2} &= -x^{1}x'^{1}T_{1}^{33} - x^{2}x'^{1}T_{2}^{33} - x'^{1}T_{3}^{33} + x^{1}T_{1}^{13} + x^{2}T_{2}^{13} + T_{3}^{13} \\
a_{21}^{3} &= x^{1}x'^{1}T_{1}^{32} + x^{2}x'^{1}T_{2}^{32} + x'^{1}T_{3}^{32} - x^{1}T_{1}^{12} - x^{2}T_{2}^{12} - T_{3}^{12} \\
a_{22}^{1} &= x^{1}x'^{1}T_{1}^{33} + x^{2}x'^{1}T_{2}^{33} - x'^{1}T_{3}^{33} - x^{1}T_{1}^{13} - x^{2}T_{2}^{13} - T_{3}^{13} \\
a_{22}^{3} &= -x^{1}x'^{1}T_{1}^{31} - x^{2}x'^{1}T_{2}^{31} - x'^{1}T_{3}^{31} + x^{1}T_{1}^{11} + x^{2}T_{2}^{11} + T_{3}^{11} \\
a_{12}^{1} &= -x^{1}x'^{2}T_{1}^{33} - x^{2}x'^{2}T_{2}^{33} - x'^{2}T_{3}^{33} + x^{1}T_{1}^{23} + x^{2}T_{2}^{23} + T_{3}^{23} \\
a_{12}^{3} &= x^{1}x'^{2}T_{1}^{31} + x^{2}x'^{2}T_{2}^{31} + x'^{2}T_{3}^{31} - x^{1}T_{1}^{21} - x^{2}T_{2}^{21} - T_{3}^{21} \\
a_{11}^{2} &= x^{1}x'^{2}T_{1}^{33} + x^{2}x'^{2}T_{2}^{33} + x'^{2}T_{3}^{33} - x^{1}T_{1}^{23} - x^{2}T_{2}^{23} - T_{3}^{23} \\
a_{11}^{3} &= -x^{1}x'^{2}T_{1}^{32} - x^{2}x'^{2}T_{2}^{32} - x'^{2}T_{3}^{32} + x^{1}T_{1}^{22} + x^{2}T_{2}^{22} + T_{3}^{22}
\end{aligned}
\tag{2.15}
$$

2.3 HOG and SVM

One of the most popular pedestrian detection methodologies is an approach which combines Histograms of Oriented Gradients and Support Vector Machines.

2.3.1 HOG

HOG is a type of feature descriptor. The purpose of a feature descriptor is to generalize an object in such a way that the same object produces as close as possible to the same feature descriptor when viewed under different conditions[7]. It was first introduced in 2005 by Dalal and Triggs and has received a lot of notice from researchers. This is due to the fact that local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions and therefore provide promising results in pedestrian detection. A lot of

variations exist but the basis of the method is described below.

First the centered and horizontal gradients are computed with no smoothing. Then the gradient orientation and magnitude is computed. If it is a color image, the color channel with the highest gradient magnitude is chosen for each pixel. Next the image is divided into small sub-images known as cells. The cells can be rectangular or circular. Then the image is divided into larger regions or "blocks" that consists of a number of cells. The blocks are generated with overlap so that a cell can belong to more than one block.

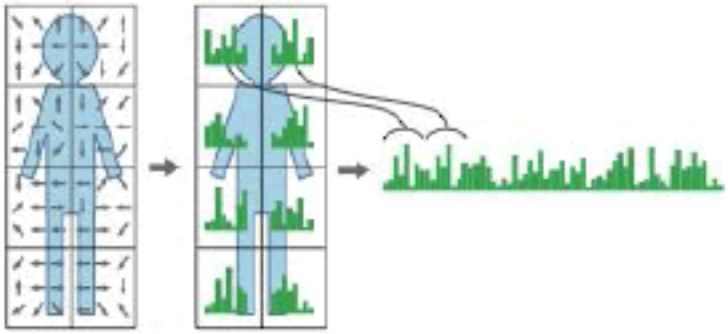

Figure 2.3: HOG calculation process visualized

The edge orientations within each cell are then accumulated in a histogram. Next the histogram entries are combined and used as the feature vector describing the object. Lastly, normalization of the cells across larger regions, multiple cells, provides for a better illumination invariance[7]. The process can be visualized in Figure 2.3.

2.3.2 SVM

Machine learning is about learning structure from data. Linear Support Vector Machine is a popular classifier. Two sets of points are linearly separable if they

can be separated by a single line in two dimensions and by a hyper-plane in more than two dimensions[8]. It should be noted that multiple separating hyperplanes may be created based on a set of training data as seen in Figure 2.4. If they cannot be separated by a line or in general hyper-plane they are said to be not-linearly separable, but can still be handled by a kernel trick in order to transform the non-linear classifier into a classifier that can work with a higher dimensional feature space. Support Vector Machine not only classifies the patterns it also optimizes the decision boundary based on maximizing the margin that separates the instances as seen in Figure 2.5.

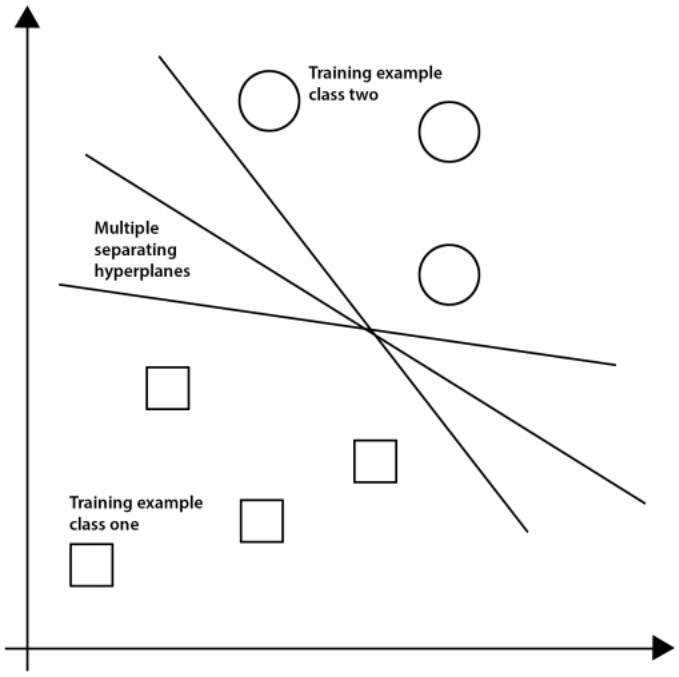

Figure 2.4: Examples of many possible separating hyperplanes

The structure of the decision-making model can take on a variety of designs. One of the most basic methods is to concatenate different types of features into a single SVM for training. This basic methodology generally work well given a sufficient set

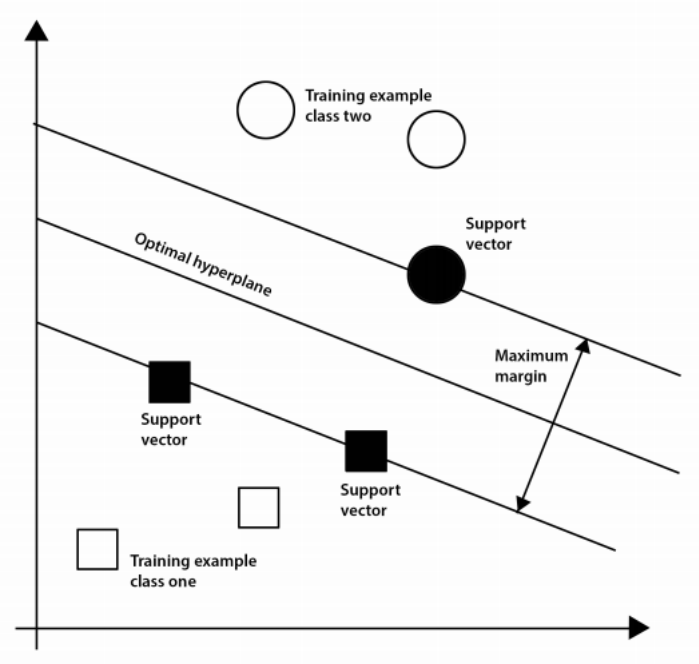

Figure 2.5: Optimized hyperplane based on the maximum margin separating the instances

of data.

Another option is to use a cascaded classifier SVM without feedback. This classifier is a multi-stage system. Cascaded SVMs may suffer from loss of information during each stage, similar to dimensionality reduction. Cascaded classifiers can be primarily used for pre-filtering or selection, in order to save on computational cost. Again this classifier results in errors that may accumulate during each stage.

To address the issue of error accumulation in the previous method, parallel cascaded SVMs with feedback structure can be used. With the feedback during training, this would eliminate the accumulation of previous stage errors. As it can be seen in Figure 2.6, the training data were split into subsets and each one is evaluated individually for support vectors in the first layer[9]. The results are combined two-by-two and entered as training sets for the next layer. Finally the resulting support vectors were tested for convergence by feeding the result of the last layer

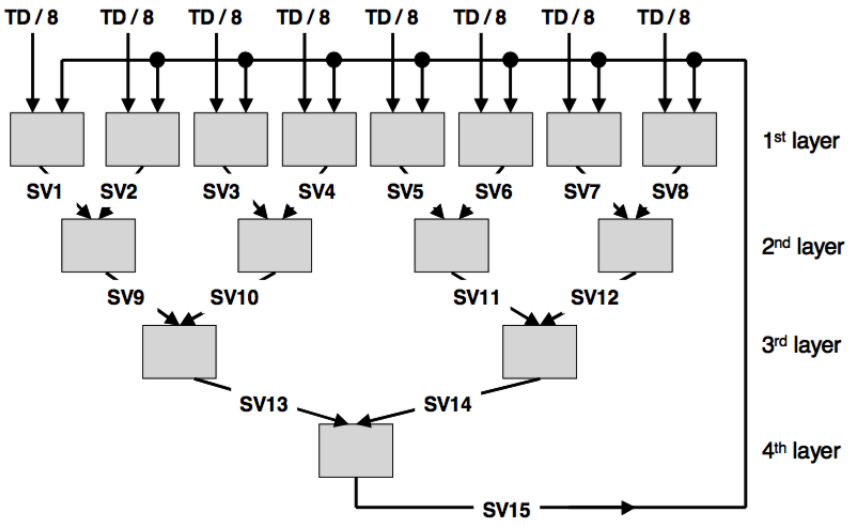

Figure 2.6: Schematic of cascaded SVM with feedback architecture

back into the first layer, along with the non-support vectors.

This classifier is only good for training, not for decision-making. The only useful SVM is the last stage. SVMs from previous layers merely give selected support vectors to the next SVM during training. Therefore they are not used in decision-making. Furthermore this classifier may pose an issue if the different features are of different dimensions. For example a 1000-dimensional HOG feature might render a 1-dimensional mean intensity feature insignificant, though both could arguably be just as useful.

Chapter 3

Related Work

This chapter will focus on related works in pedestrian detection. Trying to detect pedestrians in a traffic scene is not a trivial task. In the last decade, extensive efforts have been made to improve the performance of pedestrian detection.

3.1 Background Subtraction Method

An early work by Xu *et al.*[10] showcased a fast and effective pedestrian detection using HOG descriptor based SVM. This approach takes advantage of CodeBook background subtraction (CBBS) to produce pedestrian samples for SVM. The HOG features of the samples were extracted to train both linear and Radial Basis Function (RBF) SVM classifiers offline. The process flow chart can be seen in Figure 3.1. The group carefully investigated the influence of various ratios of positive and negative training sets on detectors performance. Furthermore they compared Linear and RBF SVM in experiments as well. It was concluded that their methodology obtained reliable detection results and was robust against pedestrian appearance and pose variations, illumination changes, background changes, shadows and etc.

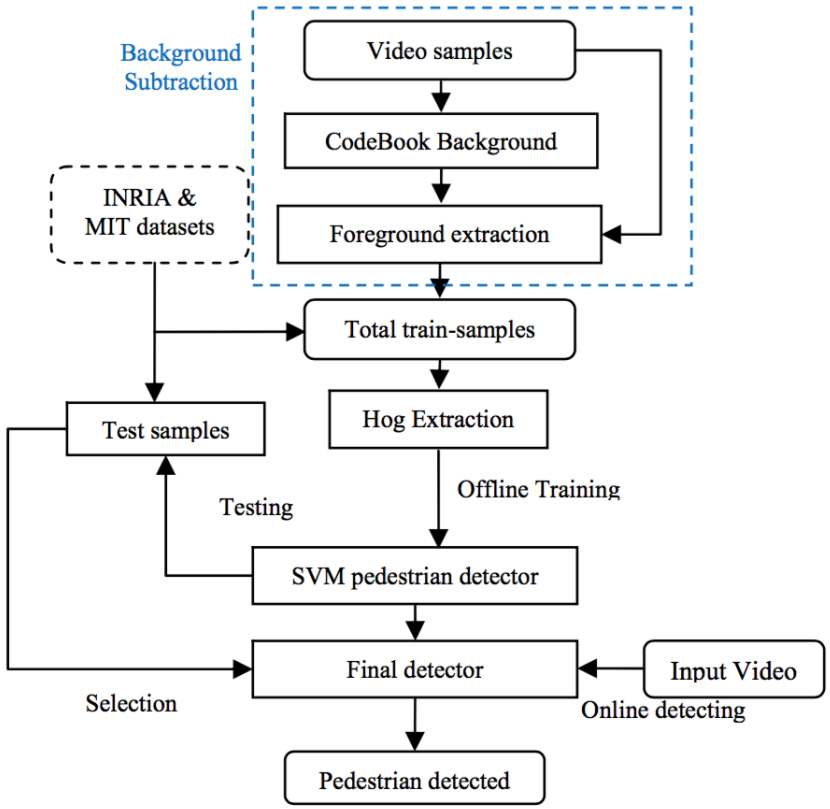

Figure 3.1: Process flow of background subtraction methodology

3.2 Color Feature Extraction Method

Van de Sande *et al.* proposed using color descriptors as an extractable feature for the purpose of object and scene recognition[11]. Different combinations of features were evaluated, and the group introduced a new feature based on the similarity of colors in different regions of the detector window. This showed a significant improvement in detection performance. It was discovered that people do exhibit some structure, in that colors are locally similar. For example, the skin color of a specific person is similar on their two arms and face, and the same is true for most peoples clothing. Therefore, color self-similarities can be encoded within the descriptor window, i.e., similarities between colors in different sub-regions.

3.3 Color and Infrared Stereo Vision

One of the most notable related works was by Krotosky and Trivedi[12] in which an analysis of color and infrared stereo approaches to pedestrian detection was explored.

A four-camera experimental test bed consisting of two color and two infrared cameras was created to allow for synchronous capture and direct frame-by-frame comparison of pedestrian detection approaches. Two pairs of color and infrared cameras were arranged so that their imaged scenes were as consistent as possible. The two pairs had identical baselines and the corresponding cameras in the color and infrared pairs are positioned as close as possible so as to maintain the same approximate fields of view. Additionally, lenses for the color cameras were selected to best match the fixed zoom of the infrared cameras. All four cameras are arranged in a single row and care was taken in aligning the pitch, roll and yaw of the cameras to maximize the similarity in field of view.

The group conducted experiments such that pedestrians walked in front of the test bed. The experiments included multiple pedestrians in the scene with varying degrees of depth, complexity, and occlusion. The experimental data was captured simultaneously with the color and infrared stereo cameras.

The fundamental algorithm utilized by this group to analyze pedestrians with stereo imagery was to detect obstacles in the scene and localize their position in 3D space from the disparity maps generated from stereo correspondence matching. In essence, the disparity images derived from stereo analysis was used to generate a list of candidate pedestrian regions in the scene.

3.4 Decision Fusion

For any pattern classification task, when there is an increase in data size, number of classes, dimension of the feature space, or interclass separability, the performance of any classifier will be affected. Typically, a single classifier is unable to handle the wide variability and scalability of the data in any problem domain. Therefore, most modern techniques of pattern classification use a combination of classifiers to fuse the decisions for the task. The problem of selection of a useful set of features and discarding the ones which do not provide class separability are addressed in feature selection and fusion tasks. A survey of these methods were presented by Mangai *et al.*[13].

Chapter 4

Data Acquisition and Processing

This chapter will discuss the process of real world data acquisition, from equipment selection to test rig construction. Furthermore, we will discuss the techniques employed to create pixel level data fusion as this is important to the task of machine learning.

4.1 Data Collection Equipment

For the purpose of this work, a custom test equipment rig was designed and assembled in order to collect real world data for pedestrian detection. This was necessary as there does not current exist an open source data set that our methodology and algorithm can be applied to. Our design constraint required that our system be mobile, easily mountable/dismountable to our test vehicle as well as maintaining calibration between data collection runs.

4.1.1 Stereo Vision Camera

The ZED StereoLabs, stereo vision camera was chosen for providing vision data as well as depth data.

Figure 4.1: StereoLabs ZED camera

This camera was chosen for multiple reasons. The ZED camera can capture high resolution side-by-side video on USB 3.0 that contains two synchronized left and right video streams, and can create a depth map of the environment in real-time using the graphics processing unit (GPU) of the host machine[14]. Furthermore, StereoLabs provides the end user an easy to use SDK that allows for access to commonly use camera controls and camera outputs with a wide array of operation modes. One of the most important factors in choosing the ZED camera is that the on board cameras are pre-calibrated and comes with known intrinsic parameters. This provides for proper image rectification and disparity map generation.

4.1.2 Thermal Camera

For our LWIR camera, the FLIR Vue Pro, as seen in Figure 4.2, was chosen for its cost to performance ratio. The FLIR Vue Pro is an uncooled vanadium-oxide (VOx) microbolometer touting a 640x512 resolution at a full 30Hz and paired with

a 13mm germanium lens providing a 45° × 35° field of view[15]. This IR camera has a wide -20 to 50 °C operation range which will allow for rugged outdoor use. The Vue Pro provides for Bluetooth wireless control and full data recording of thermal video via its onboard microSD card as well as a real-time analog video output.

Figure 4.2: FLIR Vue Pro LWIR camera

At the time of this work, this thermal camera provides for the highest resolution to cost ratio. While it is possible to obtain full digital infrared radiometric data from the Tau 2 core inside the FLIR Vue Pro, it becomes cost ineffective, but may be considered in the future if necessary.

4.1.3 Enclosure

Both the stereo vision and LWIR camera must remain fixed relative to each other for repeatability between data collection sequences. A $\frac{1}{4}$-20 threaded rod was custom cut to length and each end was threaded into the respective cameras tripod mounting hole. This provided for a rigid connection between the vision and the IR camera. Next the housing for the camera was sourced. A Carlon electrical junction box

was utilized as this was an appropriately sized, water proof box that provided high impact resistance. The top lid was replaced with a Lexan impact resistant clear acrylic sheet such the stereo vision cameras can be situated safely behind it. A circular hole was also cut into the top lid to accommodate for the thermal camera lens to fit through and mounted via the lens barrel. This was required, as even clear acrylic would block most if not all IR radiation in our spectrum of interest.

4.1.4 Mounting System and Cable Management

Finally a mounting system was designed, modeled, and built utilizing 80/20 aluminum extrusions. This structural frame provided the adequate permanent mounting points for our camera housing, as it was secured to the channels in the aluminum extrusion by numerous screws. The entire structure is now completely portable and can be mounted to any vehicle with a ski rack. The 80/20 extrusions would sit between the front and back ski rack hold-downs.

Cable management was crucial our design as long runs of cables were needed in order for communication between our laptop inside the vehicle and cameras on the roof. In order for the system to operate in adverse conditions, the cables must run down the back of the vehicle, through the trunk and into the vehicle cabin. This equated to approximately 20 feet of cable.

This created an issue for our ZED stereo vision camera, as it operated on high speed USB 3.0 protocol that allowed for a 10 feet maximum length before signal degradation and loss[16]. To resolve this issue, an active USB extension cable was used. This USB cable also required a 5V power on the female end. A two-conductor wire was soldered to it and ran parallel to the USB extension cable terminating in another USB connector for easy 5V access.

The FLIR Vue Pro IR camera utilizes a 10pin mini-USB port for power, data,

and analog video output. A special cable, as seen in Figure 4.3 was provided by FLIR that terminated in 10-pin mini-USB port and the other end terminated in a standard USB plug and composite analog plug.

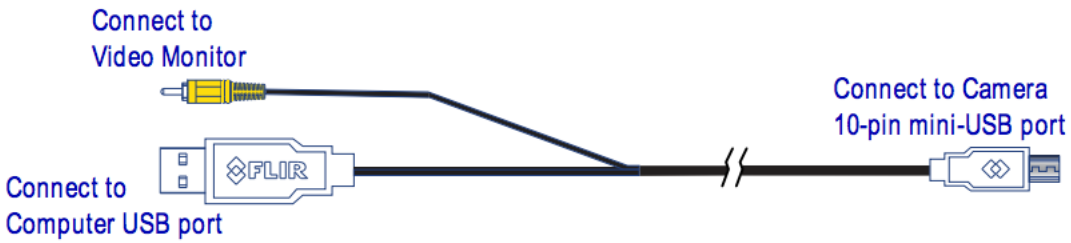

Figure 4.3: FLIR Vue Pro cable

The USB cable was further extended by 20 feet, this was acceptable as it was only USB 2.0 and did not require high-speed data transfer. The same was done with the composite analog video cable.

A total of four cables terminated from the camera setup, they were all wrapped together with braided cable sleeves to prevent tangling, ensure robustness and created an ascetically pleasing cable bundle.

4.1.5 Completed System

The completed system can be seen in Figure 4.4. Figure 4.5 shows the system having been mounted to the roof rack of a test vehicle. The data cable bundle has been routed down the back window and through the trunk of the vehicle to a laptop.

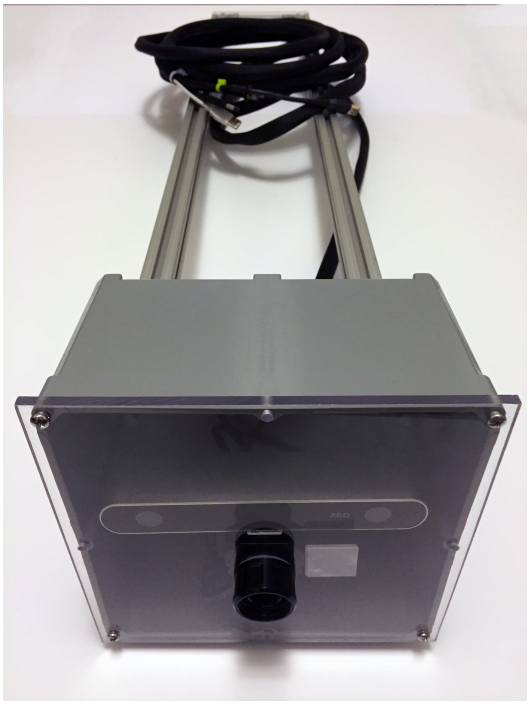

Figure 4.4: Standalone completely assembled system

Figure 4.5: Completed system mounted on test vehicle

4.1.6 Software

Software was developed in C++ in accordance to the ZED SDK in order to capture visual data. Furthermore, software was developed to convert the native raw SVO file format provided by the ZED camera into individual image sequences at 30 frames per second (FPS). These images are post rectified and can be used directly in our subsequent steps. The SDK also provided for the ability to generate raw disparity data. The SDK utilizes a proprietary disparity calculation process much like that of semi-global matching to generate the raw disparity data. This system relies on having a powerful Nvidia GPU and CUDA in order perform these calculations.

An analog frame grabber was employed to capture the real-time analog output of the IR camera instead of directly recording to the onboard microSD card. This was the solution to a problem of synchronization between the thermal camera and stereo vision camera. With analog frame grabber, we were able to precisely capture at 30 FPS. AVI files were generated using software provided along with the frame grabber. These AVI files were then converted into image sequences via Matlab.

Approximately a total of three hours of driving data was captured across multiple days and lighting conditions. This does not include any highway driving as we are unlikely to see any pedestrians on the highway. We paid close attention to capturing situations in which visual sequences are extremely difficult to see where the thermal images were very clear and also cases where thermal images are likely to fail where visual images are clearly distinguishable. This resulted in a mere 58 usable sequences totaling 4320 frames in which a person or multiple people are in clear view and un-occluded. This may seem to be a small data set given the amount of original data, but we have excluded situations where there may be high levels of noise and occlusion. Furthermore, video frames without any pedestrian were also not included in the data set.

4.2 Application of Trifocal Tensor

In order to perform data and decision fusion, proper data alignment is necessary. We first performed image fusion at a pixel level. This in essence mapped every pixel on each of our three view angles to one another to which the images were a perfect overlay on top of one another. This task is non-trivial as not only did the cameras not have the same optical field of view, it is actually physically impossible to simply stack three images taken from three different point of view and match them pixel for pixel. As seen in Figure 4.6, by simply stacking the images on top of one another, we are able to only match the foreground parts of the image where the pedestrian is, but the rest of the image is mismatched.

Figure 4.6: Simple image stacking

Since the ZED stereo vision system provides a left (Figure 4.9a) and right (Figure 4.7b) image, along with the calculated disparity map, we essentially know the point to point correspondence of two point of views. Now we simply need to map the third view, thermal imaging (Figure 4.8) to the first two. Standard methods such as semi-global matching techniques will not work in this situation as we are dealing with cross spectral imaging in which equivalent points will look completely different.

(a) Left image (b) Right image

Figure 4.7: Sample images from stereo vision camera

To solve this issue, we utilized the mathematical theory of Trifocal Tensor as discussed in section 2.2 for this application. Although there are many methods for calculation of the tensor matrix[6], the most optimal one for our application is a linear method based on direct solution of a set of linear equations.

According to equation (2.11), each point-point-point correspondence provides four linearly independent equations. Thereby seven corresponding points across the three views can uniquely determine the 27 entries of the tensor matrix. Given n point correspondences, let $\mathbf{A}$ denote a matrix of size $4n \times 27$ and $\mathbf{t}$ a vector containing all entries of the tensor, we can write

Figure 4.8: Sample thermal image of LWIR camera

$$\mathbf{At} = \mathbf{0} \tag{4.1}$$

Then the tensor can be obtained by the Singular Value Decomposition (SVD) solution to this linear system.

Hartley has shown that this kind of algorithm does not do well if all points are of the form (x_1, x_2, 1) in homogeneous coordinates with x_1 and x_2 very much larger than 1[17]. Therefore, it is necessary to normalize the points of each image separately before computing the tensor and it is necessary to de-normalize the computed tensor in order to work with the original image coordinates. In summary, the normalized linear algorithm for computation of tensor $\mathbf{T}$ given $n \geq 7$ image point

correspondences across 3 images is as follows.

1. Normalize the set of point triplets by performing transformations $\mathbf{H}$, $\mathbf{H}'$ and $\mathbf{H}''$.

2. Transform the points according to $\mathbf{x^i} \mapsto \hat{\mathbf{x}}^\mathbf{i} = \mathbf{H^i_j x^j}$

3. Compute the tensor linearly by solving a set of equations of the form $\mathbf{At} = \mathbf{0}$, where $\mathbf{A}$ expresses the equation (2.7) and $\mathbf{t}$ is the vector of entries of tensor.

4. De-normalize the tensor by $\mathbf{T_i} = \mathbf{H'^{-1}} \mathbf{\Sigma_j} (\mathbf{H^T}(\mathbf{i}, \mathbf{j}) \mathbf{T_j}) \mathbf{H''^{-T}}$

In order to reconstruct the third image base on the tensor matrix, we will also need to calculate the fundamental matrix $\mathbf{F_{21}}$ between the left and right image. The fundamental matrix is calculated utilizing the normalized eight-point algorithm[18]. This function was readily available through Matlab and therefore was not re-derived.

Now that we know both the fundamental matrix $\mathbf{F_{21}}$ and the tensor matrix $\mathbf{T}$, we can relate the $\mathbf{F_{21}} = [\mathbf{e}'] \times [\mathbf{T1}, \mathbf{T2}, \mathbf{T3}]\mathbf{e}''$.

Using this computation method, we implemented the algorithm in Matlab in order to reconstruct our thermal image such that it would perfectly overlay with our left vision image. An example of our reconstructed image of the original Figure 4.8 can be seen in Figure 4.9b. We have also overlaid the original left vision image and the now reconstructed thermal image and show that they are a perfect match for every part of the image. This anaglyph can be seen in Figure 4.10.

(a) Left vision image (b) Reconstructed thermal image

Figure 4.9: Reconstruction based on trifocal tensor algorithm

Figure 4.10: Anaglyph overlay of reconstructed thermal image and original left vision image

Chapter 5

Data Extraction

This chapter discusses the processes necessary to extract the necessary data from our raw data set such they can be use for SVM training.

5.1 Ground Truth Data Labeling

For supervised machine learning, results obtained from a given algorithm must be compared against target ground truth; a set of results determined a priori to be correct[19]. Generating ground truth outputs is called labeling data.

In order to assist in labeling a vast amount of frames and images, the Video Processing Analysis Resource (ViPER) tool is used. The tool allows for import of folders containing image sequences derived from video streams. Once we create an OBJECT ID, such as "pedestrian", we can draw a rectangular bounding box around the person of interest in such frame. The following sequences follow suit. Once all the images in a given sequence have been appropriately labeled, the label data can be saved as an extensible XML based file format. An example of this tool interface can be seen in Figure 5.1.

We choose to label only the left vision image sequences as this was sufficient

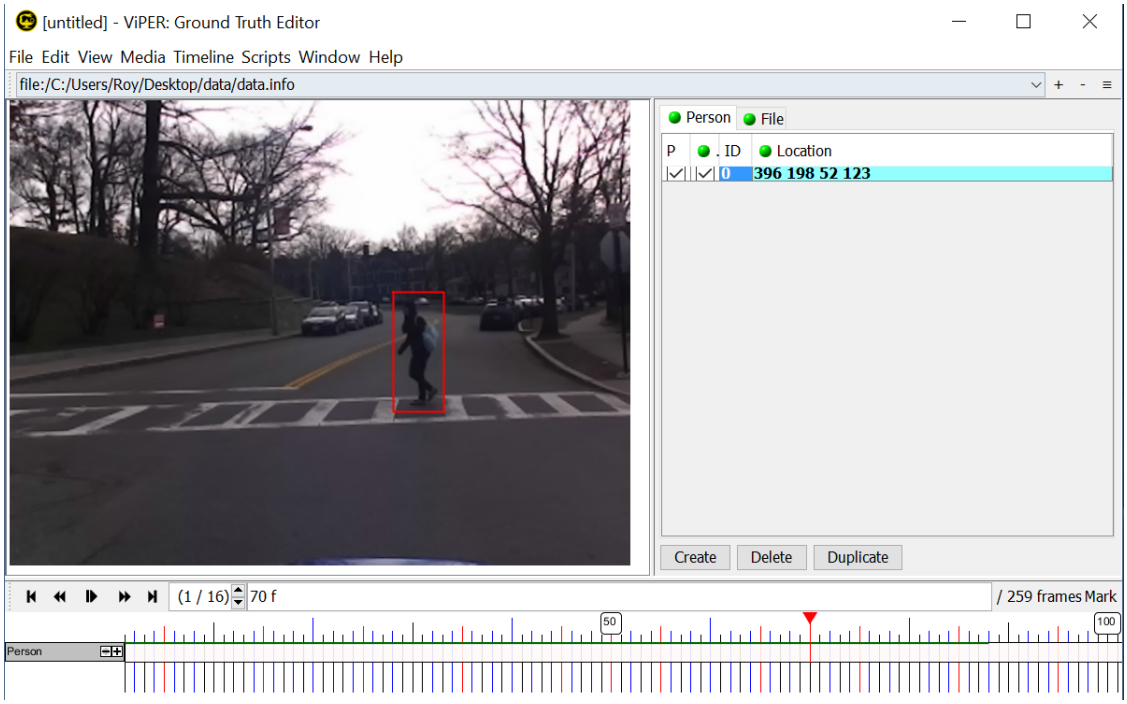

Figure 5.1: ViPER tool interface

since the thermal image sequences have been transformed and reconstructed such that the respective vision label data can be applied.

The labeled data will be then separated into two distinct groups; one group will be use for SVM training, and the other group will be used for testing the SVM. The groups will be separated by sequences and not just by frames. Since each sequence will have frames that are very similar to one another, this will ensure that the training data does not too closely resemble the testing data. The training data set consists of 39 sequences equating to 3417 frames. The testing data set consists of 19 sequences equating to 913 frames.

5.2 Positive and Negative Sample Extraction

Matlab was utilized to extract both positive and negative samples for the training data set. Using the XML data from ground truth labeling, we were able to crop out every positive sample from the training data set. This resulted in approximately 3950 positive samples. Some of these positive samples can be seen in Figure 5.2. Each roll of positive vision sample have its corresponding row of positive thermal sample. This further demonstrates that our trifocal tensor technique works as desired.

Figure 5.2: Positive samples in multispectral images

The negative samples are extracted using two different methodologies. First a set of negative samples are randomly selected from the training data set. These images may contain segments of trees, buildings, vehicle, and etc. They are obtained by

applying a random candidate selection procedure over the pedestrian-free areas, as denoted by the ground truth, in the images. We extracted an approximately 1:10 ratio of positive to negative samples.

Next, a set of hard negative samples were extracted. This was done along side the classification phase of this work. To improve the SVM accuracy, the SVM was trained and validated with the same training data set. Any false detections were then extracted and also set as negative examples, this process is called mining for hard negatives. By iterating this process, the accuracy of the SVM classification can be improved. We performed two iterations after the initial training, we did not perform any further iterations as it no longer provided any increased performance. In the end, we have built a collection of approximately 56,000 negative samples. Some of these negative samples can be seen in Figure 5.3.

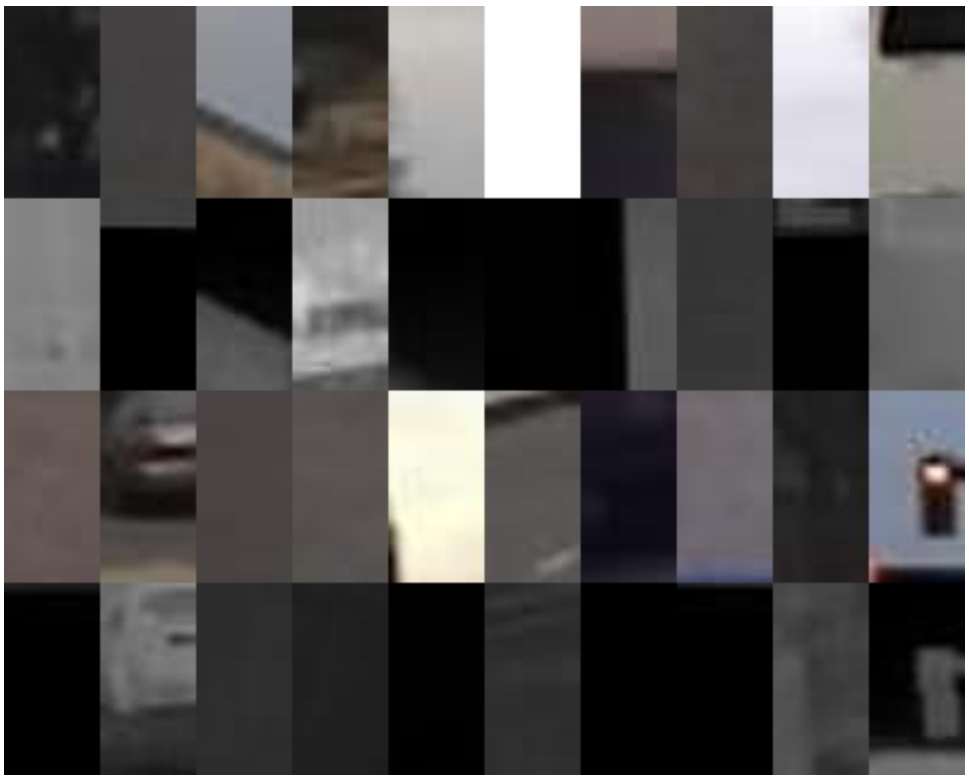

Figure 5.3: Negative samples in vision images

Chapter 6

Pedestrian Classification and Detection

This chapter is focused on methods for improving classification and detection of pedestrians by data and decision fusion.

6.1 Classification

The preliminary classification methodology that we performed provides a baseline for which we built our successive methods. This initial first pass classification simply took the extracted raw images from the left side of the stereo vision camera and performed SVM classification. Next we also performed the same style of classification with the raw images from the thermal camera. Lastly, the raw images from both the vision and thermal camera were combined and passed to the SVM for classification training. A seven-fold cross validation was also performed with the data set. The purpose of a cross validation is to assess how the results of a statistical analysis will generalize to an independent data set[20]. The results of this can be seen in Ta-

Table 6.1: Cross validated SVM classification of raw image data results

	Precision	Recall
Left Vision Only	77.23%	47.60%
Thermal Only	95.95%	88.65%
Left Vision & Thermal	98.24%	83.83%

Table 6.2: Cross validated SVM classification of HOG data results

	Precision	Recall
Left Vision Only	95.47%	75.28%
Thermal Only	98.53%	85.21%
Left Vision & Thermal	98.91%	85.09%

ble 6.1. In the table, precision is defined as $\frac{\mathbf{TruePositive}}{\mathbf{TruePositive+FalsePositive}}$, whereas recall is defined as $\frac{\mathbf{TruePositive}}{\mathbf{TruePositive+FalseNegative}}$.

Although, the results are numerically insignificant at this point, the trend is promising. We can see that the thermal images alone are significantly better performing than visual data alone. Furthermore, it can be argued that simply combining the two SVMs of Thermal and Left provides a slight improvement in precision. Although this is easily offset by the decrease in performance of recall, but that could be attributed to the extremely poor recall of vision only. We believe this is due the low quality of the visual image data, which is a limitation of our instrumentation. In the next step, we leveraged the capabilities of HOG as a feature descriptor in order to improve our classification training.

We performed the same classification training methodology as outline in the initial baseline run, except this time, instead of simply providing the SVM with the raw image data, we preprocessed the data through HOG and provided the SVM with the histogram data. Again a seven-fold cross validation was performed against the data set. The results can be seen in Table 6.2. While analyzing these results, we can see that the thermal imaging provides for what seems to be a better result on its own. These classification results should not be a direct inference on the actual

performance of this system during detection. This is because during detection, not only does positive samples need to be correctly classified but an exponentially larger number of negative data will need to be properly classified. This will result in a significant drop in actual performance. Therefore decision fusion becomes necessary.

In the next phase of this work, we have performed an intelligent data/decision fusion technique. As discussed in Section 2.3.2, a multilayer SVM architecture passes the result of the previous layer to the next layer. These results are in a binary form of yes or no. We believe this results in a significant loss of useful information. Instead we propose the use of the classification score from first layer SVMs to train the second layer SVM.

The SVM classification score for classifying observation $\mathbf{x}$ is the signed distance from $\mathbf{x}$ to the decision boundary ranging from $-\infty$ to $+\infty$[21]. A positive score for a class indicates that $\mathbf{x}$ is predicted to be in that class, a negative score indicates otherwise. Furthermore, a large value represents a higher level of confidence whereas a lower score indicates higher uncertainty. For the rest of this work, we will refer this classification score as our confidence score.

This methodology will over come some of our instrumentation limitations where the feature may be unclear in one spectra yet very evident in a different spectra. Therefore this style of decision fusion will yield a better result. Based on an overall survey of our data set, we have determined the following ranking for best to worst overall feature confidence score.

1. Thermal Imaging
2. Visual Imaging
3. Disparity Map

We have also plotted (Figure 6.1) each point in a given image frame to visualize and

compare the confidence result of the first layer of SVM for two different features.

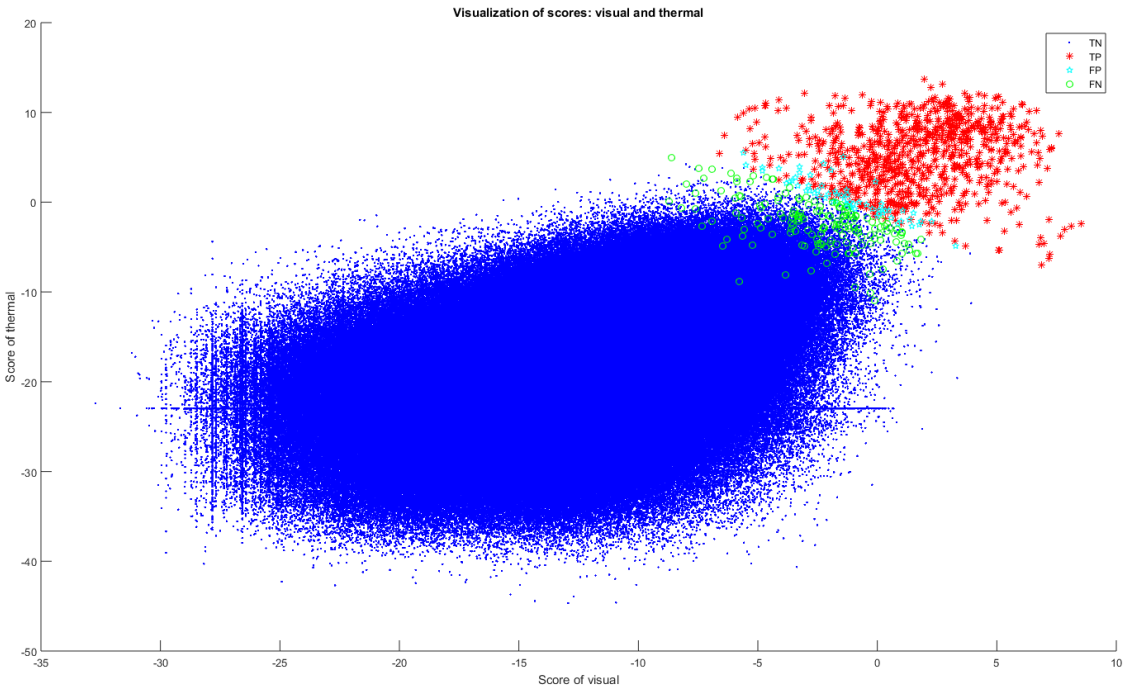

Figure 6.1: SVM feature confidence comparison of visual and thermal

The x-axis represents the confidence score of the visual image and the y-axis presents the confidence score for the corresponding thermal image. We can see that there is no distinct and clear separation of the positive samples (red, TP) vs the negative samples (blue, TN). There is an area in the middle where there are a number of false positives (cyan, FP) along with misses (green, FN). This further validates and shows the importance of our approach.

A similar plot of confidence was created for all three features of visual, thermal and disparity. We can see that there is still no clear and distinct separation of positive and negative samples.

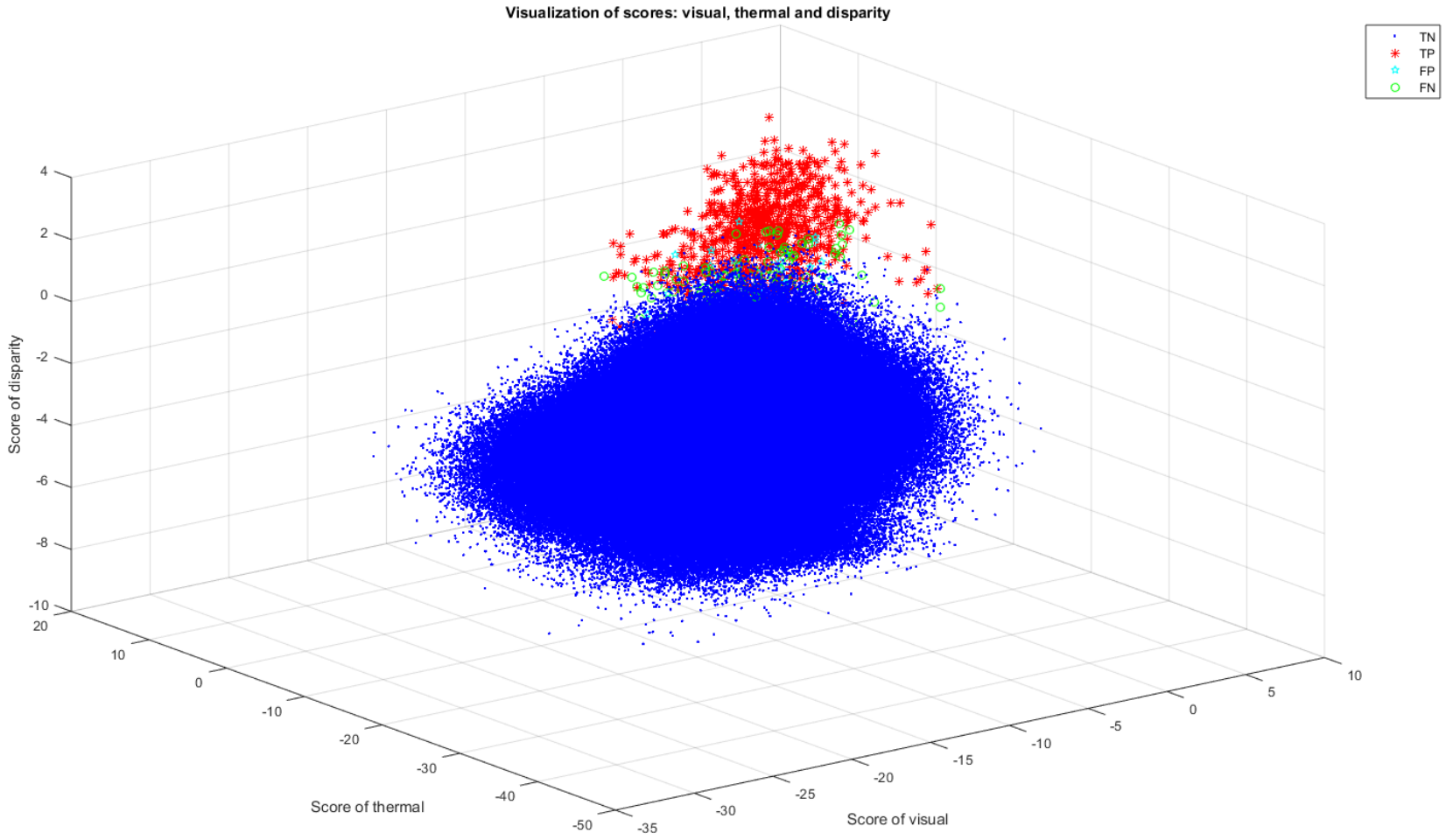

Figure 6.2: SVM feature confidence comparison of visual, thermal and disparity

6.2 Detection

In this section, we will focus on computer vision algorithms for detecting pedestrians. We begin with an overview of pedestrian detectors and examining some of the ideas introduced in detection in the last decade and how those works ultimately shaped our design.

A common methods for moving object detection proposed in the literature has been background subtraction. The goal of background subtraction is to "remove" the background in a scene by describing an adequate model of the background[10]. This results in only objects of interest left in the scene for tracking and further analysis. This technique is generally low in computational cost. However, the major flaw in this technique is the requirement of a stationary camera. Unfortunately, the cameras in our application are vehicle mounted and are non-stationary. Therefore, this method could not be applied to a moving camera situation since the background is always changing.

Color based feature extraction used to determine regions of interest, ROI, is another frequently discussed method in moving object detection. This feature extraction scheme essentially computes implicit "soft segmentations" of image regions into foreground/background[22]. Unfortunately, color by itself is of limited use, because colors vary across the entire spectrum both for people's clothing and and for the background. This creates an unsolved problem of color consistency. Therefore color information could be a good feature to use in image classification but could not be applied to detection[23].

As we have shown in our classification results, HOG features in thermal imaging alone was a strong classifier which suggests that it might be a good feature to use in detection. One could argue that the human body remains in a relatively stable

temperature range of 36.5-37.5 °C, and therefore should be a feature to utilize. The flaw in this is that though the human body core temperature remains relatively stable, the surface temperature radiated through the infrared spectrum varies wildly. Furthermore, the human body temperature often blends into the wide variety of background in a scene. This affect can be seen in Figure 6.3. Therefore using the HOG features in thermal imaging alone would not be optimal for ROI selection in detection.

Figure 6.3: Pedestrian blending into the background

Another novel approach, as described in Section 3.3, is to utilize the disparity map generated from either the stereo vision camera or the stereo thermal camera to determine a ROI. This approach is far from optimal or robust as it requires either the pedestrians to be extremely close to the test bed or for ultra high-resolution color and infrared cameras to be utilized. Neither of these are feasible at this time or cost effective. Pedestrian detection at ranges of 5 meters or less becomes ineffective

as avoidance becomes ever more challenging and collision inevitable. Furthermore, utilizing ultra-high resolution cameras creates an immense amount of data, which leads the inability to process in real time. Lastly the cost of high-resolution infrared cameras is still at an exuberant price, making a stereo infrared system unaffordable.

Train a classifier on $n \times m$ image windows. Positive examples contain the object and negative examples do not.
Choose a threshold t and steps Δx and Δy in the x and y directions.

Construct an image pyramid.

For each level of the pyramid
  Apply the classifier to each $n \times m$ window, stepping by
  Δx and Δy, in this level to get a response strength c.
  If $c > t$
    Insert a pointer to the window into a ranked list $\mathcal{L}$, ranked by c.

For each window $\mathcal{W}$ in $\mathcal{L}$, starting with the strongest response
  Remove all windows $\mathcal{U} \neq \mathcal{W}$ that overlap $\mathcal{W}$ significantly,
    where the overlap is computed in the original image by expanding windows in coarser scales.

$\mathcal{L}$ is now the list of detected objects.

Figure 6.4: Sliding window detection algorithm[1]

In this work, we propose the sliding window detector method which creates numerous windows for a given frame. Each of the window is then run through the classifier to determine if there is an instance of a pedestrian. In order to detect pedestrians at multiple scales the sliding window detector is also ran at multiple scales. The sliding window approaches appears most promising for low to medium resolution settings, under which segmentation based methods often fail. The sliding window algorithm can be seen outlined in Figure 6.4.

After implementing the sliding window algorithm in Matlab, over 90,000 win-

dows per image were created. This detection problem now has essentially become a classification problem. After applying our multilayered SVM technique using confidence scores against our entire testing data set, we were able to achieve the following results in Table 6.3. In this table, FPPI stands for false positives per image.

Table 6.3: Detection results based on multilayer & confidence score SVM approach

	Detection Rate	**Miss Rate**	**FPPI**
Visual	64.16%	35.85%	16.68
Thermal	77.70%	22.30%	14.95
Disparity	31.98%	68.02%	19.89
Visual + Thermal	83.10%	16.90%	3.37
Visual + Disparity	65.68%	34.32%	7.00
Thermal + Disparity	80.55%	19.45%	5.63
Visual + Thermal + Disparity	84.62%	15.38%	1.74

FPPI was utilized in our finding instead of false positives per window (FPPW) as a result of findings by Dollar *et al.*. It was shown that the evaluation metrics of per-window (FPPW) was flawed[24]. While the per-window methodology is useful for isolating evaluation of binary classifiers, the classification task, the ultimately the goal of pedestrian detection is to output the location of all pedestrians in an image, a detection task. Therefore, full image metrics are more appropriate for this task as it provides a natural measure of error of an overall detection system.

The results from Table 6.3 was then plotted, Figure 6.5, to visually show our improvements by comparing the miss rate and FPPI for each feature used. It is important to note that a lower miss rate and/or FPPI indicates a better performance. Furthermore, we also plotted, Figure 6.6, the performance of each feature detector against the entire test dataset. This plot shows the miss rate versus false positives per image and uses log-average miss rate as a common reference value for summarizing performance. A lower curve indicates a better performance.

Using this detection system, we created a script in Matlab to draw green boxes

around areas where the system had identified the image as having a pedestrian. Below in Figures 6.7 and 6.8 are sample result images of our pedestrian detection system shown in both spectra.

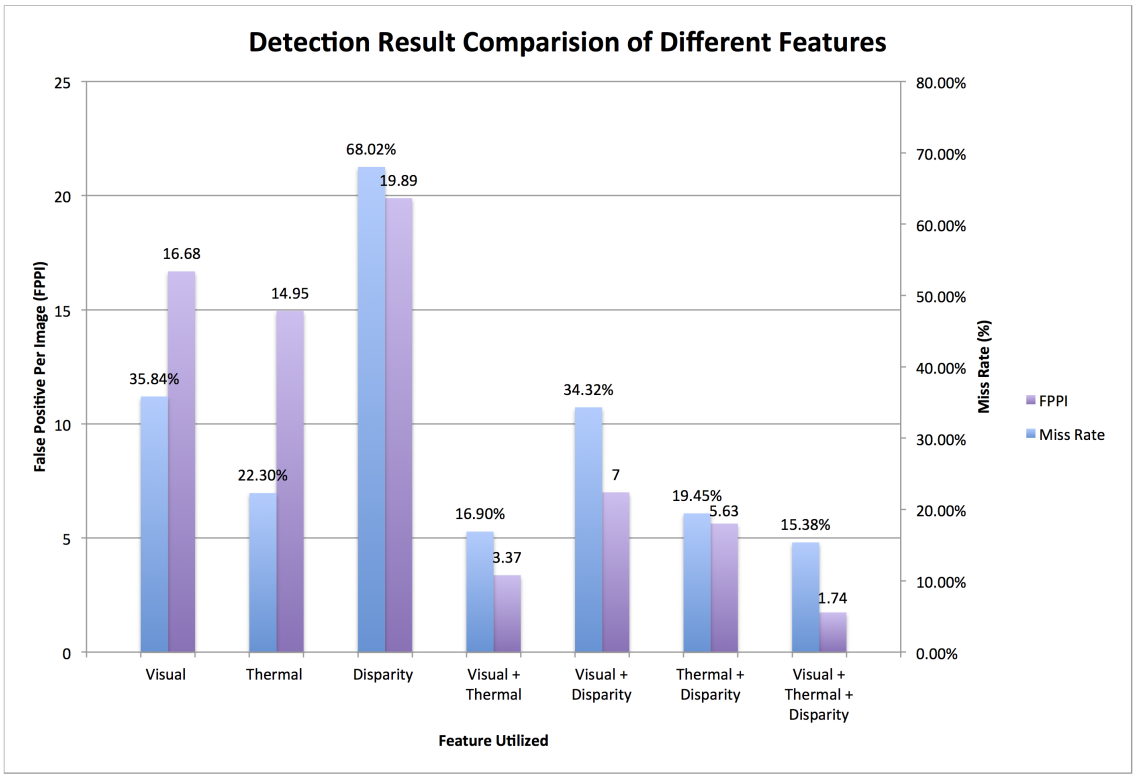

Figure 6.5: Detection result comparison of different features

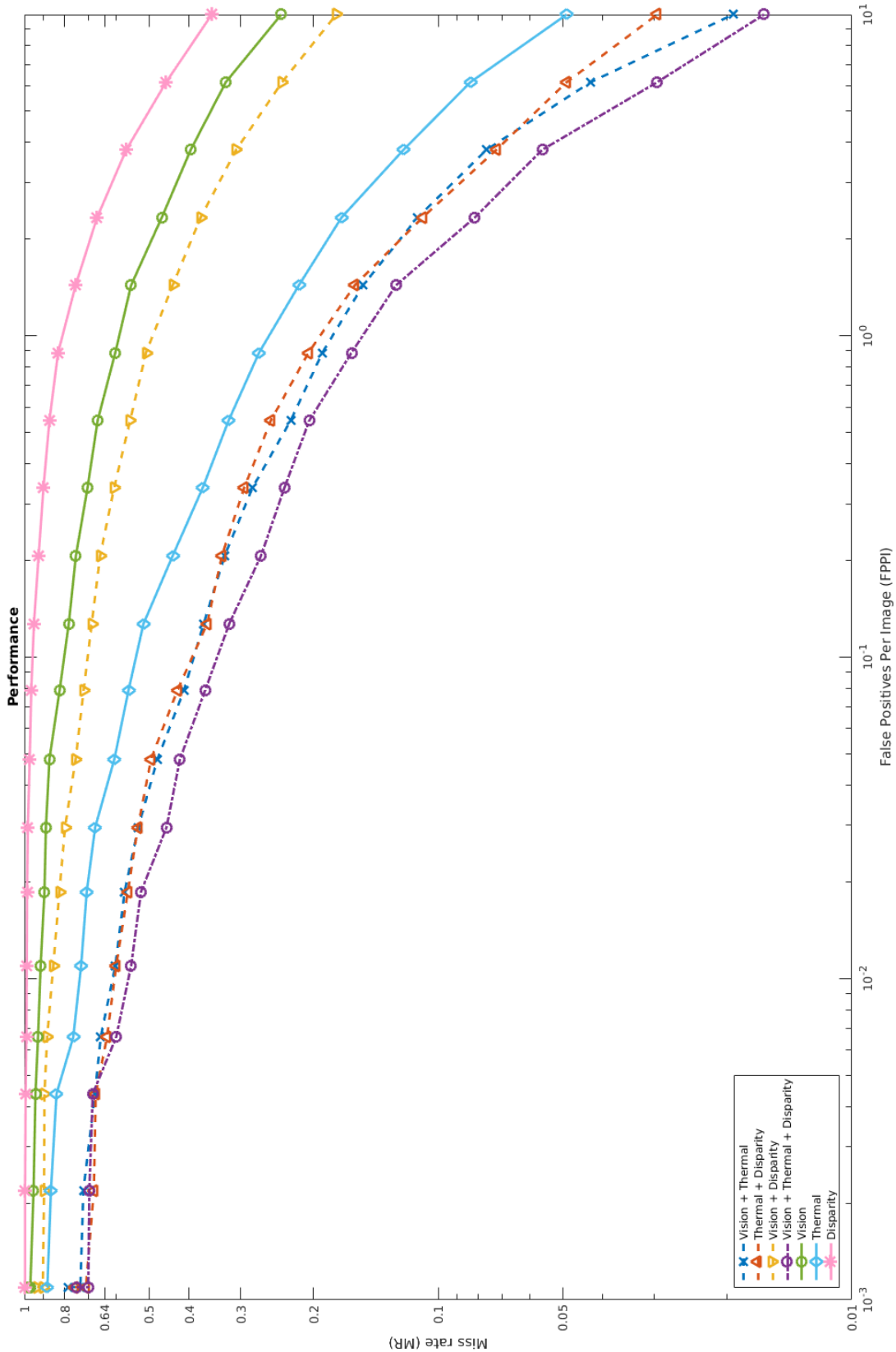

Figure 6.6: Performance plot of different features

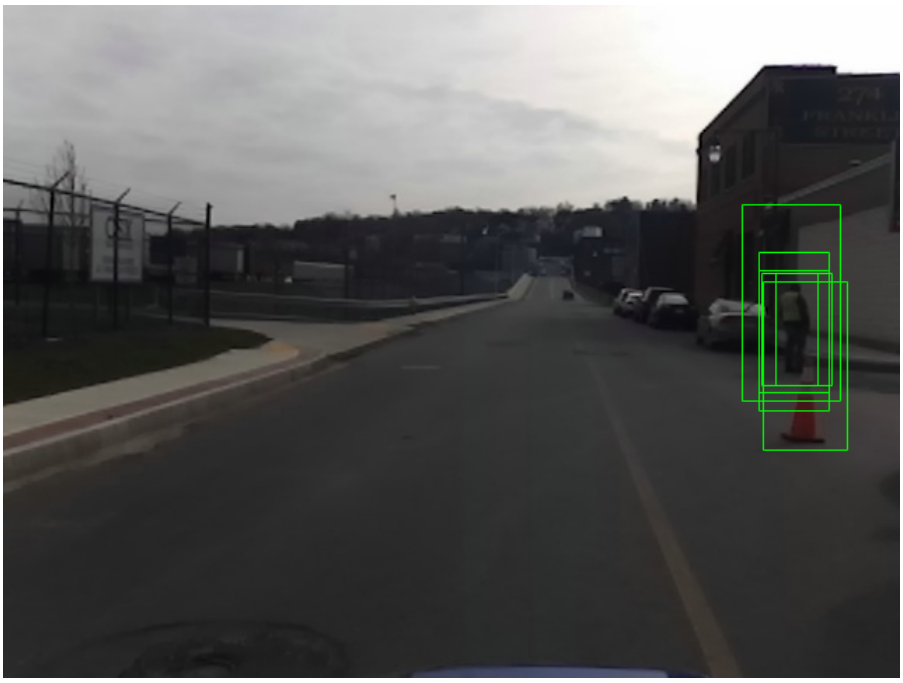

Figure 6.7: Detection result on visual image

Figure 6.8: Detection result on thermal image

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This thesis work has presented a novel approach to data and decision fusion of a multispectral imaging system. We have designed and built a testing rig which is not only portable but reusable on any vehicle with a ski rack. This thesis work has also made our data set publicly available to enable future research on computer vision. Furthermore, we have provided a permanent platform for which more data can be easily collected. This data will maintain the same perspective views and will allow for quick and effective post processing with very little overhead. With even more data, the accuracy of our SVM will undoubtedly increase.

An real-world application of the mathematical theory of trifocal tensor was presented and validated. This allowed for pixel level data fusion across a multispectral set of data. Different architecture of SVMs were presented as well as showing the flaws of decision based cascaded SVMs. We have shown that a confidence score based cascaded SVM was more appropriate in fusion of multiple features. Lastly, this was validated through testing of pedestrian detection in our data set. We have

shown a significant improvement through the use of feature fusion. It should also be noted that we utilized only linear SVMs for computational efficiency, we believe RBF SVMs could provide for even better results.

7.2 Future Work

Further improvements can be made to our current work. There are three major categories in which improvements can be made.

The first category is in the quality of our equipment. As the cost of LWIR imaging continues to drop, more and more higher resolution thermal cameras will be come readily available and affordable. Stereo vision cameras are also improving consistently, the current problem does not lie with the resolution of the imaging sensor, but with the data transfer protocol. Unfortunately, even with USB 3.0, we are unable to achieve enough bandwidth to stream simultaneous high definition videos without frame loss. Furthermore, data storage and data write speed becomes a significant problem as we are unable to record the data fast enough or have enough space to store this vast amount of data. At 1080p resolution, approximately four seconds of footage requires 1GB of storage. With improvements to technology we hope to be able to work with higher resolution images in the future as this will fully utilizes the power of HOG features.

The second category of improvements lies with discovery of more meaningful features. Features in which we can fuse to our current set to further improve SVM classification and detection accuracy. One of these features we can look at is the correlation between disparity map and pedestrian image size. These two values are proportional to each other since a person's height should increase as they get closer to our cameras which leads to a higher disparity value. This could then form a

constraint, with a certain threshold to account for people of different height.

The last category of improvement to our current work is to able to enable this system to perform in real-time. A pedestrian detection system is not useful in a real-world situation if it cannot operate in real-time at a minimum of four frames per second. Four FPS corresponds to 250 ms which is the average human reaction time. This acceleration can be achieved on graphics processing units (GPUs) or field programmable gate arrays (FGPAs) as these equipment are perfect for vast parallel processing.

Bibliography

[1] A. David and P. Jean, "Computer vision: a modern approach," 2002.

[2] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?," in *Computer Vision-ECCV 2014 Workshops*, pp. 613–627, Springer, 2014.

[3] ASIRT, "Road crash statistics," 2015.

[4] C. Ibarra-Castanedo, S. Sfarra, M. Genest, and X. Maldague, "Infrared vision: Visual inspection beyond the visible spectrum," in *Integrated Imaging and Vision Techniques for Industrial Inspection*, pp. 41–58, Springer, 2015.

[5] R. Gade and T. B. Moeslund, "Thermal cameras and applications: a survey," *Machine vision and applications*, vol. 25, no. 1, pp. 245–262, 2014.

[6] R. H. A. Zisserman, *Multiple view geometry in computer vision.* Cambridge, 2004.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.

[8] H. Pant, "Support vector machines (svm) fundamentals part-i," 2013.

[9] H. P. Graf, E. Cosatto, L. Bottou, I. Dourdanovic, and V. Vapnik, "Parallel support vector machines: The cascade svm," in *Advances in neural information processing systems*, pp. 521–528, 2004.

[10] Y. Xu, L. Xu, D. Li, and Y. Wu, "Pedestrian detection using background subtraction assisted support vector machine," in *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*, pp. 837–842, IEEE, 2011.

[11] K. E. Van De Sande, T. Gevers, and C. G. Snoek, "Evaluating color descriptors for object and scene recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1582–1596, 2010.

[12] S. J. Krotosky and M. M. Trivedi, "On color-, infrared-, and multimodal-stereo approaches to pedestrian detection," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 8, no. 4, pp. 619–629, 2007.

[13] U. G. Mangai, S. Samanta, S. Das, and P. R. Chowdhury, "A survey of decision fusion and feature fusion strategies for pattern classification," *IETE Technical review*, vol. 27, no. 4, pp. 293–307, 2010.

[14] "Zed - 3d camera for ar/vr and autonomous navigation," 2016.

[15] "Introducing flir vue pro," 2015.

[16] J. Axelson, *USB complete: the developer's guide*. Lakeview research LLC, 2015.

[17] R. I. Hartley, "Lines and points in three views and the trifocal tensor," *International Journal of Computer Vision*, vol. 22, no. 2, pp. 125–140, 1997.

[18] R. I. Hartley, "In defense of the eight-point algorithm," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 6, pp. 580–593, 1997.

[19] "Guide to authoring media ground truth with viper-gt," 2007.

[20] H. Madsen and P. Thyregod, *Introduction to general and generalized linear models*. CRC Press, 2010.

[21] "Classification score documentation."

[22] Q. Wang, J. Pang, L. Qin, S. Jiang, and Q. Huang, "Justifying the importance of color cues in object detection: a case study on pedestrian," in *The Era of Interactive Media*, pp. 387–397, Springer, 2013.

[23] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pp. 1030–1037, IEEE, 2010.

[24] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 4, pp. 743–761, 2012.