AKITA PREFECTURAL UNIVERSITY

STUDY ON HIGH ACCURACY METHOD OF PEDESTRIAN DETECTION FOR VEHICLE CAMERA IMAGES

（車載カメラ画像に対する歩行者検知手法の高精度化に関する研究）

Xingguo Zhang

（張　興国）

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Akita Prefectural University, Japan

Date　3 / 2015

ABSTRACT

Over the past twenty years, research has moved toward intelligent systems that predict dangerous situations and anticipate accidents. An effective and efficient visual word selection method based on Bag-of-Features (BoF), which can be applied to the pedestrian detection problem, is proposed in this thesis. We first calculate the difference in the total appearance frequency of each visual word in pedestrian and non-pedestrian images. Visual words that exhibit greater absolute values are more efficient for pedestrian detection, and are thus selected. The effectiveness of the proposed method is validated by analyzing the distribution of selected feature points. Through this analysis, we find that discriminative feature points for pedestrian images are mainly located about the lower body, whereas those for non-pedestrian images are mainly located in background areas. Experimental results show that, using the proposed method, the detection rate for the Daimler-DB datasets exceeds 92.5%, whereas the miss rate is less than 6.8%. More-over, the time required for learning and detection can be reduced by approximately 50%, with no significant degradation in precision, using the proposed method, even if only 40% of the visual words are selected. Overall, our experiments offer insights into what makes current systems work well, and state-of-the-art results on several image recognition benchmarks.

ACKNOWLEDGEMENTS

I would like to thank the following people who have helped so much with my research. Firstly, my supervisors, Prof. Guoyue Chen and A.P. Kazuki Saruta for all the invaluable advice they have provided over the last three years, and for all their help in proof-reading this thesis. I’m especially grateful to Mr. Yuki Terata for his continued involvement in my work. Secondly, I’d like to thank all the people I’ve worked with at the Akita Prefectural University. Thirdly, I’d like to thank everyone who has provided feedback and suggestions on my work and on this thesis. Finally, I thank my family and friends for encouraging me and bearing with me during all these years.

Xingguo Zhang

TABLE OF CONTENTS

Introduction

Over the past twenty years, research has moved toward intelligent systems that predict dangerous situations and anticipate accidents [1][2][3]. Intelligent machines have been engineered by humans since the appearance of early civilizations. The formalization of Artificial Intelligence around the 1950s brought intelligent machines to a new dimension, in which their role in human lives has progressively gained importance. At this moment, humans are assisted everywhere: from hazard alarms, medical technology, communications, transportation, etc. Pedestrian detection is one of the most challenging tasks in computer vision, and has received a lot of attention in the last years. However, pedestrian detection also is an extremely challenging task due to the large intra-class variability caused by different articulated poses and clothing, cluttered backgrounds, abundant partial occlusions and frequent changes in illumination. The research in this thesis is focused on the role of Computer Vision for driver assistance, which not only represents a hot research topic nowadays but also is of crucial importance for human societies, as it is argued along this chapter.

1.1 Advanced driver assistance systems

Automobile is one of the most important vehicles in the world, and its invention has greatly affected people's lives. Since their popularization during the 20th century, automobiles have changed societies in many aspects: demographic distribution, urbanism, social interactions, industry growth, environmental alterations, economy development, etc. Moreover, their potential to provide independent, flexible and fast movement to people has lead to new trends in city planning, traveling and employment. According to [4], around 50 million passenger cars and 20 million commercial vehicles are being produced worldwide every year. If car numbers keep increasing at the present rate, there will be more than a billion on the road by 2025, specially due to emerging economies like India and China. Unfortunately, together with the many benefits, such a technology has also carried a dark side since the very beginning: traffic accidents. According to a recent report by the W.H.O., road accidents represent the 6th cause of death in high-income countries and the 11th worldwide [5]. Every year almost 1.2 million people are killed in traffic crashes while the number of injured rises to 50 million. Unfortunately, this number is still growing. And behind every accident is a family tragedy.

In order to improve safety, in the last twenty years, research has moved toward intelligent systems able to predict dangerous situations and anticipate the accidents. They are referred as advanced driver assistance systems (ADAS), in the sense that they help the driver by providing warnings, assisting to take decisions and even taking automatic evasive actions in extreme cases. They differ from the previous safety technologies in the sense that they do not only can rely on physical/mechanical cues from the host vehicle but in addition they understand the exterior world up to some extent. As will be devised during this thesis, Artificial Intelligence plays a key role when pursuing this understanding of the vehicle surroundings.

The first research in the area of ADAS was put by E. Dickmanns group in 1986 with an autonomous highway driving system [6]. They presented a system able to drive through closed highways at speeds of up to 96 km/h by vehicle cameras, simple image processors and Kalman filtering. Nowadays many ADAS have already been commercialized and can be found from some premium vehicles (e.g. Lexus, Mercedes, Volvo).

1.2 Pedestrian protection systems

In view of above mentioned terrible statistics, during the last twenty years companies have progressively turned their safety efforts also to pedestrian protection. In these early stages, research was focused on optimizing the physical parts of the vehicle in order to minimize the risk and degree of injury. Some examples of this research direction, often referred to as improving safety through design, are collapsing fenders, hood and windshield, or increasing the space between hood and the engine to accommodate the pedestrian's head in the case of a crash. The first pedestrian protection system which using machine vision was conducted in the 1990s by Papageorgiou (MIT), Gavrila (University of Amsterdam and Daimler Chrysler), Broggi and Bertozzi (University of Parma). Nowadays, pedestrian safety has become an interesting research and development topic for companies, governments and research centers.

Pedestrian Protection Systems (PPSs) are a particular type of ADAS devoted to pedestrian safety. A PPS is formally defined as a system that detects both static and moving people in the surroundings of the vehicle (typically in the front area) in order to provide information to the driver and perform evasive or braking actions on the host vehicle if needed. Pedestrian detection before the impact (either long or short term) is crucial given that the severity of injuries for the pedestrian decreases with speed of the crashing vehicle. Thus, any reduction in the speed can drastically reduce the severity of the crash. As mentioned at [7], pedestrians have a 90% chance of surviving to car crashes at 30km/h or below, but less than 50% chance of surviving to impacts at 45 km/h or above.

Figure 1.1 illustrates the potential of PPSs. They can anticipate the potential accident they can not only provide warnings to the driver in a reduced time but also control the different active measures like airbags or brakes. Hence, the distance where pedestrians can be severely damaged is significantly reduced.

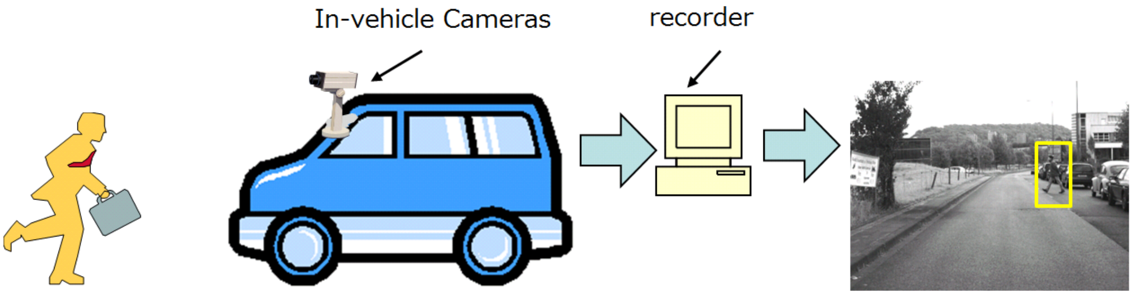


Figure 1.1. The outline of the pedestrian protection Systems

1.3 Generic framework

The simplest technique for determining the initial location of an object is the sliding window method, whereby detector windows at various scales and locations are shifted over the image. However, the computational cost of high-precision detection in every sliding window is often too high to allow for real-time processing[8][9]. After that, they soon started to include other stages aimed at both reducing the number of false positives and to accelerate the processing. For example, tracking techniques are also being included to the systems recently.

Some researcher proposed a generic architecture to be used as a framework for pedestrian detection from a vehicle camera [4]. The architecture consists of six conceptual modules each one with its own responsibilities:

- **Preprocessing**, which takes the input data from the camera and prepares it to the further processing, such as exposure time, gain adjustments and calibration, to mention a few.

- **Foreground segmentation** extracts regions of interest or candidates from the image to be sent to the classification module, avoiding as many background regions as possible.

- **Object classification**, which receives a list of candidates likely to contain a pedestrian. In this stage, they are classified as pedestrian or non-pedestrian with the aim of minimizing the number of false positives as well as the false negatives.

- **Verification and refinement**. Many systems contain one step that verifies and refines the ROIs classified as pedestrians, referred to as detections. The verification filters false positives using criteria not overlapped with the classifier while the refinement performs a fine segmentation of the pedestrian (not necessarily silhouette-oriented) so to provide an accurate distance estimation or to support the following module, tracking.

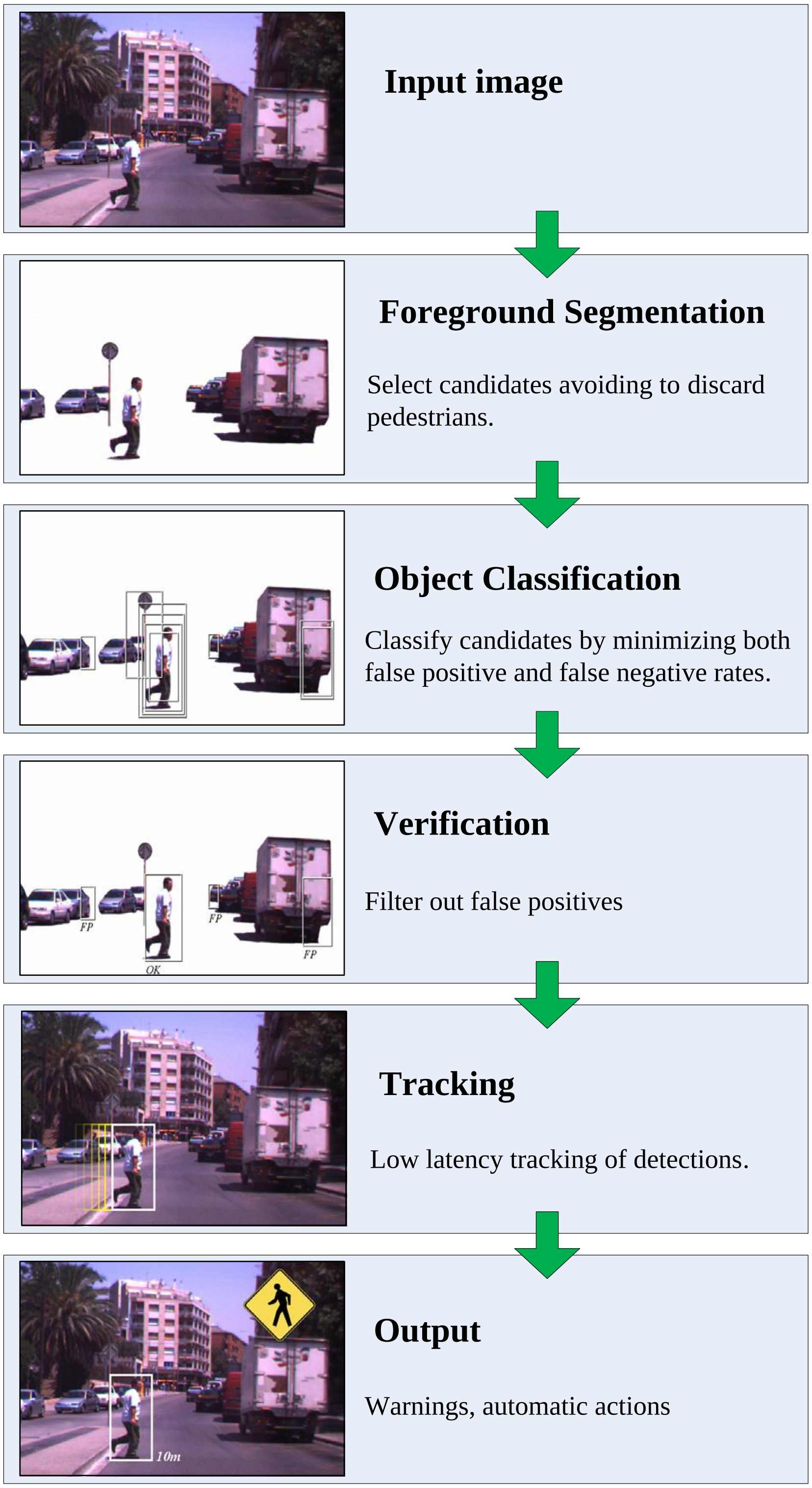

Input image
Foreground Segmentation
Select candidates avoiding to discard pedestrians.
Object Classification
Classify candidates by minimizing both false positive and false negative rates.
Verification
FP
OK
FP
FP
Filter out false positives
Tracking
Low latency tracking of detections.
Output
10m
Warnings, automatic actions

Figure 1.2. The architecture of the pedestrian detection system (Figure form[4]**)**

- **Tracking**, which follows the detected pedestrians along time with several purposes such as avoiding spurious false detections, predict the next pedestrian position and direction and even other high-level tasks, like inferring pedestrian behavior.

- **Application**, which takes high level decisions by making use of the information provided by the previous modules. This module represents a complete area of research, which includes not only driver monitoring or vehicle speed but also psychological issues, human-machine-interaction, etc.

Figure 1.2 shows a schematic overview.

1.4 State of the Art

1.4.1 Preprocessing

The preprocessing module includes tasks such as exposure time, gain adjustments, histogram equalization, spatial alternation and camera calibration, etc.

Although low-level adjustments, such as exposure or dynamic range are normally not described in ADAS literature, some recently published papers have focus on image enhancements for these systems. Real-time adjustments are a recurring difficulty, specially in urban scenarios. For example, short tunnels, narrow streets and the fast motion of the scene (common conditions in PPSs) can result in images with over/under saturated areas or poorly adjusted dynamic range, which creates additional difficulties for the latter algorithms of the system. Although not specifically devoted to ADAS, Nayar et al. [10] present some approaches for performing a locally adaptive dynamic range: fusion of different exposures, spatial filter mosaicing and pixel exposures, multiple image/pixel sensors, etc. Besides, during last years solutions exploiting high dynamic range images [11] are gaining interest in driver assistance due to their potential to provide high contrast in the aforementioned scenarios. In fact, these cameras cover both VS and NIR spectra so they are also useful for night time vision.

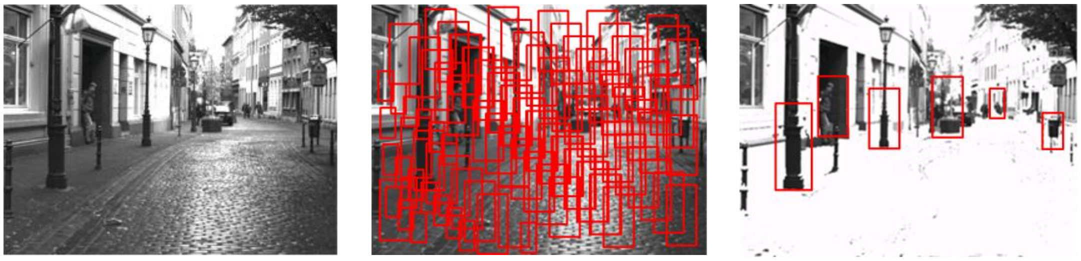
The existing approaches can be divided into two categories: monocular-based and stereo-based. In the former case, the algorithms are mainly based on the study of visual features. In [12], Broggi et al. correct the vertical image position by relying on the detection of horizontal edges oscillations: the horizon line is computed according to the previous frames. A comparative study of different monocular camera pose estimation approaches is presented in [13]. It includes horizontal edges, features-based and frame difference algorithms. Recently, [14] presents a probabilistic framework for 3D geometry estimation based on a monocular system. A training process, based on a set of 60 manually labeled images, is applied to form a prior estimation of the horizon position and camera height (i.e., camera pose values).

High dynamic range sensors provide the possibility of obtaining highly contrasted images in outdoor scenarios. In the next years, this technology will be of crucial importance in PPSs in order to avoid the over/under-saturated regions that are typically seen in ADAS imagery. In fact, many of the failures of the current detection algorithms is related to poorly contrasted images so this technology will undoubtedly benefit the system performance.

The more recent of the reviewed works show a clear trend towards using stereo-based approaches to obtain accurate camera pose estimates in spite of the additional CPU time required for disparity/depth estimation.

1.4.2 Foreground segmentation

Foreground segmentation, sometimes referred to as candidate generation, extracts regions of interest (ROIs) from the image to be sent to the classification module, avoiding as many background regions as possible. The simplest technique to obtain the initial object location hypotheses is the sliding window technique, where detector windows at various scales and locations are shifted over the image. The computational costs of the high-precision detection approach for every sliding window are often too high to allow for real-time processing[4]. In this thesis we added a so-called Candidate Generation Pruning (CGP) step to our system, that extract specific windows in the image. These techniques are of remarkable importance not only to reduce the number of candidates but also to avoid scanning regions like the sky. The key to this stage is to avoid missing pedestrians; otherwise the subsequent modules will not be able to correct the error. While describing this module we will often use the term pedestrian size constraints (PSC), which refers to the aspect ratio, size and position that candidates shall fulfill to be considered to contain a pedestrian.



(a) (b) (c)

Figure 1.3. Foreground Segmentation Schemes

An exhaustive scanning approach [15] that selects all of the possible candidates in an image according to PSC, without explicit segmentation. This method is known as sliding window. For instance, in [15], the authors start by scanning the image with candidate windows of 64 × 128 pixels, placing these windows every 8 pixels. Then they reduce the image size by a factor of 1.2, and perform the same scan again. This procedure has two main drawbacks: 1) the number of candidates is large (see Fig 1.3(b)), which makes it difficult to fulfill real-time requirements, although some proposals have recently studied this problem; and 2) many irrelevant regions are passed to the next module (e.g., sky regions or ROIs inconsistent with perspective), which increases the potential number of false positives. As a result, other approaches are used to perform explicit segmentation.

The exhaustive scan is typically used in general human detection systems, for example, image retrieval, whereas PPSs tend to use some kind of segmentation. In fact, the latter can take advantage of some application prior knowledge (e.g., it is not necessary to search the top area of the image), so that the number of ROIs to process can be greatly reduced. For example, a typical exhaustive scan on a 640 × 480 image can provide from 400,000 to 3,000,000 ROIs, depending on the sampling step and the minimum candidate size.

According to the literature, stereo is the most successful option. 2D-based analysis does not provide convincing results at this stage. For instance, symmetry is not very reliable so extra-cues such as depth are necessary, hotspot analysis seems to be ruled by heuristics and attentional bottom-up pixel-based algorithms do not provide accurate ROI positions, so the reduction of the number of candidates is not as large as expected. More sophisticated appearance based techniques are likely to be used during classification, not during candidates generation. In addition, the accuracy of motion-based approaches depends on driving speeds, and the reliability of those approaches has not been demonstrated under the wide range of ADAS conditions.

1.4.3 Object classification

The object classification module receives a list of ROIs that are likely to contain a pedestrian. In this stage, they are classified as pedestrian or non-pedestrian with the goal of minimizing the number of false positives and false negatives.

Silhouette matching: The simplest approach is the binary shape model. [16], in which an upper body shape is matched to an edge modulus image by simple correlation after symmetry-based segmentation. A more sophisticated approach is the Chamfer System, a silhouette-matching algorithm proposed by Gavrila et al. in [17]. This system consists of a hierarchical template-based classifier (Fig 1.4) that matches distance-transformed ROIs with template shapes in a coarse-to-fine manner. The shape hierarchy is generated offline by a clustering algorithm. This technique has also been exploited for

TIR images in [18]. Also in the TIR spectrum, Nanda et al. [19] perform probabilistic template matching on a multiscale basis, by using just three templates (each for a defined scale).

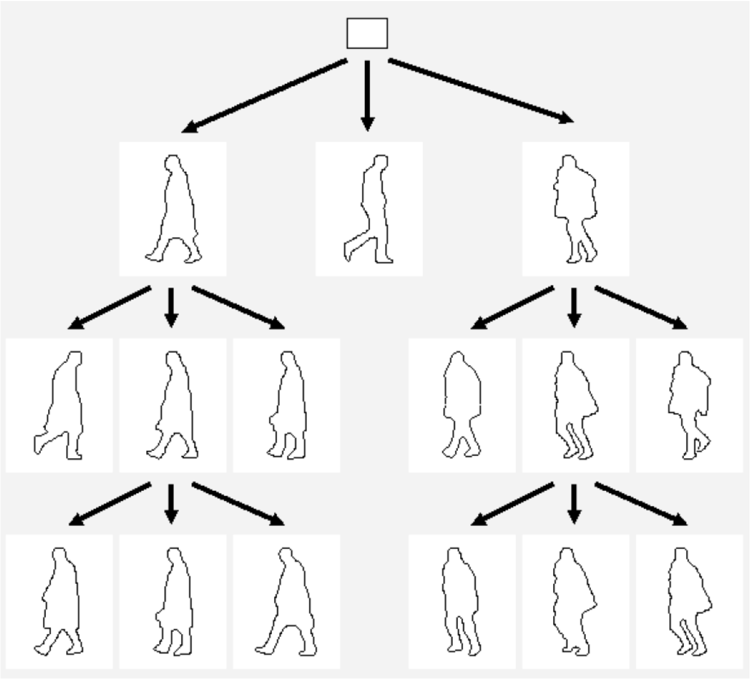


Figure 1.4. Hierarchy of templates used in the Chamfer System (figure from [20]**).**

Appearance: The methods included in this group define a space of image features (also known as descriptors), and a classifier is trained by using images known to contain examples (pedestrians) and counter-examples (non-pedestrians). The seminal work of Dalal and Triggs [15] showed the importance of using rich block-based descriptors such as the Histograms of Oriented Gradients (HOG) representation, which provides both robustness and distinctiveness. Based on this work, other authors have proposed additional features that enrich the visual representation, including the use of color through self-similarity features (CSS) [21], texture through block-based Local Binary Patterns (LBP) [22], and the design of efficient gradient-based features via integral channels [23].

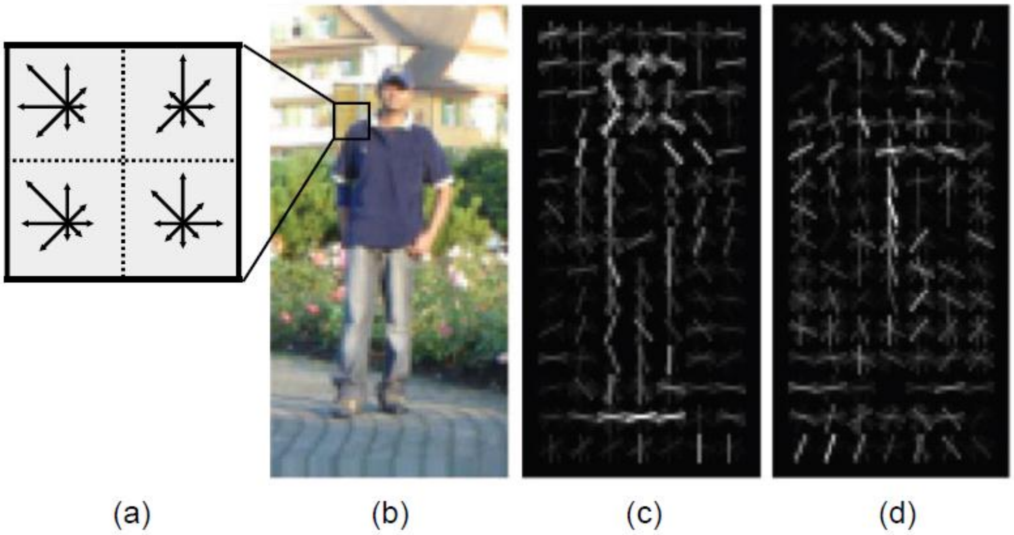


Figure 1.5. Histograms of Oriented Gradients by Dalal and Triggs (figure from [8]**).**

Following a holistic approach (i.e., target is detected as a whole), in [24][25], Gavrila et al. propose a classifier that uses image grayscale pixels as features and a neural network with local receptive fields (NN-LRF [26]) as the learning machine that classifies the ROIs generated by the Chamfer System.

Dalal and Triggs [8] present a human classification scheme that uses SIFT inspired [27] features, called histograms of oriented gradients (HOG), and a linear SVM as a learning method. A HOG feature also divides the region into k orientation bins (in this case, k = 9), but instead of computing the ratio between two bins, they define 4 different cells that divide the rectangular feature, as illustrated in Fig. 1.6. In addition, a Gaussian mask is applied to the magnitude values in order to give more weight the center pixels, and the pixels are interpolated with respect to pixel location within a block (both factors disallow the use of the integral image). The resulting feature is a 36-dimensional vector containing the summed magnitude of each pixel cells, divided into 9 bins. These features have been extensively exploited in the literature.

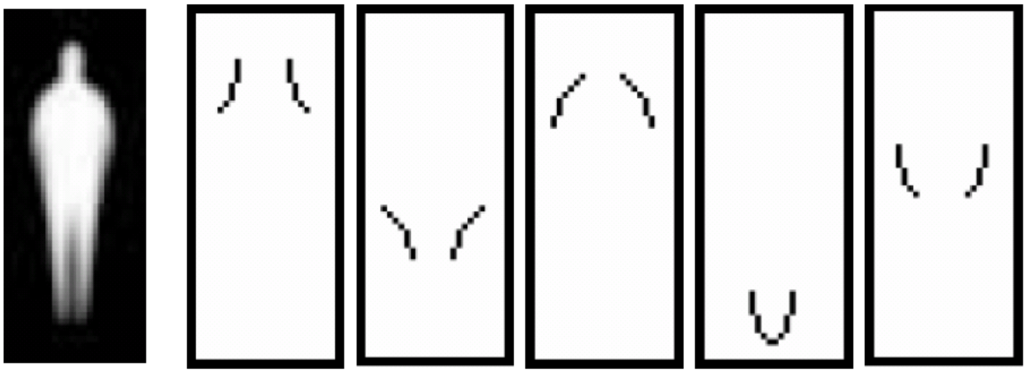


Figure 1.6. First five edgelet features selected by AdaBoost in the approach by Wu and Nevatia (figure from [28]**).**

Wu et al. [28] study the performance of short segments (up to 12 pixels long) of lines or curves, referred to as edgelets, as features for AdaBoost for VS images. In this case, a mask is attached to each feature in order to provide pixel-wise segmentation (Fig 1.7). The same authors study edgelets and HOG together with AdaBoost and SVM learning algorithms in both the VS and TIR [29]. Each of the selected cells is referred to as a shapelet feature.

Other features and learning algorithms used in the literature include the gradient magnitude and quadratic SVM, Four Directional Features and Gaussian kernel SVM, and intensity image with Convolutional Neural Networks or with an SVM.

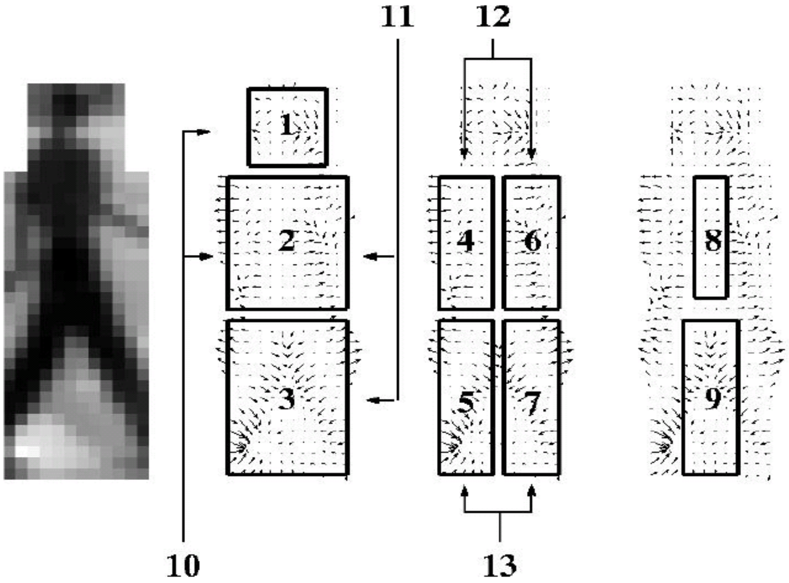


Figure 1.7. Part-based classification using gradient-based features (figure from[30]**)**

Part-based approaches [30], contrary to the previous techniques, combine the classification of different parts of the pedestrian body (e.g., head and legs), instead of classifying the entire candidate as a single entity.

Lin and Davis [31] have recently proposed a technique that combines some of the aforementioned paradigms to a greater or lesser extent, i.e., silhouette, appearance, holistic and parts-based. First, HOG descriptors are computed for the whole image following [8]. Then, the descriptors are used to extract a silhouette, which is fed to a probabilistic hierarchical part-matching algorithm. Finally, HOGs are again computed for the closest regions of the matched silhouette, serving as features for a radial basis function (RBF) kernel SVM.

Other approaches: Following recent research in object detection, Leibe et al. [32] present a technique termed the implicit shape model, which avoids the candidate generation step. During recognition, each detected keypoint is matched to a cluster, which then votes for an object hypothesis using Hough voting, thus avoiding a candidate generation step. The Chamfer distance is used to provide a fine silhouette segmentation of the pedestrian. In [33], Seeman et al. improve this technique with multi-aspect (viewpoint and articulation) detection capabilities, extending the hypothesis voting to object shapes, rather than just objects.

Silhouette matching methods are not applicable as stand-alone techniques. Even the elaborate Chamfer System needs an extra appearance-based step. In contrast, methods that exploit appearance seem to indicate the current direction of research, specifically revolving around the continuous development of new learning algorithms and features for use in these algorithms, not only in pedestrian detection but also in general object classification.

Despite the large number of papers, approaches tend to be poorly compared to one another in PPSs research. Wojek et al. [34] try shed light on the comparison of classifiers with a study on some popular features and learning methods. Two conclusions are highlighted: HOGs and shape context features are the best option, independent of the learning algorithm, and feature combination significantly improves detector performance. In recent years, however, the lack of comparisons has been amended thanks to Dalal's proposal, which has been established as a defacto baseline.

1.5 Objectives

In this thesis, in order to reduce the number of traffic accidents, injuries, and deaths, we focusing on verification task which is foremost stage in aforementioned PPS systems.

A comprehensive review of recent work and existing techniques for pedestrian detection is carried out. The survey represents a crucial part of the research in the sense that it helps to visualize what has and has not been made and the current needs in this area.

We also describe the process of pedestrian detection using a BoF approach, and examine some existing problems. In order to address these issues, we propose a very simple method to reduce the dimension of the classifier by setting a limit value to remove irrelevant and redundant visual words.

According to the reports almost half of pedestrian who die in road traffic crashes are occurred during the night. Thus, a study of pedestrian detection via our visual words select method from near-infrared images is made. In addition, we investigate the distribution of discriminative feature points which belong to the selected visual words from NIR images and VS images.

Of course, there are many interesting aspects that are left unexplored since they are out of the thesis scope. We think that enumerating some of these aspects can help to provide a better focus on the aim of the thesis.

At first, although a very strong emphasis in time consumption and realistic computational requirements is made along the thesis, and in fact is a key piece of it, we do not spend efforts on real-time optimizations.

And then, the algorithms are not specifically trained with children examples. As will be seen, the foreground segmentation is thought to work on adults, although the system will be able to detect pedestrians of very different sizes and proportions. Hence, children younger than 10 years old are not taken into account in the statistics, not for good nor for bad. Young children are expected to go with an adult, as in fact it is seen in the presented sequences, and according to NHTSA [35], children younger than 10 represent just the 4% of the killed and the 10% of the injured, so they are also left for future investigation.

1.6 Thesis outline

The thesis is organized in the following chapters. Chapter 2 focuses on the study of pedestrian detection based Bag-of-Features. Chapter 3 introduce our proposed visual words selection strategy for pedestrian detection. Chapter 4 presents the results of distribution of the selected feature points. Chapter 5 presents an improved visual words selection method via set two thresholds. Chapter 7 provide formal conclusions, formal discussion and perspectives on the research area.

CHAPTER 2

Pedestrian Detection Based on Bag-of-Features

In this chapter, we will introduce the flow of Bag-of-Features to detect the pedestrian.

2.1 Why using the BoF to detect the pedestrian?

The simplest technique to obtain the initial object location hypotheses is the sliding window technique, where detector windows at various scales and locations are shifted over the image. The computational costs of the high-precision detection approach for every sliding window are often too high to allow for real-time processing[4]. To speed up the detection process, the pedestrian detection process is often decomposed into the potential target's detection and classification to reduce the search area[5]. First, the system defines a region of interest (ROI), which is possibly associated with a potential pedestrian. Second, detection is validated by a high-precision identifying method for the ROI.

The seminal work of Dalal and Triggs [6] showed the importance of using rich block-based descriptors such as the Histograms of Oriented Gradients (HOG) representation, which provides both robustness and distinctiveness. Based on this work, other authors have proposed additional features that enrich the visual representation, including the use of color through self-similarity features (CSS) [7], texture through block-based Local Binary Patterns (LBP) [8], and the design of efficient gradient-based features via integral channels [9].

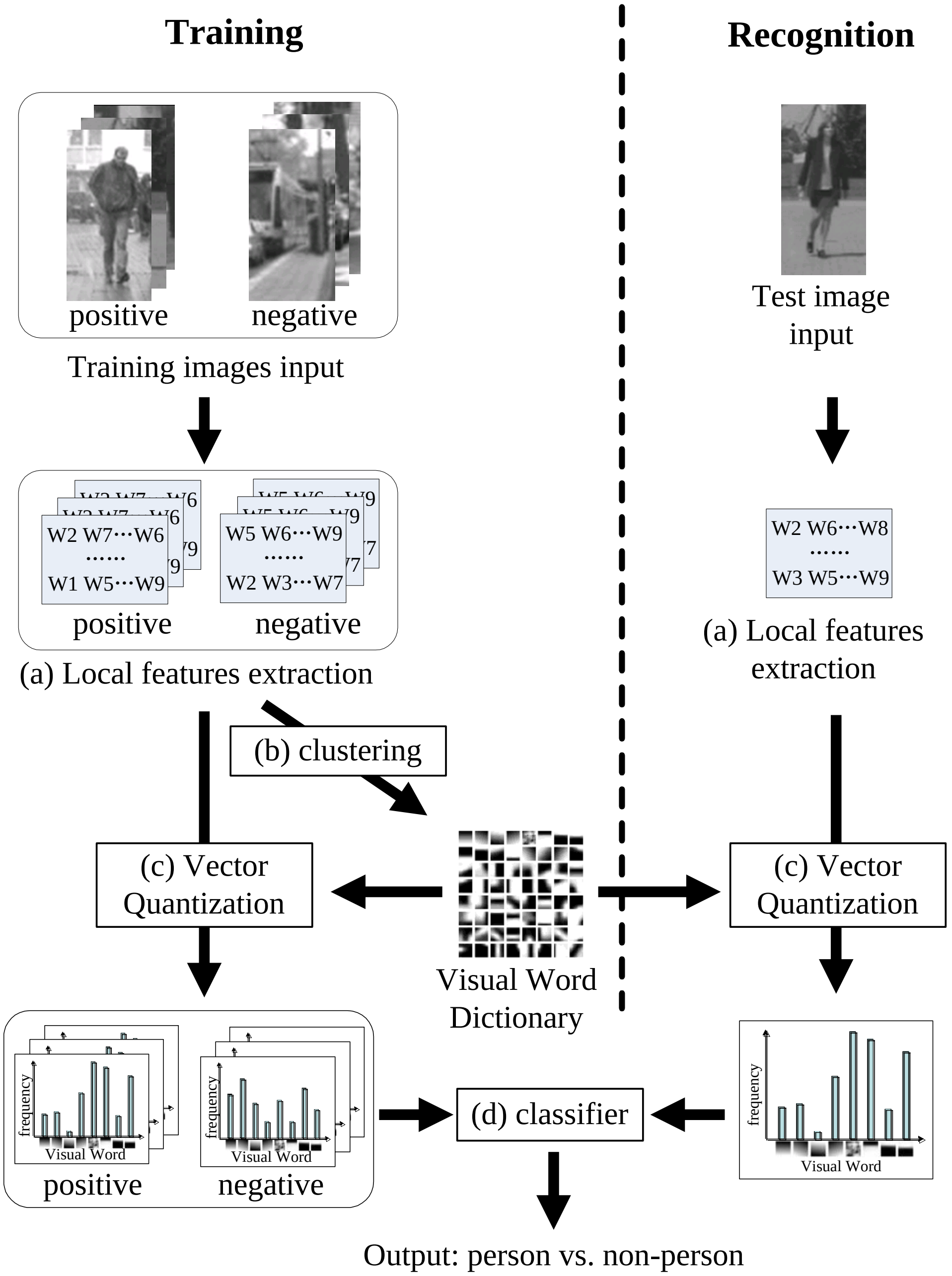


Figure 2.1. Pedestrian detection progress of BoF

All of these approaches are holistic, in the sense that the whole pedestrian is described by a single feature vector and is classified at once. Because of this, those approaches also brings some problems, especially when the pedestrian's position cannot be in the middle of the ROI search window.

Recently, some authors have proposed successful methods for combining local detectors [36][28] and integrating the evidence from multiple local patches [37][16]. This type of approaches provides more flexibility in the spatial configuration of the different parts of the object, which leads to higher adaptability to the different poses of the pedestrian. However, those methods are usually restricted when pedestrian at far scales.

Many of the current methods for image classification represent images as collections of independent patches characterized by local visual descriptors, and vectors quantize them by the K-means method to produce so-called visual words [38]. The introduction of such visual codebooks has allowed significant advances in image classification, especially when combined with bag-of-features (BoF) models[39]. One advantage of this method is that the frequency histogram is irrelevant to the location of the local feature and is very useful in detecting the image of the pedestrian with position shift.

However, not every visual word created by K-means is efficient for classification. A compact visual codebook has advantages in both computational efficiency and memory usage.

A compact and discriminative visual vocabulary has been proposed by pioneering[16]. The work hierarchically merges the visual words in a large-sized initial vocabulary, and requires the new histograms to maximize the conditional probability of the true labels of the training images. This is a rigorous but complicated criterion that involves nontrivial computation after each merging operation. These lead to a heavy computational load when dealing with large-sized initial visual words.

Other authors have proposed a visual words select method via Principle Component Analysis (PCA) algorithm [40]. However, it seems to be difficult to achieve both speed and good discrimination.

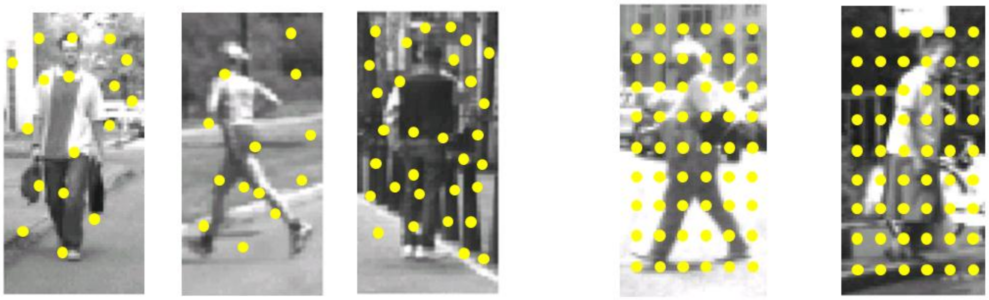
In this section, we describe the pedestrian detection approach by BoF. BoF is an object classification method which ignores the positional information of the local features extracted from the images and uses the occurrence frequency of visual words.

Figure 2.1 shows the basic flowchart of pedestrian detection by BoF, which consists of feature extraction (a), building of visual vocabulary (b), building a frequency histogram (c), and training the classifier (d).

In the training stage, local features are extracted from the training samples, and are clustered into X groups with the K-means algorithm. After clustering, the visual vocabulary is built, and the frequency histogram of each visual word, which records the num-ber of its occurrences, is calculated. The frequency histogram based on the visual vocabulary is considered as the input classifier, which is trained with the support vector machine (SVM) algorithm. In the recognition stage, the frequency histogram of local features extracted from the test samples is calculated in the same manner, and the constructed classifier will make the decision based on this frequency histogram. The detail of each process is described below.

2.2 Local feature extraction

To describe the image, we must extract features from the input images. Typically, three local feature extraction methods are used: interest point sampling, regular dense sampling, and random sampling. Interest point detection is often used because of its good performance in some fields. The scale invariant feature transform (SIFT) [14] and speeded up robust features (SURF) [15] are two typical methods based on interest point feature extraction. SIFT can robustly identify objects even among clutter and under partial occlusion, because the SIFT feature descriptor is invariant to uniform scaling, orientation, and partially invariant to affine distortion and illumination changes. However, SIFT needs to structure the Gaussian scale space to find interest points, and therefore cannot always extract enough features for low-resolution pedestrian images.



(a) SIFT feature points

(b) Dense-SIFT

Figure 2.2. Example of feature points

To obtain better discriminative power, we utilize a regular dense sampling method, known as dense-SIFT descriptors. This is roughly equivalent to running SIFT on a dense grid of locations at a fixed scale and orientation, but without the need to structure the Gaussian scale space to find interest points.

2.3 Forming the visual vocabulary and frequency histograms

Various clustering methods can be used to form the visual vocabulary, such as k-means [41], affinity propagation [42], self-organizing maps [43], fuzzy c-means [44], etc. Each method has its strengths, but inevitably has some weaknesses. In terms of a combination of efficiency and accuracy, k-means is a satisfactory clustering method.

Visual vocabularies are created as follows. After extracting a large number of local patch descriptors (here, dense-SIFT descriptors) from a set of training images, k-means clustering is used to group these descriptors into k clusters, where k is predefined. The center of each cluster is called the “visual word,” and a set of visual words forms a “visual vocabulary.” Each image descriptor is then labeled with the most similar visual word, according to the Euclidean distance between the two, and the image is characterized by a k-dimensional histogram of the number of occurrences of each visual word. The frequency histogram of each visual word forms the training data that is input to the SVM.

2.4 SVM classifier

SVM is a well-known statistical learning method [20]. The objective of SVM learning is to find a hyperplane that maximizes the inter-class margin of the training samples. Feature vectors are projected into a high-dimensional space by a kernel function. The final SVM classifier is given by the following expression

$$f(x) = \sum_i \omega_i K(x, x_i) \qquad (1)$$

where ω_i are support vectors and $K(x,y)$ is the kernel function. There are several common kernel functions, such as a linear kernel, polynomial kernel, radial basis function (RBF) kernel [21], etc. The choice of kernel function is dependent on the data and application. We tested each kernel function, and found that an RBF kernel gives the best performance without any obvious deterioration in efficiency. Thus, in our experiments, we use an RBF kernel.

CHAPTER 3

A Visual Words Selection Strategy for Pedestrian Detection

A compact visual codebook has advantages in both computational efficiency and memory usage. For example, when linear or nonlinear SVMs are used, the complexity of computing the kernel matrix, testing a new image, or storing the support vectors is all proportional to the codebook size, n. Also, many algorithms working well in a low dimensional space will encounter difficulties such as singularity or unreliable parameter estimate when the dimensions increase. This is often called the “curse of dimensionality”.

A compact visual codebook provides a lower-dimensional representation and can effectively avoid these difficulties. Moreover, in patch-based object recognition, the histogram used to represent an image is essentially a discrete approximation of the distribution of visual words in that image. A large-sized visual codebook may over-fit this distribution, as pointed out in [45].

One disadvantage of the dense regular grid is that a large number of redundant features are included in the visual vocabulary, meaning more time will be spent on feature extraction and classification during the training and recognition stage. A simple and efficient visual vocabulary is expected to speed up learning and classification.

Recently work of creating a compact and discriminative visual codebook has been seen in [46], which hierarchically merges the visual words in a large-sized initial codebook.

To minimize the loss of discriminative ability, the work in [46] requires the new histograms to maximize the conditional probability of the true labels of training images (or image regions in their work). This is a rigorous but complicated criterion that involves nontrivial computation after each merging operation. Moreover, at each level of the hierarchy, the optimal pair of words to be merged are sought by an exhaustive search. These lead to a heavy computational load when dealing with large-sized initial codebooks. Creating a compact codebook is essentially a dimensionality reduction problem.

To preserve the discriminative power, any classification performance related criterion may be adopted, for example, the rigorous Bayes error rate, error bounds or distances, class separability measure, or that used in [46].

In this chapter, we consider two-class classification problems, and propose a very simple method to reduce the dimension of the classifier by setting a limit value to remove irrelevant and redundant visual words. Our method calculates the difference in the total appearance frequency for each visual word of the pedestrian and non-pedestrian images. The visual words that exhibit greater absolute values are considered to be more efficient for pedestrian detection, and are selected. Experimental results show that the proposed method retains almost the same detection accuracy when only 40% of the visual words are selected.

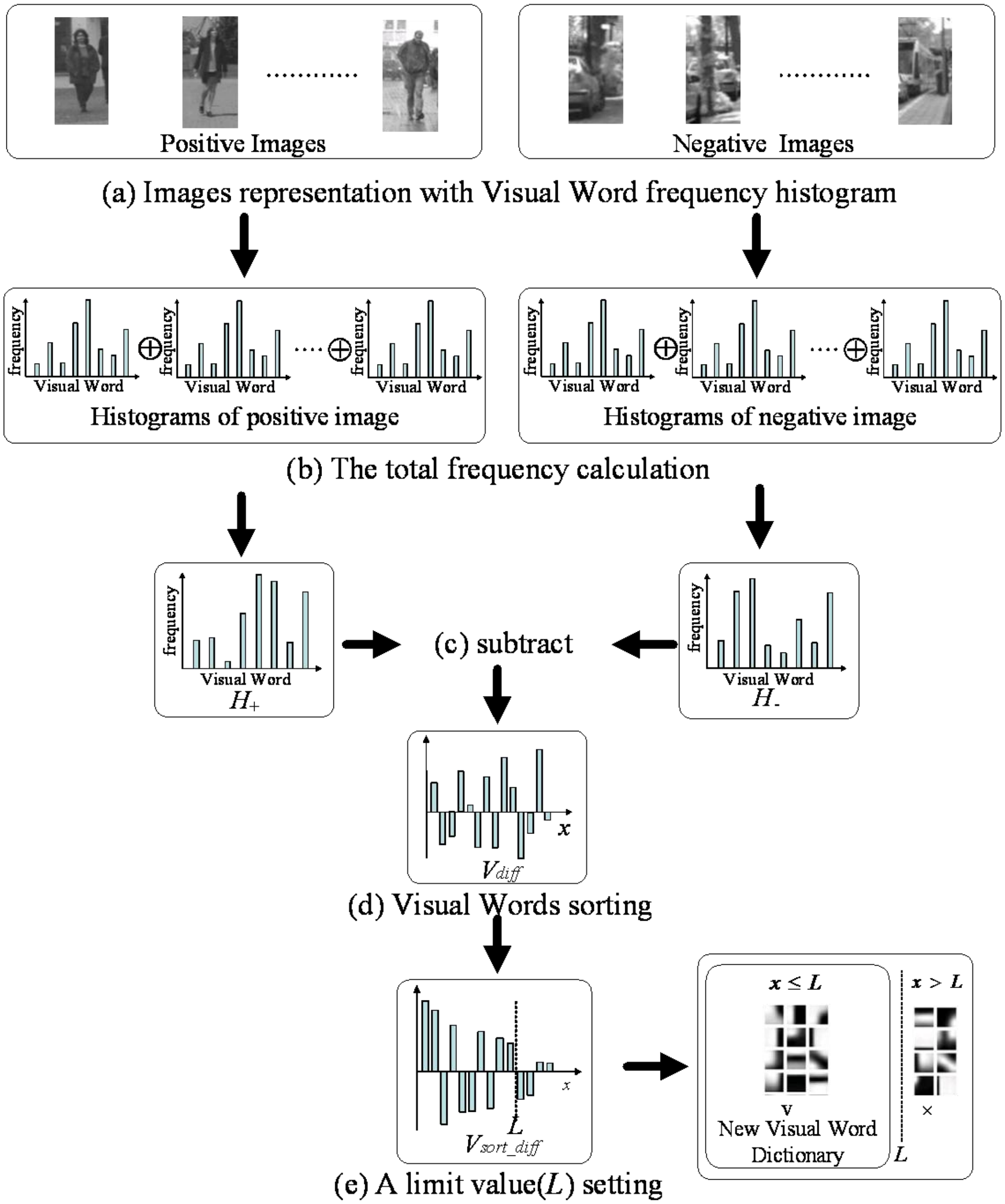

Positive Images
....
Negative Images
....
(a) Images representation with Visual Word frequency histogram
frequency
Visual Word
....
Histograms of positive image
Histograms of negative image
(b) The total frequency calculation
H_+
H_-
(c) subtract
x
V_{diff}
(d) Visual Words sorting
L
V_{sort_diff}
$x \leq L$
$x > L$
v
×
New Visual Word Dictionary
(e) A limit value(L) setting

Figure 3.1. Flowchart of proposed method

3.1 Flowchart of visual words select method

A brief overview of this approach is given in Figure 3.1. First, the quantization histograms obtained from each training image are divided into positive images and negative images (Figure 2.1(a)). The total frequency histograms for positive sample $\mathbf{H_1}$ and negative sample $\mathbf{H_0}$are then computed by the following equation, as shown in Figure 2.1(b).

$$\begin{cases} H_1(x) = \sum_{i \in N,\ x \in X} h_{1,i}(x) \\ H_0(x) = \sum_{j \in M,\, x \in X} h_{1,i}(x) \end{cases} \qquad (2)$$

$$\mathbf{H_1} := \left[H_1(0), H_1(1), \ldots, H_1(X)\right]^{\mathrm{T}}$$
$$\mathbf{H}_0 := \left[H_1(0), H_1(1), \ldots, H_1(X)\right]^{\mathrm{T}}$$

where X is the number of visual words in a dictionary. N and M represents the number of pedestrian and non-pedestrian training samples, respectively. $h_{1,i}(x)$ represents the frequency of x'th visual word on i'th pedestrian image. $h_{0,j}(x)$ represents the frequency of x'th visual word on j'th non-pedestrian image.

Next, we normalize the two total frequency histograms as

$$\begin{cases} H_1(x) = \dfrac{H_1(x)}{\sum_{x=1}^{X}\left(H_1(x)\right)} \\ H_0(x) = \dfrac{H_0(x)}{\sum_{x=1}^{X}\left(H_0(x)\right)} \end{cases} \qquad (3)$$

The difference between $\mathbf{H}_1$ and $\mathbf{H}_0$ is calculated to obtain the difference vector (Figure 2.1(c))

$$V_{diff}\left(x\right)=H_1\left(x\right)-H_0\left(x\right) \qquad (4)$$

If $V_{diff}(x)$ is positive, this visual word is effectively classified as a positive sample, and vice versa. The larger the absolute value of $V_{diff}(x)$, the more beneficial the *x*-th feature to the classification.

The visual words are sorted in descending order of absolute value, and a limit value L is set to determine the expected size of the new visual vocabulary to be preserved (Figure 2.1(d)). Visual words for which $V_{diff}(x)$ is below the limit value L are considered to be redundant, and are screened out of the original dictionary. The remaining L visual words comprise a new visual vocabulary (shown in Figure 2.1(e)). Next, the corresponding dimensions of the original histograms h are removed according to the new visual vocabulary. The new frequency histogram of the visual vocabulary forms the input to the classifier, which is trained by the SVM.

Based on our experimental results, we found that L could be set as 0.4X with little change in detection accuracy (e.g., if X = 500, then L = 200).

3.2 Benchmarking

In order to comparison to other state-of-the-art proposals, multiple public pedestrian datasets have been collected. Public datasets are necessary for two reasons: 1) to evaluate algorithms with different example sets, taken at different places under different conditions, but specifically from different research groups (which adds extra variability); and 2) to compare new algorithms with existing ones, that is, given that it is hard to reproduce algorithms, the easiest way of establishing comparisons is to compare results from the same datasets following the same criteria.

There are some specific requirements that a pedestrian dataset shall fulfill to be specifically used in PPSs. Some of them are a must in any set while others make easier the task of evaluating different aspects of classifiers. They can be summarized in the following points: **Topic significance**. This first one may seem obvious, but it is important that the test data is the most similar to the final application as possible in this case ADAS environments. This means that pedestrians must be standing approximately in the same plane as the on-board camera placed at a realistic height from the ground. **Quantity**. Given the variability of the target, the number of examples shall be high, for example, at least 1, 000 positive samples for training. **Resolution**. As has been seen, the range of pedestrians sizes in the image is large due to perspective and distance. Given that algorithms can either make use of the resized or the original size (depending on the classifier) it is desirable to make both approaches available. By providing this data researchers get a well defined set in both cases avoiding to have to reconstruct these cases.

Sequences. Cropped samples are useful for the object classification module, but in order to benchmark the whole system full annotated video sequences are required.

Multiple public pedestrian datasets have been collected over the years; INRIA [1], ETH [2], TUD-Brussels [3], Daimler [4] (Daimler stereo [5]), Caltech-USA [6], and KITTI [7] are the most commonly used ones. They all have different characteristics, weaknesses, and strengths.

INRIA is amongst the oldest and as such has comparatively few images. It benefits however from high quality annotations of pedestrians in diverse settings (city, beach, mountains, etc.), which is why it is commonly selected for training (see also §4.4). ETH and TUD-Brussels are mid-sized video datasets. Daimler is not considered by all methods because it lacks color channels. Daimler stereo, ETH, and KITTI provide stereo information. All datasets but INRIA are obtained from video, and thus enable the use of optical flow as an additional cue.

Daimler Pedestrian Classification Benchmark (DC-01)[47] and the Computer Vision Center Pedestrian Dataset are the first specifically ADAS oriented pedestrian datasets, containing images taken from cameras mounted on a vehicle. The samples are significantly smaller (36 and 24 pixels high, respectively), all taken from street scenarios, so in contrast with INRIA, so in this case there are not out-of-the-topic images.

Today, Caltech-USA and KITTI are the predominant benchmarks for pedestrian detection. Both are comparatively large and challenging. Caltech-USA stands out for the large number of methods that have been evaluated side-by-side. KITTI stands out because its test set is slightly more diverse, but is not yet used as frequently. For a more detailed discussion of the datasets please consult [8,7]. INRIA, ETH (monocular), TUD-Brussels, Daimler (monocular), and Caltech-USA are available under a unified evaluation toolbox; KITTI uses its own separate one with unpublished test data. Both toolboxes maintain an online ranking where published methods can be compared side by side.

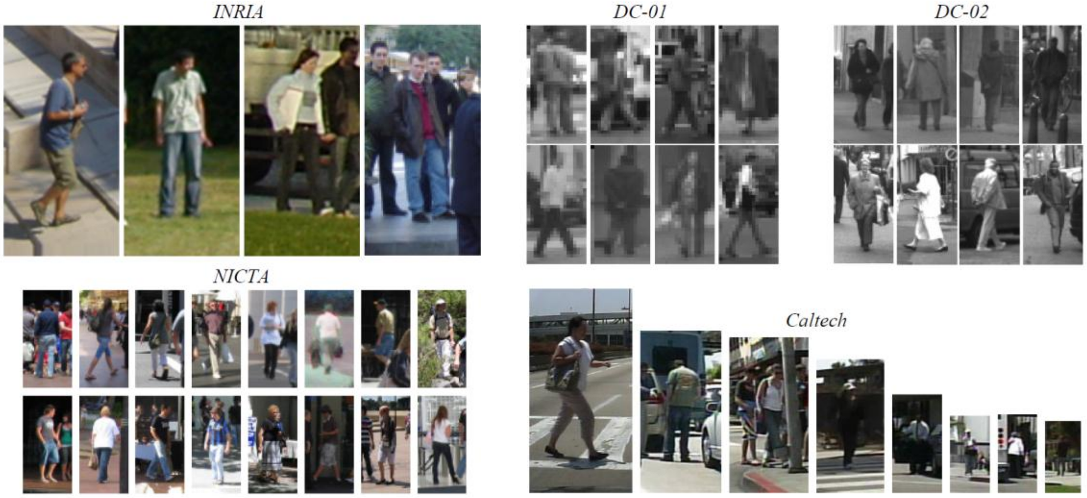


Figure 3.2. Positive examples from different pedestrian datasets

From all these sets, just DC-01, CVC-01 and NICTA are ADAS-specific, so although the others can already provide an intuition on detection algorithms performance, just the former provide relevant statistics for ADAS. Two pedestrian datasets have been presented aimed at resolving this lacks. The first one is the Caltech Pedestrian Dataset [2], which contains on-board video sequences containing instance annotation. Although this dataset seems promising at a first glance given the spectacular number of samples stated, the number of single pedestrians is similar to the previous ones, and are not so precisely annotated as for example NICTA. Moreover, the authors do not provide testing data, which represents a big inconvenience for researchers. The second dataset is Daimler Pedestrian Detection Benchmark (DC-02), which contains grayscale resized training examples and fully annotated video sequences. Fig 3.2 illustrates some examples of these datasets.

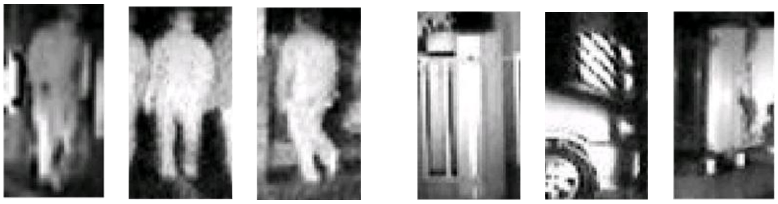


Figure 3.3. Positive examples from different pedestrian datasets

In this thesis we present a near infrared reflection dataset (NIR). For NIR images, we collected a set of video sequences containing pedestrians from multiple view points and of multiple sizes, using a monochrome board camera KPC-EX500BA and a NIR lamp RM-240 (spectral wavelength in 0.7–2.5 microns). Images were captured at night and the height of the persons in the images ranged from 50 to 300 pixels. Some of the training and testing images are shown in Fig. 2.7.

3.3 Experimental Results

In this section, we evaluate the performance of our proposed method in terms of its detection rate and processing time. In addition, we analyze the distribution of discriminative features by visualizing the selected features. We implement our proposal method using Matlab, and use the open-source toolbox VLFeat to extract SIFT features and form the visual vocabulary. We use the LibSVM to train the classifier, which is integrated software for support vector classification.

3.3.1 Experimental setup

We used the Caltech Pedestrians [24], DaimlerChrysler Pedestrian Classification Benchmark (Daimler-CB) [25], and Daimler Pedestrian Detection Benchmark (Daimler-DB) [26] datasets to conduct a series of experiments. These datasets were collected by the on-board camera within a vehicle, and include images of pedestrians from different viewpoints.

In our experiments, we applied training to these three datasets, and created the detectors respectively. The sizes of both the training and test images were uniformly fixed at 48×96 pixels.

We sampled SIFT features densely over 4-pixel intervals, with a block size of 8×8 pixels. This results in 256 SIFT feature points being extracted from each 48×96 pixel sample. Using the dense-SIFT feature descriptors calculated from all training samples, we undertook *k*-means clustering of the features to form a visual vocabulary. The SVM detectors were then trained using RBF kernels. The performance of SVM classifier depends on the choice of the regularization parameter C and the kernel parameters γ. We use the grid search with cross-validation to determine the optimal values of the parameters C and γ. Experimental results show that the classifier attains optimal performance when $C = 16$ and $\gamma = 1$.

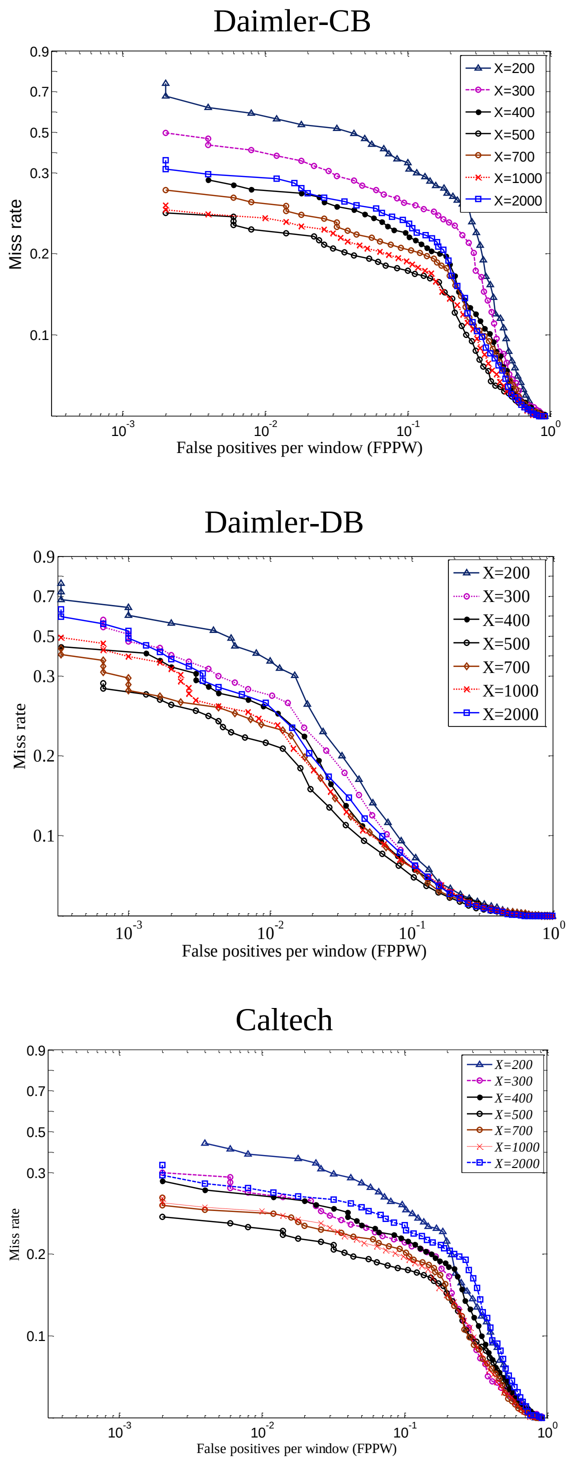


Figure 3.5. Relation between visual word *X* and detection precision.

3.3.2 Detection accuracy with various sizes of visual vocabulary

To clarify the relationship between the detection accuracy and the size of the visual dictionary, and determine the initial number of visual words in the proposed method, we conducted the following experiment.

We randomly selected 3000 pedestrian and 3000 non-pedestrian images from the Caltech, Daimler-CB, and Daimler-DB datasets as the training samples. For the test samples, 3000 pedestrian and 3000 non-pedestrian images were selected from the remainder of each dataset.

We varied the size of the initial visual vocabulary X from 200 to 2000, and tried to ascertain the optimal value for each dataset. The detection accuracies for different X are shown in Figure 3.

The results in Figure 3. show that optimal detection accuracy of each dataset is achieved when X is around 500. Thus, in the following evaluations, the initial number of visual words X is set to 500.

3.3.3 Detection accuracy with various sizes of selected visual word

In this section, we describe the relation between the limit size L of the visual vocabulary and the detection accuracy for the three datasets.

Using an initial visual vocabulary size $X = 500$ for training and testing images, the value of L was varied from 400 to 100. The detection accuracy at different L values is shown in Figure 3.. The results show similar results for the three datasets, with small variations in accuracy when $L \geq 200$.

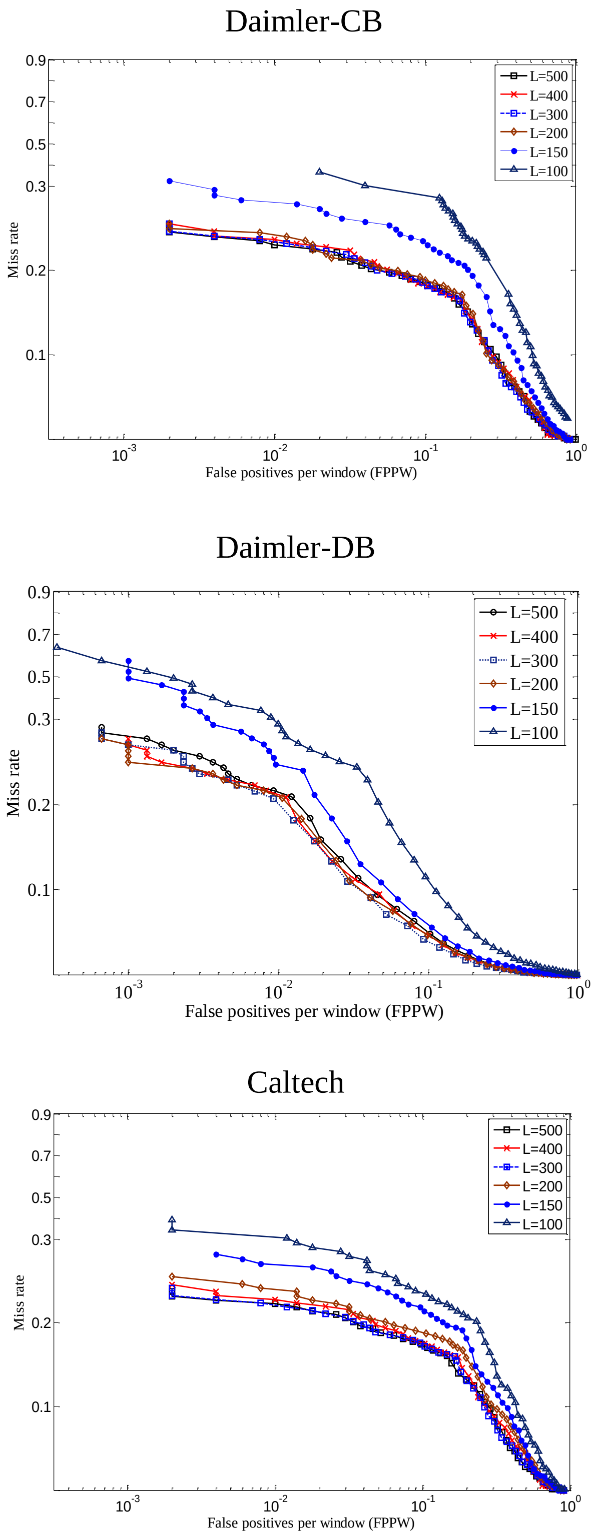


Figure 3.6. Relation between limit value *L* and detection precision.

This implies that 200 efficient visual words in the original visual vocabulary produces almost the same performance as using all 500 visual words. In addition, the detection precision decreases more quickly when $L < 200$.

These results show that the proposed method retains similar detection accuracy when only 40% of the visual words are selected. Thus, we can set $L = 0.4X$ without significantly affecting the accuracy.

3.3.4 Evaluation of pedestrian detection by cross experiments

In this section, we evaluate the proposed method for pedestrian detection using the following cross experiment. First, we set $X = 500$ and $L = 200$. We randomly selected 3000 pedestrian images for Groups A and B, and 3000 non-pedestrian images for Groups C and D from the Daimler-DB dataset. Different combinations of these groups were then used to perform the cross experiment.

TABLE I.
EVALUATION OF THE PROPOSED METHOD

Training Data	**Test Data**	**True Positive**	**False Positive**
A C	B D	92.8%	6.3%
A D	B C	92.5%	6.6%
B C	A D	93.1%	6.0%
B D	A C	92.6%	6.8%

Table 1 shows the experimental results. We found that the detection rate for each group was greater than 92.5%, and the miss rate was less than or equal to 6.8%. This confirms that the proposed method is effective for pedestrian detection applications.

3.3.5 Processing time performance

In this section, we study the relationship between the processing time and the number of selected visual words *L*. We compare the SVM runtime on the same desktop with an Intel i3-540 CPU and 2 GB RAM.

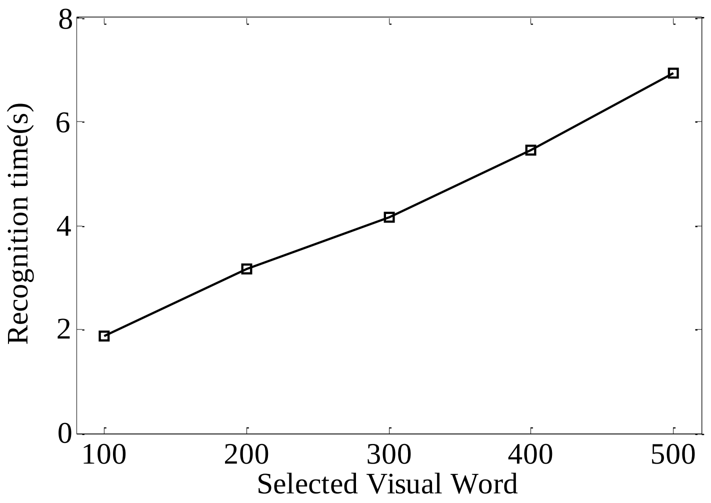


Figure 3.7. Change in recognition time with *L* using 6000 images.

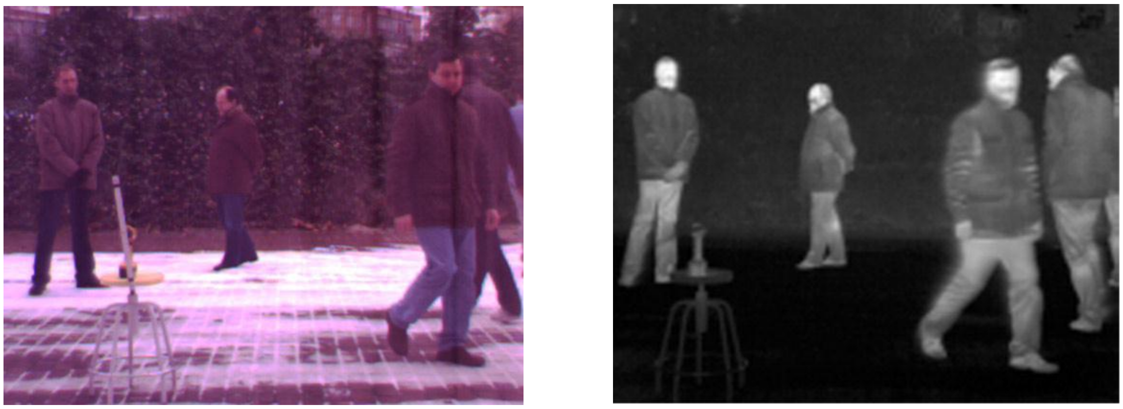
Figure 3. shows that the recognition time increases with the value of *L* for 6000 images. As mentioned above, the detection accuracy is not obviously reduced as *L* is decreased to 200. Thus, the experimental results show that the classification time required by the SVM can be reduced by about 50% using the proposed method, without any significant degradation in accuracy, even if only 40% of the visual words are selected.

3.4 Pedestrian Detection from Near-Infrared

Normal cameras based on visible spectrum images (hereafter called VS images) are not very satisfactory in the absence of plenty of illumination. In the day time, this illumination can come from the sun, but at night, artificial illumination is required. Important areas of interest could be lit with bright lights but undesirable activities are more likely to occur in darker areas. Infrared (IR) cameras are ideally suited to imaging under these conditions, as they sense emitted radiation from the objects of interest, such as pedestrians. However, IR cameras are still expensive to deploy on a large scale. Therefore, in this paper, we attempt to detect pedestrians at night by near-infrared (NIR) cameras which are cheaper.

3.4.1 Sensors

Cameras can be divided according to their working range in the electromagnetic spectrum. Visible spectrum (VS) is in the 0.4-0.74 μm range, near infrared (NIR) covers 0.75-1.4 μm and thermal infrared1 (TIR) captures in 6-15 μm. Fig 3.8 illustrates the same image in VS and TIR for an in-door scenario. However, as will be seen in this chapter, these ideal laboratory conditions differ from the real outdoor ones used in actual PPSs. Pedestrian detection is typically focused on daytime, hence VS cameras are the most extensively used ones. Some papers make also use of NIR, as will be seen later, and in fact they are cheaper than TIR ones. NIR cameras capture relative temperature, which is very convenient for distinguishing hot targets like pedestrians or vehicles from cold ones like asphalt or trees, hence they are used for pedestrian detection at night. Without enough ambient light, VS cameras provide too dark and poorly contrasted scenes, so pedestrian detection is not possible. Along the review we will assume that VS sensors are used if not stated the contrary.



(a) VS

(b) TIR

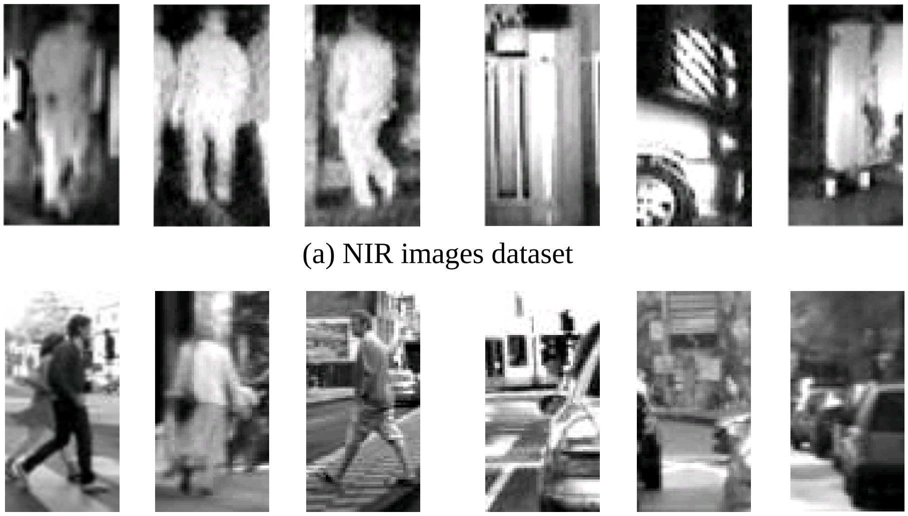
Figure 3.8. Appearance of VS and TIR for the same scene(photo by ADAS group).

3.4.2 Experimental setup

We now describe the datasets used in our experiments.

NIR image data: For NIR images, we collected a set of video sequences containing pedestrians from multiple view points and of multiple sizes, using a monochrome board camera KPC-EX500BA and a NIR lamp RM-240 (spectral wavelength in 0.7–2.5 microns). Images were captured at night and the height of the persons in the images ranged from 50 to 300 pixels. Some of the training and testing images are shown in Fig 3.9(a).

VS image data: For VS images, we tested our experiments on the Daimlar-DB which is a publicly available dataset [22]. The dataset is collected by an on-board camera within a vehicle that includes pedestrians from different viewpoints Fig 3.9 (b)).



(a) NIR images dataset

(b) VS images dataset

Figure 3.9. Example images of NIR images dataset and VS images dataset used in this experiment.

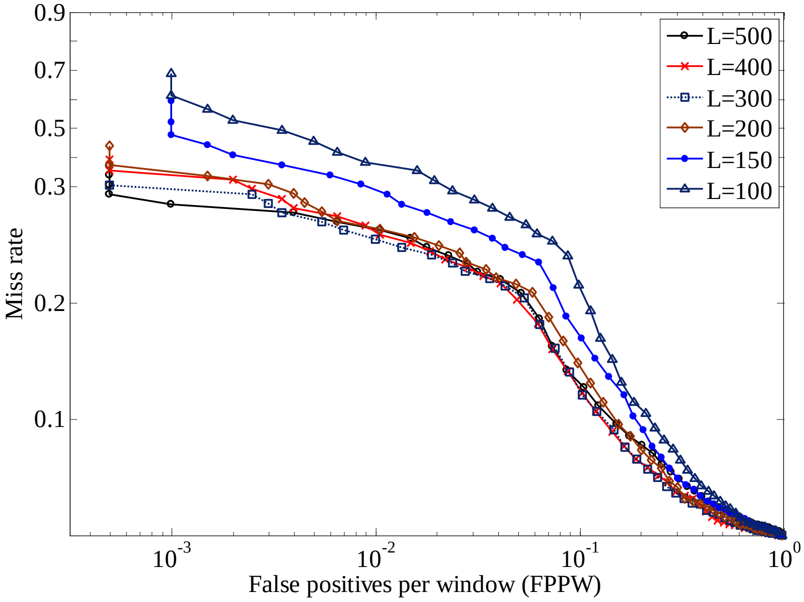
For both the NIR images and the VS images, we selected 3000 pedestrian and 3000 non-pedestrian images randomly as the training samples respectively (we also tried more train data, but the accuracy was no significant increase). And the test samples included 3000 pedestrian and 3000 non-pedestrian images selected from the rest of the dataset.

3.4.3 The number of discriminative visual words in NIR and VS image

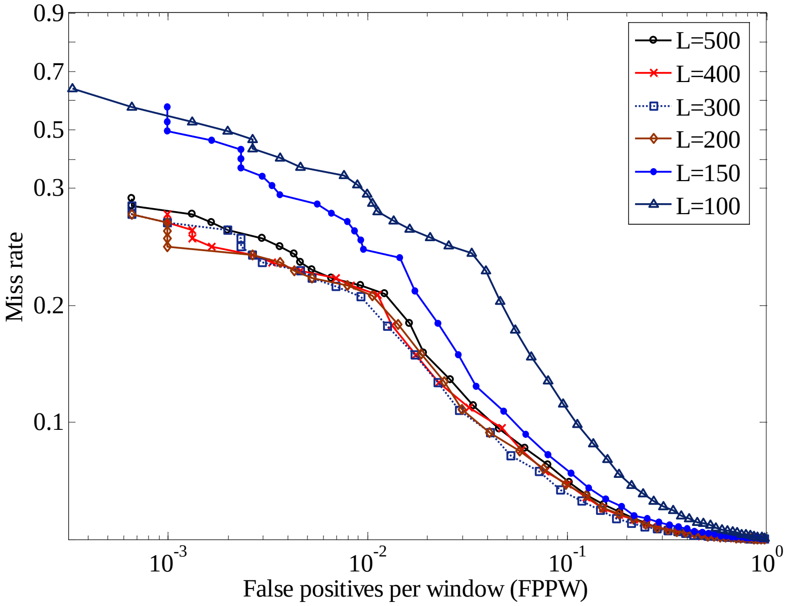
In this section, we describe the relation between the selected visual word L and the detection precision.

We set the initial visual vocabulary size X =500 for NIR and VS images and then set the selected visual word L from 400 to 100. The DET curves of the detection accuracy with difference L are shown in Fig 3.10. We found that, in both of the two model images, the precision variation was very small when $L \geq 200$, which means that 200 efficient visual words in the original visual vocabulary can result in almost the same performance as with all 500 visual words. In addition, the detection precision decreases more quickly as L decreases for $L<200$.

The experiment results show that the proposed method keeps nearly the same detection accuracy even if only 40% of the visual words are selected.



(a) DET curves on NIR images



(b) DET curves on VS images

Figure 3.10. The relation between selected visual word *L* and detection precision.

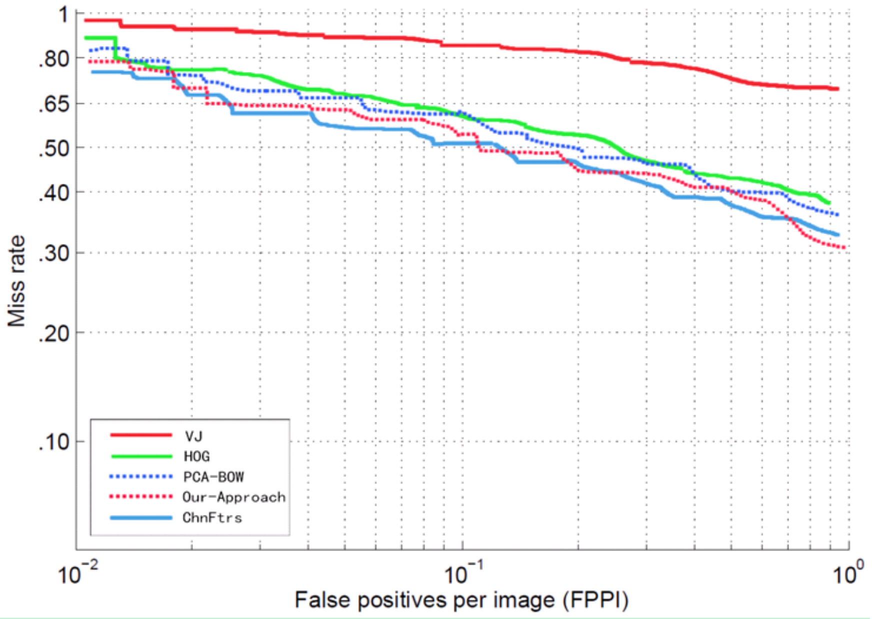
3.4.4 Comparison with the state-of-the-art

In this section, Our final detectors were evaluated with other state-of-the-art methods using our NIR dataset and Daimlar -DB.

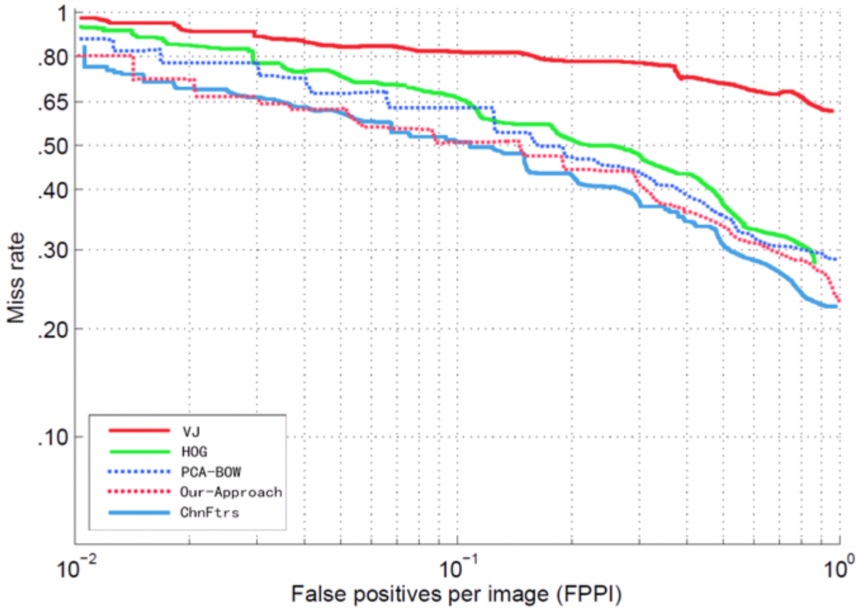
We performed the standard per-image evaluation used in pedestrian detection [2]. We added a so-called Candidate Generation Pruning (CGP) step to our system, in order to obtain a fair comparison with the best performer in this dataset [48]. Making use of projective geometry, the CGP algorithm forecasting the possible arisen location so as to confirm the scope of target searching. This permits us to both accelerate the detection of pedestrians (as fewer windows are evaluated by the classifier), and remove false positives standing on non-plausible locations of the targets, thus improving the resulting accuracy.

Results are shown in **Fig. 3.11**. As a matter of fact, the relative ordering of methods seems roughly preserved across different pedestrian datasets. For the both of two datasets, our detector is competitive in terms of the detection quality with respect to ChnFtrs and provides significant improvement over HOG+SVM.

It is worth mentioning that many of recent works are focus on integrate different features (e.g., gradient magnitude, LUV color channels, motion) in order to feed more relevant information, so that they can be integrated in our framework with moderate changes. This would further increase the accuracy of our method.



(a) Results on the NIR pedestrian dataset



(b) Results on the Daimlar pedestrian dataset

Figure 3.11. Performance of our method.

3.5 Summary

This chapter presented a method of obtaining a compact and discriminative visual vocabulary for pedestrian detection. Our visual word selection method calculates the difference in the total appearance frequency of each visual word in pedestrian and non-pedestrian images. The visual words that exhibit greater absolute values are considered to be more efficient for pedestrian detection, and are thus selected. The experiments also showed that the learning and detection process can achieve similar precision in about 50% of the time using the proposed method, even if only 40% of the visual words are selected. Furthermore, we detect pedestrians at night by near-infrared (NIR) cameras. We found our method is also For the both of two datasets, our detector is competitive in terms of the detection quality with other state-of-the-art.

CHAPTER 4

Distribution of the Selected Feature Points

In this chapter, we analyze the distribution of selected feature points from pedestrian and non-pedestrian images, and present the average number of discriminative features in the pedestrian images. In addition, we discuss the causes of false positive (FP) and false negative (FN) results.

4.1 Define the discriminative visual words

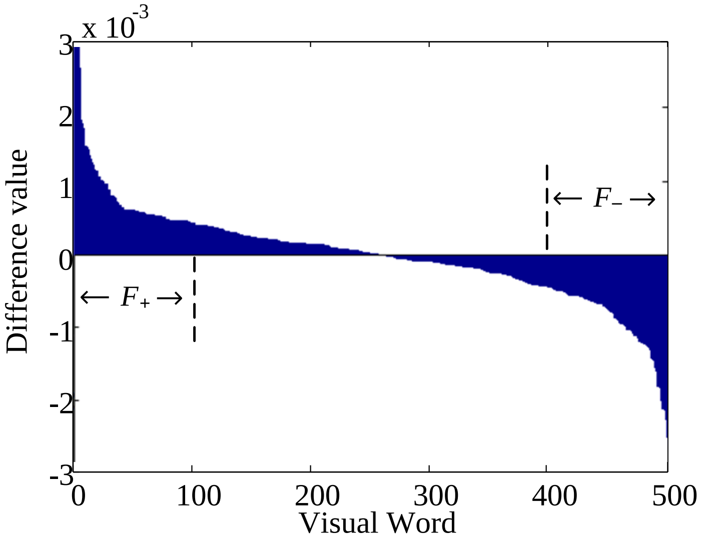


Figure 4.1. Justification of F_+ and F_- with visual words.

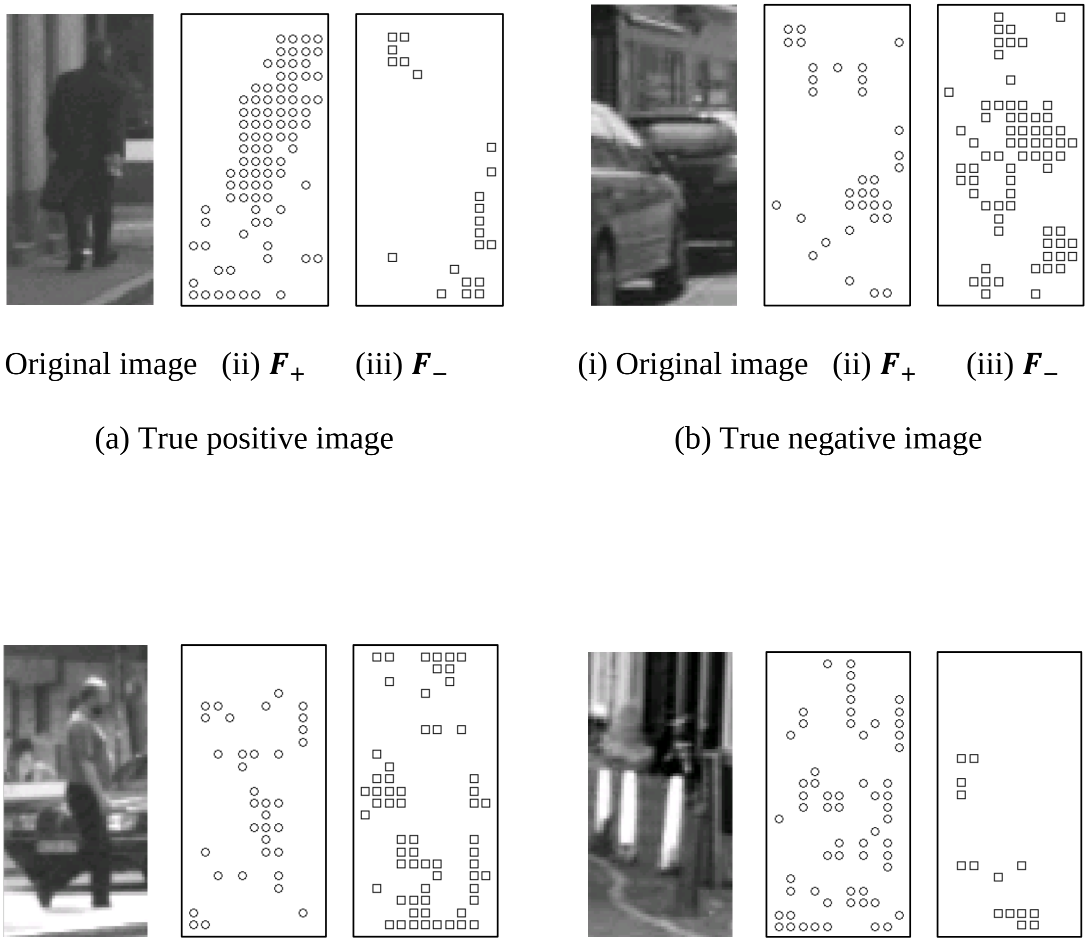
Figure 4.1 shows the result of sorting the difference vector obtained from the process in Figure 3.1(d) with an initial visual dictionary size X = 500. The horizontal axis represents the number of visual words, and the vertical axis represents the difference in the total occurrence frequency of each visual word between the pedestrian and non-pedestrian images. Visual words for which the difference is positive are considered to contribute to the determination of pedestrians, whereas negative values imply that the visual word contributes to the determination of non-pedestrian objects. As shown in Figure 4.1, we defined the top 100 visual words for determining pedestrians as F_{+}, and the 100 visual words that best represent non-pedestrian objects as F_{-}.

4.2 Visualization of selected feature points

Figure 4.2 illustrates examples of the distribution of F_{+} ('○' in the figure) and F_{-} ('□' in the figure) in the case of true positive (TP), true negative (TN), FN, and FP detections.

In the pedestrian images in Figure 4.2(a), it can be see that F_{+} feature points are mainly located about the body, whereas F_{-} are mainly located in the background. In the rejected non-pedestrian image shown in Figure 4.2(b), the distribution of feature points varies widely, but there are many more F_{-} feature points than F_{+}.

In the FN image in Figure 4.2(c), the F_{+} points are again mainly located about the body, but there are fewer than in the TP image. We suspect this is because FN images have complicated backgrounds, which will affect the detection accuracy. In contrast, in the FP image, the F_{+} points are in the majority, and so this image was determined to contain a pedestrian.



(i) Original image (ii) $\boldsymbol{F}_{+}$ (iii) $\boldsymbol{F}_{-}$

(a) True positive image

(i) Original image (ii) $\boldsymbol{F}_{+}$ (iii) $\boldsymbol{F}_{-}$

(b) True negative image

(i) Original image (ii) $\boldsymbol{F}_{+}$ (iii) $\boldsymbol{F}_{-}$

(c) False negative image

(i) Original image (ii) $\boldsymbol{F}_{+}$ (iii) $\boldsymbol{F}_{-}$

(d) False positive image

Figure 4.2. Visualization of selected feature points.

4.3 Average distribution of the selected feature points

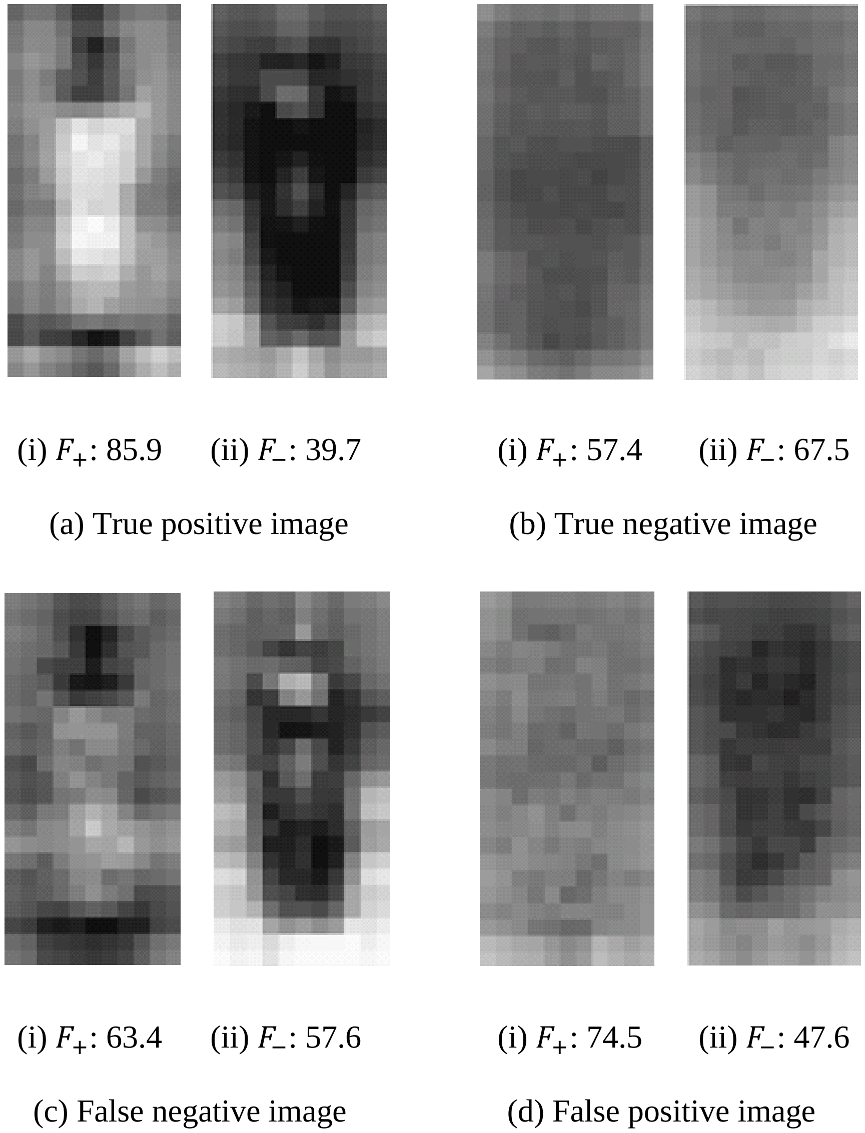
In this section, we analyze the average distribution of selected feature points from TP, TN, FN, and FP images.

We randomly selected 500 correctly classified pedestrian and non-pedestrian images for the TP and TN dataset, and 500 incorrectly classified pedestrian and non-pedestrian images for the FN and FP dataset from the classification results of Section IV-D. We then examined the average number of selected discriminative features of each set.

To compare the difference in the feature point distribution of pedestrian and non-pedestrian images, we divided each detection window into 11 × 22 grids, and computed the average number of feature points in each cell.

Figure 4.3 shows the average distribution of F_+ and F_- feature points in TP, TN, FN, and FP image sets. The white area indicates a large number of feature points in the figure, and the black area indicates the opposite.

The figure shows that, for pedestrian images (TP or FN), F_+ feature points are mainly located about the body, and F_- points are primarily in the background. This is because features located in the background are highly consistent with those extracted from non-pedestrian images. In the non-pedestrian images (TN or FP), both F_+ and F_- are uniformly distributed in the images. This is because the position of F_- changes with the content of the non-pedestrian images.



(i) F_+: 85.9 (ii) F_-: 39.7

(a) True positive image

(i) F_+: 57.4 (ii) F_-: 67.5

(b) True negative image

(i) F_+: 63.4 (ii) F_-: 57.6

(c) False negative image

(i) F_+: 74.5 (ii) F_-: 47.6

(d) False positive image

Figure 4.3. Average distribution and number of selected features.

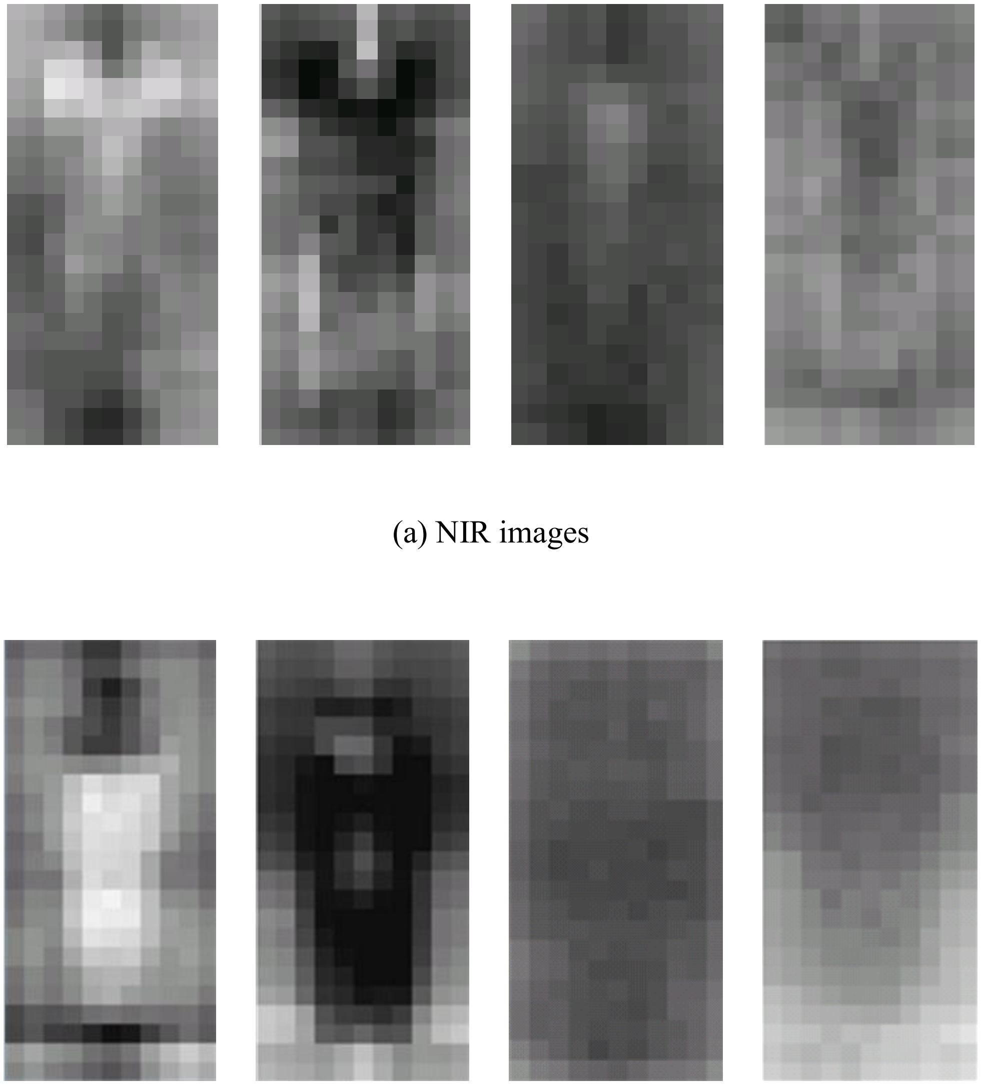
Furthermore, in the pedestrian images, F_+ are mainly located in the lower body areas. We speculate that this is because, in pedestrian images, feature points located around the shoulder areas suffer more interference with the background, so the discriminative features are mainly distributed in the lower part of the body.

Figure 4.3 also shows the average number of selected discriminative feature points for each image set. It can be seen that the average number of F_+ points in TP images (85.9) far outweighs that of F_- (39.7), but in the TN images, the number of F_- points (67.5) outweighs that of F_+ (57.4). This is the greatest difference between the TP and TN images. In contrast, in the FN images, the count of F_+ (63.4) is close to F_- (57.6), whereas in the FP images, F_+ (74.5) far outweighs F_- (47.6).

Overall, it can be said that the number of discriminative features is a significant factor in the detection accuracy. Furthermore, an unknown image can be determined to contain a pedestrian when the F_+ count is greater than the F_- count, but if F_- is in the majority, the classification result is more likely to be a non-pedestrian image.

4.4 Selected feature points in NIR and VS images

In this section the distributions of selected feature points from pedestrian images and non-pedestrian images will be analyzed and the average counts of selected discriminative features of the NIR and VS images are given.



(a) NIR images

(b) VS images

Figure 4.4. Average distribution of selected feature points.

To compare the difference in the feature points distribution between the NIR and VS images, we divided each detected window into 11 × 22 grids, and computed the average value of the number of feature points in each grid.

Figure 4.4 show the visualization of the average distribution of the selected discriminative feature points from 3000 images for each type. The white area indicates a large number of feature points in the figure, and the black area indicates the opposite. In both the NIR and VS images, F_{+} feature points are mainly located in the body areas, F_{-} are mainly located in the background, and in the non-pedestrian images, both F_{+} and F_{-} are uniformly distributed in the images. The reason is that the position of F_{-} changes with the content in the non-pedestrian images.

Furthermore, in the NIR pedestrian images, F_{+} are mainly distributed in the upper body areas, that is, the shoulders and the back, but in the VS pedestrian images, F_{-} are mainly located in the lower body areas. This is the greatest difference between the NIR and VS images. We speculate that this is because in the NIR pedestrian images, since the NIR irradiates, the pedestrian’s back is shown as white and the background areas are mainly black, so the discriminative features are mainly distributed in these regions. But, in VS pedestrian images, the image resolution is higher than that in the NIR images, and the feature points located around the shoulder areas are more easily interfered with by the background, so the discriminative features are mainly distributed in the lower part of the body.

TABLE II.
AVERAGE COUNT OF SELECTED FEATURES

	Positive		Negative	
	F_+	F_-	F_+	F_-
NIR	74.3	30.2	49.6	66.7
VS	85.9	39.7	57.4	67.5

Table 2 gives the average count of the selected discriminative feature points for each modality. From the table it can be seen that the average number of F_+ far outweighs the F_- in the pedestrian images, but in the non-pedestrian images the number of F_- outweighs the F_+, in both the NIR and VS images. In addition, in both of the pedestrian and non-pedestrian images, the average number of F_+ and F_- in VS images is more than in the NIR images. This is due to the VS images having a higher resolution than the NIR images.

4.5 Summary

In this chapter, we investigated the distribution of discriminative feature points belonging to the selected visual words from pedestrian images and non-pedestrian images. As shown by the experimental results, discriminative feature points in the pedestrian images are mainly located in body areas, whereas the feature points are uniformly distributed in non-pedestrian images. In addition, we also investigated the distribution of discriminative feature points from NIR images and VS images, respectively. We found that, in the NIR pedestrian images, the discriminative feature points are mainly distributed in the upper body areas, but in the VS pedestrian images, they are mainly located in the lower body areas.

CHAPTER 5

An Improved Visual Codebook Selection Method by set two thresholds

In this chapter, to minimize its loss of discriminative power, we propose an improved code words selection method based on Bag-of-Features(BoF). We first calculate the difference in the total appearance frequency of each visual word in pedestrian and non-pedestrian images. Then two thresholds are set for selected code words which have greater absolute values. The experiment results show that the proposed method is comparable with state-of-the-art methods for pedestrian detection. Furthermore, the effectiveness of the proposed method is validated by analyzing the distribution of selected feature points.

We have proposed a method to reduce the dimension of the classifier [49]. In [49], the difference in the total appearance frequency for each visual word of the pedestrian and non-pedestrian images are calculated. The visual codebook that exhibit greater absolute values are considered to be more efficient for pedestrian detection, and are selected. However, this method can only obtain a total number of visual codebook by setting one limit value to measure the effective of visual codebook. But selected visual codebook includes discriminative visual word for both pedestrian and non-pedestrian images. Just one limit value is different to know how many visual word have discrimination for pedestrian and non-pedestrian, respectively. In this paper, we propose an improved method to reduce the dimension of the classifier by setting two limit value to remove irrelevant and redundant visual codebook. By using two variable limit value, it can be obtained the optimal number of visual word for pedestrian and non-pedestrian, respectively. Experimental study shows that the compact visual codebook created in this way can achieve excellent classification performance even after a considerable reduction in size.

5.1 Overview of our approach

In this chapter, we investigated the distribution of discriminative feature points belonging to the selected visual words from pedestrian images and non-pedestrian images.

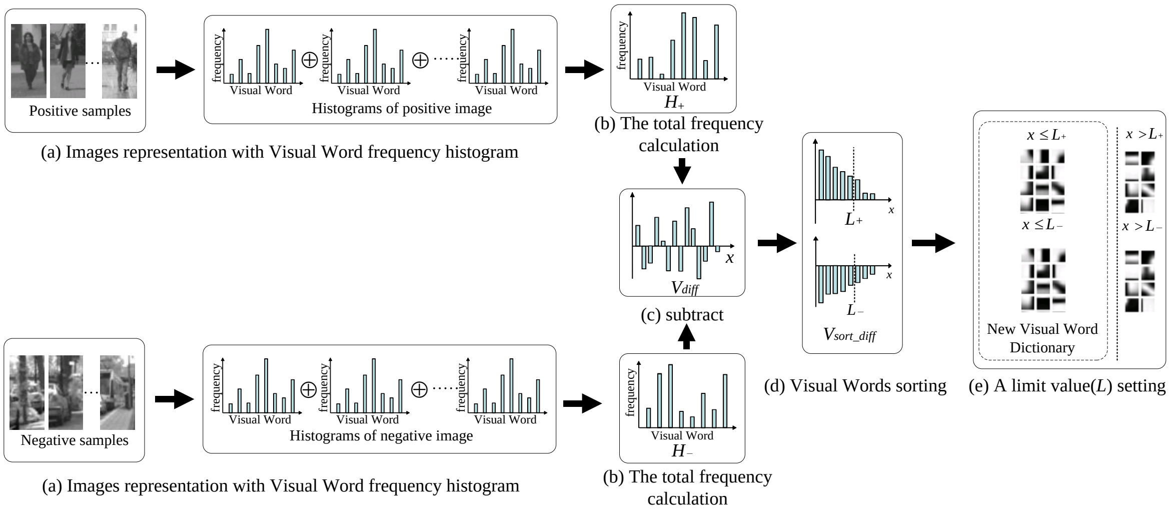

Positive samples
frequency
Visual Word
Histograms of positive image
H_+
(a) Images representation with Visual Word frequency histogram
(b) The total frequency calculation
V_{diff}
x
(c) subtract
L_+
L_-
V_{sort_diff}
(d) Visual Words sorting
$x \le L_+$
$x > L_+$
$x \le L_-$
$x > L_-$
New Visual Word Dictionary
(e) A limit value(L) setting
Negative samples
Histograms of negative image
H_-
(a) Images representation with Visual Word frequency histogram
(b) The total frequency calculation

Figure 5.1. Flowchart of the proposed method.

As mentioned at [49], there is a disadvantage of the dense regular grid is that a large number of redundant features are included in the visual vocabulary. A compact visual codebook has advantages in both computational efficiency and memory usage. For example, when linear or nonlinear SVMs are used, the complexity of computing the kernel matrix, testing a new image, or storing the support vectors is all proportional to the codebook size. Also, many algorithms working well in a low dimensional space will encounter difficulties such as singularity or unreliable parameter estimate when the dimensions increase. A compact visual codebook provides a lower-dimensional representation and can effectively avoid these problems. In other words, creating a compact codebook is essentially a dimensionality reduction problem.

We have proposed a method to reduce the dimension of the classifier by setting a limit value to remove irrelevant and redundant visual words. Firstly, the difference in the total appearance frequency for each visual word of the pedestrian and non-pedestrian images are calculated. The visual codebook that exhibit greater absolute values are considered to be more efficient for pedestrian detection, and are selected. However, this method can only obtain a total number of visual codebook by setting one limit value to measure the effective of visual codebook. But selected visual codebook includes discriminative visual word for both pedestrian and non-pedestrian images. Just one limit value is different to know how many visual word have discrimination for pedestrian and non-pedestrian, respectively. Hence, we propose an improved method to reduce the dimension of the classifier by setting two limit value to select the discriminative visual words for pedestrian and non-pedestrian, respectively. By using two variable limit value, it can be obtained the optimal number of visual word for pedestrian and non-pedestrian, respectively. Details of each process are described below.

A brief overview of this approach is given in Fig. 3. First, the quantization histograms obtained from each training image are divided into positive images and negative images (Fig. 5.1 (a)). The total frequency histograms for positive sample H_+ and negative sample $H_-$are then computed by the following equation, as shown in Fig. 5.1 (b).

$$\begin{cases} H_+(x) = \sum_{x \in X \cap y_\varepsilon = +1} h(x) \\ H_-(x) = \sum_{x \in X \cap y_\varepsilon = -1} h(x) \end{cases} \quad 2)$$

where ε is the number of training samples, and X is the number of visual words in a dictionary. y_ε represents the label of the pedestrian or non-pedestrian in the training sample $y \in \{+1, -1\}$.

Then we normalize two total frequency histograms as

$$\begin{cases} H_+(x) = \frac{H_+(x)}{\sum_{i=1}^{X} (H_+(i))} \\ H_-(x) = \frac{H_-(x)}{\sum_{i=1}^{X} (H_-(i))} \end{cases} \quad 3)$$

And the difference between $H_+(x)$ and $H_-(x)$ is calculated to obtain the difference vector (Fig. 5.1 (c))

$$V_{diff}(x) = H_+(x) - H_-(x) \quad 4)$$

If $V_{diff}(x)$ is positive, this visual word is effectively classified as a positive sample, and vice versa. The larger the absolute value of $V_{diff}(x)$, the more beneficial the x-th feature to the classification.

And then, the corresponding visual codebook which V_{diff} value are positive and negative are sorted by descending order of absolute value, respectively. And two limit values L_+ and L_- are set to determine the expected size of the new visual vocabulary to be preserved (Fig. 5.1 (d)). Visual words for which $V_{diff}(x)$ is below the limit value L_+ and L_- are considered to be redundant, and are screened out of the original dictionary. The remaining $L_+ + L_-$ visual words comprise a new visual vocabulary (shown in Fig. 5.1 (e)). Next, the corresponding dimensions of the original histograms h are removed according to the new visual vocabulary. The new frequency histogram of the visual vocabulary forms the input to the classifier, which is trained by the SVM.

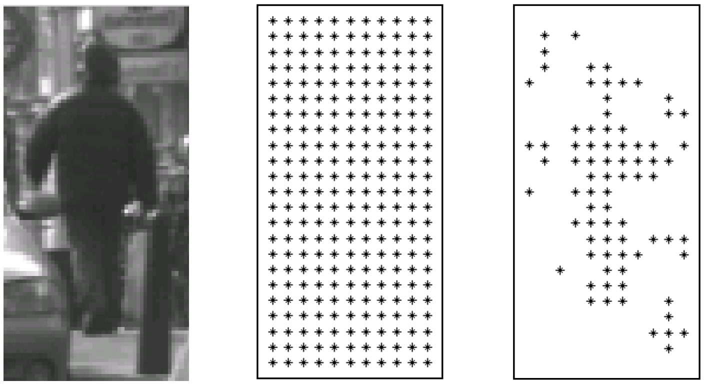


Figure 5.2. The selection and visualization examples of features.

Fig.5.2 shows an example of the distribution of selected feature points by our method. (a) is an original image of a pedestrian. (b) is dense-SIFT feature points distribution which extracted from the image (a). (c) is selected feature points distribution by our method.

5.2 Experiments and results

In this section, we evaluate the performance of our proposed method in terms of its detection rate. In addition, we analyze the distribution of discriminative features by visualizing the selected features.

5.2.1 Detection accuracy with various sizes of selected visual word

In this section, we describe the relation between the limit size L_+ and L_- of the visual vocabulary and the detection accuracy.

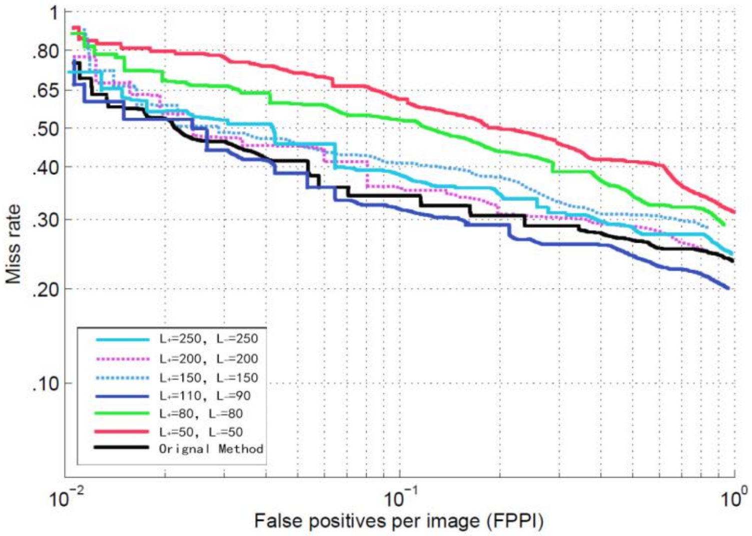


Figure 5.3. Performance of our method.

We performed the standard per-image evaluation used in pedestrian detection [2]. We added a so-called Candidate Generation Pruning (CGP) step to our system, in order to obtain a fair comparison with the best performer in this dataset [48]. Making use of projective geometry, the CGP algorithm forecasting the possible arisen location so as to confirm the scope of target searching. This permits us to both accelerate the detection of pedestrians (as fewer windows are evaluated by the classifier), and remove false positives standing on non-plausible locations of the targets, thus improving the resulting accuracy.

Using an optimal initial visual vocabulary size X = 500 for training and testing images, the value of L_+ and L_- was varied from 250 to 0, respectively. In our experiments, we found that the accuracy variation was very small when $\mathrm{L}_+, \mathrm{L}_- > 100$, and when $\mathrm{L}_+ = 110, \mathrm{L}_- = 90$ with negligible loss in detection accuracy, which means that 200 efficient visual codebook in the original visual vocabulary can result in almost the same performance as with all 500 visual words. In addition, the detection precision decreases more quickly when $\mathrm{L}_+ < 110, \mathrm{L}_- < 90$. The detection accuracy at different L_+ and L_- values is shown in Fig. 6.

Furthermore, the results show that the miss rate reduce 3% by proposed method than [49].This confirms that the proposed method is effective for pedestrian detection applications.

5.3 Summary

In this chapter, we propose an efficient code words selection method based on Bag-of-Features(BoF), which can be applied to the pedestrian detection problem. Our visual word selection method calculates the difference in the total appearance frequency of each visual word in pedestrian and non-pedestrian images. Then two thresholds are set for selected code words which have greater absolute values. The experiment results show that the proposed method is comparable with state-of-the-art methods for pedestrian detection. Furthermore, the effectiveness of the proposed method is validated by analyzing the distribution of selected feature points. As shown by the experimental results, discriminative feature points in the pedestrian images are mainly located in body areas, whereas the feature points are uniformly distributed in non-pedestrian images. The experiments also showed that the miss rate reduce 3% by our proposed method than original method.

CHAPTER 6

Conclusions

Pedestrian protection systems represent a key technology to reduce the number of accidents between pedestrians and vehicles. Given the difficulties that such systems shall overcome, that is, realtime detection of changing targets in uncontrolled outdoor scenarios, pedestrian protection is by no means an easy task. From our point of view, research was too focused on specific tasks of the system like classification and forgot the relation between them. In this thesis we have developed the research from a global viewpoint.

6.1 Summary and contributions

Although each chapter contains a specific discussion section that analyses and points out the most relevant advantages and disadvantages of the explored algorithms and of their combination between the modules of the proposed architecture, in this chapter we highlight some more global conclusions, which are in fact linked with the contribution of the thesis.

In the survey of the state of the art, we have extensively reviewed the literature by first introducing a general architecture that consists of different modules, each with its own objectives, in which we fit every analyzed technique in the literature. As has been seen, this general architecture is of crucial importance to analyze the literature in a sensible and ordered way. In the chapter we have highlighted a set of interesting points that lead the thesis studies and in fact will lead the future lines of research, like problems usually omitted in the literature (e.g., foreground segmentation) or techniques that have demonstrated their classification capabilities (e.g., histograms of oriented gradients).

According to the review, it can be said that there is a clear research trend in every module. For example, the promising algorithms in foreground segmentation are the road based ones; the research in classification is mostly focused on gradient-based features and several typical learning algorithms, but recent multiclass/multipart approaches are also gaining importance; and the Kalman filter is the most used algorithm for the tracking module. In the survey we have also highlighted the lack of datasets. In order to evaluate all the different proposals, we have introduced a pioneer multi-purpose dataset aimed at being utilized as an evaluation framework for different modules of a PPS: foreground segmentation, classification and whole system.

In addition, we presented a method of obtaining a compact and discriminative visual vocabulary for pedestrian detection. Our visual word selection method calculates the difference in the total appearance frequency of each visual word in pedestrian and non-pedestrian images. The visual words that exhibit greater absolute values are considered to be more efficient for pedestrian detection, and are thus selected. The experiments also showed that the learning and detection process can achieve similar precision in about 50% of the time using the proposed method, even if only 40% of the visual words are selected. Furthermore, we detect pedestrians at night by near-infrared (NIR) cameras. We found our method is also For the both of two datasets, our detector is competitive in terms of the detection quality with other state-of-the-art.

we investigated the distribution of discriminative feature points belonging to the selected visual words from pedestrian images and non-pedestrian images. As shown by the experimental results, discriminative feature points in the pedestrian images are mainly located in body areas, whereas the feature points are uniformly distributed in non-pedestrian images. In addition, we also investigated the distribution of discriminative feature points from NIR images and VS images, respectively. We found that, in the NIR pedestrian images, the discriminative feature points are mainly distributed in the upper body areas, but in the VS pedestrian images, they are mainly located in the lower body areas.

6.2 Perspectives

Driver assistance systems, and particularly pedestrian protection systems, are a very young area of research. Hence, the future research possibilities are so numerous and diverse that they can easily occupy a chapter on its own. We condense the lines we consider of key importance in a few general points.

Short term challenges. The pursuit of a perfect PPS based on Computer Vision only shall be considered a long term goal. The development of a PPS that works under restricted conditions is already useful. For instance, a system that works only at daytime, under good weather conditions (no heavy rain/snow/fog), over a range of distances up to 50 m is, from our viewpoint, the first intermediate challenge for the community. According to [4], these conditions represent a very relevant scenario in accidents.

Long term goals: focus on the real problem. It is clear that many new proposals are tested on too easy data. Developing systems capable of working under restricted conditions is different from developing techniques that just work on high-resolution near non-occluded pedestrians, because this can lead to a loss of perspective of the real problem. Although it can seem strange to provide statistics of 10% DR or 10 FPPI, specially in front of other more traditional areas like face detection or object classification, this poor performance in realistic complex examples is more useful for the community than presenting a 99% in nearly toy examples.

Face the problem globally. In addition to developing the individual parts of a PPS, which is one of the keys to reach good performance rates, a global view of the problem can lead us to interesting conclusions as the ones assessed in this thesis.

Overall vision of future ADAS. When talking about a global viewpoint, one also has to have in mind that a PPS is likely to work with many other ADAS. This leads to a point in which the different systems have to share sensors and computation time, which in fact has both disadvantages such as the restrictions when choosing sensors, but also advantages in the sense that techniques like stereo reconstruction, free space analysis and even sensor fusion data can be shared between them.

APPENDIX A

Performance Evaluation

The statistical validation of any decision process is crucial to determine the performance of a wide variety of applications, for example from medical treatments to spam filtering or object detection. Let us define the null hypothesis as the default state of some phenomenon, for instance that a patient does not have a disease or that a given window in an image contains just background clutter. If we label the natural state of the null hypothesis as a negative, then the opposite state (i.e., the patient does have a disease and a pedestrian has been found in the image) is referred to as a positive. Since the decision process is aimed at rejecting or not rejecting the null hypothesis, then there exist two basic sources of error:

- false positives (FP) when the null hypothesis is incorrectly rejected (i.e., finding a disease in a healthy patient or detecting a pedestrian in a background image), or
- false negatives (FN) when the null hypothesis is incorrectly not rejected (i.e., finding healthy an ill patient or failing to find a pedestrian when in fact there is one).

Given these two errors, the performance measurement of a decision process consists in counting the number of FP and FN in the context of for example the total number of real positives or negatives, the total number of decisions, etc. On the contrary, a true positive (TP) is found if the null hypothesis is correctly rejected, whereas a true negative (TN) is found when the null hypothesis is correctly not rejected. Figure D.1 illustrates these concepts.

A.1 Visualization of selected feature points

Although basic measurements can be defined as real numbers, for example in terms of true positive rate (number of true positives out of the total number of positives), most classification algorithms typically provide a confidence value that shall be thresholded to take a decision. Thus, by varying such threshold we can plot different curves that show the classifiers performance in terms of the behavior of the basic measurements.

The Receiver Operating Characteristic (ROC) curve takes two measures into account, false positive rate (FPR) and true positive rate (TPR):

$$FPR = \frac{FP}{TN + FP} \tag{A.1}$$

$$TPR = \frac{TP}{TP + FN} \tag{A.2}$$

where FPR is plotted on the x-axis and TPR on the y-axis. The perfect classifier would have TPR=1 and FPR=0, placing the performance point at the top-left corner of the ROC. This curve has been used in a large number of papers [114, 68]. Dalal et al. [35] makes use of a complementary curve, called Detection Error Trade-off (DET), which plots Miss Rate (MR = 1 − TPR) on the y-axis and FPR on x-axis, both axes using a logarithmic scale instead of a linear one.

Another widely used plot is Recall-Precision (RP) curve:

$$Recall = TPR \tag{A.3}$$

$$Precision = \frac{TP}{TP + FP} \tag{A.4}$$

with Recall on the x-axis and Precision on the y-axis, although sometimes their positions are interchanged and 1−Precision is used instead of the regular Precision. The perfect classifier in this case would have Recall= 1 and Precision= 1, which means that neither false positives nor false negatives exist. For instance, this curve is used in [47].

The final curve is Sensitivity-Specificity, defined as

$$Sensitivity = TPR \tag{A.5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{A.6}$$

plotting sensitivity on x-axis and specificity on y-axis. Again, both Sensitivity and Specificity would be 1 in a perfect classifier.

A.2 Window-based versus image-based evaluation

The common procedure while evaluating classifiers (either for medical tests or object classification, for instance) is to select a training set containing both positive and negative samples and a testing set with also positives and negatives. In order to plot a typical curve like ROC we just have to count the number of TP, FN, TN and FP on the testing set and compute the corresponding rates. This is possible because all the measures are well defined. In the case of pedestrian classification, a common procedure is to work with cropped images, say 1 000 test positives and 5 000 test negatives. However, some authors (e.g., [35]) just provide pedestrian-free images to randomly crop (both train and test) negatives, so the set is not well defined, that is, it is difficult to reproduce the exact results that the researchers present. In this case, different authors count False Positives Per Window (FPPW), which is equivalent to FPR but leave open the number of test negatives. If a range of confidence testing thresholds to count TPR and FPPW is used, this performance measure leads to a ROC in which the axes are TPR and FPPW, this latter one often using a logarithmic scale. We call this evaluation window-based. As it is made in the literature, we analyze the whole curve but special attention will be put to the point of the curve, which represents one false positive each 10,000 tested negatives.

Another type of evaluation is focused on detection rather than classification, i.e., placing the performance in the context of frames rather than on isolated examples. In this case, since there are not cropped samples anymore, the way that a detection is considered as a TP or FP depends on the similarity of the detection with the annotations of a set of test frames. In this case, a ROC has the axes TPR and false positives per image (FPPI). The most used similarity measure at this moment is the detection-annotation overlap: a detection d is marked as TP if its overlap with a window annotation a exceeds a certain threshold Γ, where

$$overlap(d,a) = \frac{area(a \bigcap d)}{area(a \bigcup d)} \tag{A.7}$$

otherwise the detection is marked FP. This evaluation is called image-based. In this case, the curve shall be read in a more global manner than in the image-based in the sense that the preferred FPPI working range will depend on the requirements for the given module/system. We focus on the DR value when FPPI= 100, which corresponds to one average single false positive per frame, which is assumable for a system consisting of a classifier and cluster given that a tracking process is likely to absorb most of the spurious false positives. In this thesis, the case in which multiple detections fulfilling this criterion for a single annotation is treated differently if we evaluate the classifier or the whole system (also the clustering). For example, in the case of the system, each annotation account for just a TP, the additional detections associated to the annotation are marked as FP, whereas in the case of the classification they are all marked as TP since there is not any clustering technique involved. This is detailed in the corresponding chapters.

APPENDIX B

List of Acronyms

ABS	antilock braking system
ACC	adaptive cruise control
ADAS	advanced driver assistance systems
AFL	advanced front lighting
CGP	candidate generation performance
CPA	candidates per annotation
DR	detection rate
EOH	edge orientation histograms
ESC	electronic stability control
FN	false negatives
FP	false positives
FPPI	false positives per image
FPPW	false positives per window
FPR	false positives rate
HF	Haar feature
HFOV	horizontal field of view
HOG	histograms of oriented gradients
II	integral image
NIR	near infrared

NN	neural network
NPC	non pedestrian candidates
PPS	pedestrian protection system
ROI	region of interest
SIFT	scale invariant feature transform
SVM	support vector machine
SHOG	simplified HOG
TIR	thermal infrared
TN	true negatives
TP	true positives
TPR	true positives rate
VFOV	vertical field of view
VS	visible spectrum

APPENDIX C

List of State of the Art

Method	MR	Family	Features	Classifier	Context	Deep	Parts	M-Scales	More data	Feat. type
VJ[50]	94.73%	DF	✓	✓						Haar
Shapelet [51]	91.37%	−	✓							Gradients
PoseInv [31]	86.32%	−					✓			HOG
LatSvm-V1 [52]	79.78%	DPM					✓			HOG
ConvNet[53]	77.20%	DN				✓				Pixels
FtrMine [14]	74.42%	DF	✓							HOG+Color
HikSvm [54]	73.39%	−		✓						HOG
HOG [15]	68.46%	−	✓	✓						HOG
MultiFtr [55]	68.26%	DF	✓	✓						HOG+Haar
HogLbp[22]	67.77%	−	✓							HOG+LBP
AFS+Geo[56]	66.76%	−			✓					Custom
AFS[56]	65.38%	−								Custom
LatSvm-V2[36]	65.38%	DPM		✓			✓			HOG
Pls[57]	62.10%	−	✓	✓						Custom
MLS[58]	61.03%	DF	✓							HOG
MultiFtr+CSS[21]	60.89%	DF	✓							Many
FeatSynth[56]	60.16%	−	✓	✓						Custom
pAUCBoost[59]	59.66%	DF	✓	✓						HOG+COV
FPDW[60]	57.40%	DF								HOG+LUV
ChnFtrs[23]	56.34%	DF	✓	✓						HOG+LUV
CrossTalk[61]	53.88%	DF			✓					HOG+LUV
DBN─Isol[62]	53.14%	DN				✓				HOG
ACF[63]	51.36%	DF	✓							HOG+LUV
RandForest [64]	51.17%	DF		✓						HOG+LBP
MultiFtr+Motion[21]	50.88%	DF	✓						✓	Many+Flow
SquaresChnFtrs[65]	50.17%	DF	✓							HOG+LUV
Franken[66]	48.68%	DF			✓					HOG+LUV
MultiResC[48]	48.45%	DPM			✓		✓	✓		HOG
Roerei [65]	48.35%	DF	✓					✓		HOG+LUV
DBN─Mut[67]	48.22%	DN			✓		✓			HOG

Method	MR	Family	Features	Classifier	Context	Deep	Parts	M-Scales	More data	Feat. type
MF+Motion+2Ped[68]	46.44%	DF			✓				✓	Many+Flow
MOCO[69]	45.53%	−	✓		✓					HOG+LBP
MultiSDP[70]	45.39%	DN	✓		✓	✓				HOG+CSS
ACF-Caltech[63]	44.22%	DF	✓							HOG+LUV
MultiResC+2Ped[68]	43.42%	DPM			✓		✓	✓		HOG
WordChannels[71]	42.30%	DF	✓							Many
MT-DPM[1]	40.54%	DPM					✓	✓		HOG
JointDeep[70]	39.32%	DN			✓					Color+Gradient
SDN[72]	37.87%	DN				✓	✓			Pixels
MT-DPM+Context[1]	37.64%	DPM			✓		✓	✓		HOG
ACF+SDt[73]	37.34%	DF	✓						✓	ACF+Flow
SquaresChnFtrs[65]	34.81%	DF	✓							HOG+LUV
InformedHaar[74]	34.60%	DF	✓							HOG+LUV

REFERENCES

[1] J. Yan, X. Zhang, and Z. Lei, “Robust multi-resolution pedestrian detection in traffic scenes,” *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3033–3040, 2013.

[2] P. Dollar and C. Wojek, “Pedestrian detection: An evaluation of the state of the art,” *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 743–761, 2012.

[3] R. Benenson, O. Mohamed, J. Hosang, and B. Schiele, “Ten Years of Pedestrian Detection , What Have We Learned ?,” in *European Conference on Computer Vision - ECCV*, 2014.

[4] V. P. Detection, “Universitat Autònoma de Barcelona A Global Approach to Vision-Based Pedestrian Detection for Advanced Driver Assistance Systems,” 2009.

[5] M. Cameron, “World Report on Road Traffic Injury Prevention.,” *Inj. Prev.*, vol. 10, pp. 255–256, 2004.

[6] E. D. Dickmanns and A. Zapp, “A Curvature-based Scheme for Improving Road Vehicle Guidance by Computer Vision,” *Proc. SPIE*, vol. 0727. pp. 161–168, 1987.

[7] J. Ashton and G. M. Mackay, “Benefits from Changes in Vehicle Exterior Design-Field Accident and Experimental Work in Europe,” *Injury*, 2000.

[8] Q. Zhu, S. Avidan, M. C. Yeh, and K. T. Cheng, “Fast Human Detection Using a Cascade of Histograms of Oriented Gradients,” *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2. pp. 1491–1498, 2006.

[9] B. Wu and R. Nevatia, “Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors,” *Int. J. Comput. Vis.*, vol. 75, no. 2, pp. 247–266, Jan. 2007.

[10] S. K. Nayar and V. Branzoi, “Adaptive dynamic range imaging: optical control of pixel exposures over space and time,” *Proc. Ninth IEEE Int. Conf. Comput. Vis.*, 2003.

[11] P. Knoll, “HDR vision for driver assistance,” in *High-Dynamic-Range (HDR) Vision*, 2007, pp. 123–136.

[12] D. Gerónimo, A. M. López, A. D. Sappa, and T. Graf, “Survey of pedestrian detection for advanced driver assistance systems,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, pp. 1239–1258, 2010.

[13] L. Bombini, P. Cerri, P. Grisleri, S. Scaffardi, and P. Zani, “An evaluation of monocular image stabilization algorithms for automotive applications,” *2006 IEEE Intell. Transp. Syst. Conf.*, 2006.

[14] R. Benenson, M. Mathias, R. Timofte, L. Van Gool, E. Ibbt, and K. U. Leuven, “Pedestrian detection at 100 frames per second.”

[15] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, pp. 886–893, 2005.

[16] M. Enzweiler and D. M. Gavrila, “A multilevel mixture-of-experts framework for pedestrian classification,” *IEEE Trans. Image Process.*, vol. 20, pp. 2967–2979, 2011.

[17] D. M. Gavrila and S. Munder, “Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle,” *Int. J. Comput. Vis.*, vol. 73, no. 1, pp. 41–59, Jul. 2006.

[18] M. Mählisch, M. Oberländer, O. Löhlein, D. Gavrila, and W. Ritter, “A multiple detector approach to low-resolution FIR pedestrian recognition,” in *IEEE Intelligent Vehicles Symposium, Proceedings*, 2005, vol. 2005, pp. 325–330.

[19] H. Nanda and L. Davis, “Probabilistic template based pedestrian detection in infrared videos,” *Intell. Veh. Symp. 2002. IEEE*, vol. 1, 2002.

[20] D. M. Gavrila, “Pedestrian Detection from a Moving Vehicle,” *ECCV Eur. Conf. Comput. Vis.*, vol. 2, pp. 37–49, 2000.

[21] S. Walk, N. Majer, K. Schindler, and B. Schiele, “New features and insights for pedestrian detection,” *2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1030–1037, Jun. 2010.

[22] X. Wang, T. X. Han, and S. Yan, “An HOG-LBP human detector with partial occlusion handling,” *2009 IEEE 12th Int. Conf. Comput. Vis.*, pp. 32–39, Sep. 2009.

[23] P. Dollár, “Integral Channel Features,” pp. 1–11.

[24] D. M. Gavrila, J. Giebel, and S. Munder, “Vision-based pedestrian detection: the PROTECTOR system,” *IEEE Intell. Veh. Symp. 2004*, 2004.

[25] D. Gerónimo, A. D. Sappa, A. López, and D. Ponsa, "Adaptive image sampling and windows classification for on-board pedestrian detection," *Proc. Int. Conf. Comput. Vis. Syst.*, 2007.

[26] S. Munder and D. M. Gavrila, "An experimental study on pedestrian classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, pp. 1863–1868, 2006.

[27] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[28] B. Wu and R. Nevatia, "Detection and Segmentation of Multiple, Partially Occluded Objects by Grouping, Merging, Assigning Part Detection Responses," *Int. J. Comput. Vis.*, vol. 82, no. 2, pp. 185–204, Dec. 2008.

[29] B. Wu and R. Nevatia, "Pedestrian Detection in Infrared Images based on Local Shape Features," *2007 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1–8, Jun. 2007.

[30] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–20, 2009.

[31] Z. Lin and L. S. Davis, "A pose-invariant descriptor for human detection and segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2008, vol. 5305 LNCS, pp. 423–436.

[32] B. Leibe, A. Leonardis, and B. Schiele, "Robust Object Detection with Interleaved Categorization and Segmentation," *Int. J. Comput. Vis.*, vol. 77, no. 1–3, pp. 259–289, Nov. 2007.

[33] E. Seemann, B. Leibe, and B. Schiele, "Multi-aspect detection of articulated objects," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 1582–1588.

[34] C. Wojek, G. Dorkó, A. Schulz, and B. Schiele, "Sliding-windows for rapid object class localization: A parallel technique," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 2008, vol. 5096 LNCS, pp. 71–81.

[35] M. P. McKay, T. Thoma, C. Kahn, and C. S. Gotschall, "National Highway Traffic Safety Administration (NHTSA) Notes," *Ann. Emerg. Med.*, vol. 52, pp. 453–454, 2008.

[36] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models.,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–45, Sep. 2010.

[37] D. Mochizuki, Y. Yano, T. Hashiyama, and S. Okuma, “Pedestrian detection with a vehicle camera using fast template matching based on background elimination and active search,” *Electron. Commun. Japan (Part II Electron.*, vol. 90, no. 10, pp. 115–126, Oct. 2007.

[38] G. Csurka and C. Dance, “Visual categorization with bags of keypoints,” *Proc. Eur. Conf. Comput. Vis.*, pp. 59–74, 2004.

[39] J. Uijlings, “Real-time visual concept classification,” *IEEE Trans. Multimed.*, vol. 12, no. 7, pp. 665–681, 2010.

[40] X. Han and Y. Chen, “Image Categorization by Learned PCA Subspace of Combined Visual-words and Low-level Features,” pp. 7–10, 2009.

[41] J. a Hartigan and M. a Wong, “Algorithm AS 136: A K-Means Clustering Algorithm,” *Appl. Stat.*, vol. 28, p. 100, 1979.

[42] B. J. Frey and D. Dueck, “Clustering by Passing Messages Between Data Points,” *Science*, vol. 315. pp. 972–976, 2007.

[43] T. Kohonen, “Self-organized formation of topologically correct feature maps,” *Biological Cybernetics*, vol. 43. pp. 59–69, 1982.

[44] M. N. Ahmed, S. M. Yamany, N. Mohamed, A. A. Farag, and T. Moriarty, “A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data,” *IEEE Trans. Med. Imaging*, vol. 21, pp. 193–199, 2002.

[45] M. Varma and A. Zisserman, “A statistical approach to texture classification from single images,” *Int. J. Comput. Vis.*, vol. 62, pp. 61–81, 2005.

[46] L. Wang, L. Zhou, C. Shen, L. Liu, and H. Liu, “A hierarchical word-merging algorithm with class separability measure,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 417–435, 2014.

[47] M. Enzweiler, S. Member, and D. M. Gavrila, “Monocular Pedestrian Detection : Survey and Experiments,” vol. 31, no. 12, pp. 2179–2195, 2009.

[48] C. Park, Dennis and Ramanan, Deva and Fowlkes, “Multiresolution models for object detection,” *Comput. Vis. – ECCV 2010*, vol. 6314, pp. 1–14, 2010.

[49] X. Zhang, G. Chen, K. Saruta, and Y. Terata, “A Simple Visual Words Selection Strategy for Pedestrian Detection,” *Adv. Vis. Comput.*, pp. 658–667, 2014.

[50] P. Viola and M. Jones, “Robust real-time face detection,” *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[51] P. Sabzmeydani and G. Mori, “Detecting Pedestrians by Learning Shapelet Features,” *2007 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1–8, Jun. 2007.

[52] P. Felzenszwalb, D. Mcallester, D. Ramanan, and U. C. Irvine, “A Discriminatively Trained , Multiscale , Deformable Part Model.”

[53] P. Sermanet, “Pedestrian Detection with Unsupervised Multi-Stage Feature Learning.”

[54] S. Maji, U. C. Berkeley, and A. C. Berg, “Classification using Intersection Kernel Support Vector Machines is Efficient ∗,” 2008.

[55] C. Wojek and B. Schiele, “A Performance Evaluation of Single and Multi-feature People Detection,” pp. 82–91, 2008.

[56] D. Levi, S. Silberstein, and A. Bar-Hillel, “Fast multiple-part based object detection using KD-ferns,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 947–954, 2013.

[57] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, “Human detection using partial least squares analysis,” *Comput. Vision, 2009 IEEE 12th Int. Conf.*, no. Iccv, 2009.

[58] W. Nam, B. Han, and J. H. Han, “Improving object localization using macrofeature layout selection,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 1801–1808.

[59] S. Paisitkriangkrai, C. Shen, and A. Van Den Hengel, “Efficient Pedestrian Detection by Directly Optimizing the Partial Area under the ROC Curve,” *2013 IEEE Int. Conf. Comput. Vis.*, pp. 1057–1064, 2013.

[60] P. Dollar, S. Belongie, and P. Perona, “The Fastest Pedestrian Detector in the West,” *Procedings Br. Mach. Vis. Conf. 2010*, pp. 68.1–68.11, 2010.

[61] P. Doll, “Crosstalk Cascades for Frame-Rate Pedestrian Detection,” pp. 1–14.

[62] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3258–3265.

[63] R. Appel, S. Belongie, P. Perona, and P. Doll, "Fast Feature Pyramids for Object Detection," pp. 1–14.

[64] J. Marin, D. Vazquez, A. M. Lopez, J. Amores, and B. Leibe, "Random forests of local experts for pedestrian detection," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2592–2599, 2013.

[65] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool, "Seeking the strongest rigid detector," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3666–3673.

[66] M. Mathias, R. Benenson, R. Timofte, and L. Van Gool, "Handling Occlusions with Franken-Classifiers," *2013 IEEE Int. Conf. Comput. Vis.*, pp. 1505–1512, 2013.

[67] W. Ouyang, X. Zeng, and X. Wang, "Modeling mutual visibility relationship in pedestrian detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3222–3229.

[68] W. Ouyang and X. Wang, "Single-Pedestrian Detection aided by Multi-pedestrian Detection."

[69] G. Chen, Y. Ding, J. Xiao, and T. X. Han, "Detection Evolution with Multi-order Contextual Co-occurrence," *2013 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1798–1805, Jun. 2013.

[70] X. Zeng, W. Ouyang, and X. Wang, "Multi-stage Contextual Deep Learning for Pedestrian Detection," *2013 IEEE Int. Conf. Comput. Vis.*, pp. 121–128, 2013.

[71] A. D. Costea, "Word Channel Based Multiscale Pedestrian Detection Without Image Resizing and Using Only One Classifier," pp. 4321–4328, 2014.

[72] P. Luo, Y. Tian, X. Wang, and X. Tang, "Switchable Deep Network for Pedestrian Detection," *Conf. Comput. Vis. Pattern Recognit.*, pp. 899–906, 2014.

[73] D. Park, U. C. Irvine, C. L. Zitnick, D. Ramanan, and P. Doll, "Exploring Weak Stabilization for Motion Feature Extraction," vol. 1, no. c.

[74] S. Zhang, C. Bauckhage, and A. B. Cremers, "Informed Haar-like Features Improve Pedestrian Detection," in *CVPR*, 2014.