2013

Pedestrian detection in far infrared images

Olmeda Reino, Daniel

Universidad Carlos III de Madrid
Escuela Politécnica Superior

Departamento de Ingeniería de Sistemas y Automática
Doctorado en Ingeniería Eléctrica, electrónica y Automática

Tesis Doctoral

Pedestrian Detection in Far Infrared Images

Daniel Olmeda Reino

Dirigida por

José María Armingol

Arturo de la Escalera

Madrid, 2013

Web page: http://www.uc3m.es/islab

E-mail: dolmeda@ing.uc3m.es

Address:

Laboratorio de Sistemas Inteligentes
Departamento de Ingeniería de Sistemas y Automática
Escuela Politécnica Superior
Universidad Carlos III de Madrid
C/ Butarque, 15
Leganés 28991 — Spain

Pedestrian Detection in Infrared Images

Autor: Daniel Olmeda Reino
Directores: José María Armingol
Arturo de la Escalera

Firma del Tribunal Calificador:

	Nombre y Apellidos	Firma
Presidente: D.	.....	.....
Vocal: D.	.....	.....
Secretario: D.	.....	.....

Calificación: .....

Leganés, ... de ..... de 2014.

A Verónica

Contents

Abstract xv

Resumen xvii

Agradecimientos xix

1. Introduction 1

1.1. Motivation . . . 2
1.2. Advanced Driver Assistance Systems . . . 7
1.3. People safety in Intelligent Transportation Systems . . . 8
1.4. Thermal Imaging . . . 10
1.5. IVVI Research Platform . . . 11
1.6. Objectives . . . 13
1.7. Outline of the dissertation . . . 14

2. State of the art 15

2.1. Overview . . . 15
2.2. Preprocessing . . . 16
2.3. Selection of regions of interest . . . 19
2.3.1. Stereo-based segmentation. . . . 19
2.3.2. Far Infrared Spectrum . . . 21
2.3.3. Integration of visible-infrared imagery . . . 22
2.3.4. Sliding Window approach . . . 22
2.4. Silhouette Matching . . . 23
2.5. Pedestrian Descriptors . . . 24
2.5.1. Holistic Methods . . . 24
2.5.2. Part-based descriptors . . . 28
2.5.3. Multi-feature Methods . . . 29
2.6. Classification . . . 30
2.7. Verification and Refinement . . . 31

2.8. Tracking 31
2.8.1. Kalman 32
2.8.2. Particle Filters 32
2.8.3. Other techniques 33
2.9. Other Important Issues 33
2.9.1. Sensors and Fusion 33
2.9.2. Applications 35
2.10. Other surveys in pedestrian detection and recognition 35

3. Classification **37**
3.1. Introduction 37
3.1.1. Chapter Structure 38
3.2. Classification Dataset 38
3.2.1. Pedestrian Datasets 38
3.3. Probabilistic models 42
3.4. Histograms of Oriented Phase Energy 46
3.4.1. Phase Congruency 47
3.4.2. Descriptor specifications 50
3.4.3. Evaluation of Descriptor Parameters 53
3.4.4. Evaluation of the Classifier Parameters 57
3.4.5. Other considerations 60
3.5. Integral Features 66
3.5.1. Square, non-overlapping features 68
3.5.2. Rectangular, overlapping features 70
3.6. Comparative Results 72
3.6.1. Classification Methods 73
3.6.2. Features 74
3.6.3. Discussion of the Comparative Results 77
3.6.4. Statistical Significance of the Results 77
3.7. Conclusions and Discussion 79

4. Detection **83**
4.1. Introduction . . . 83
4.1.1. Chapter Structure . . . 84
4.2. Detection Dataset . . . 84
4.3. Sliding Window Approach . . . 85
4.3.1. Evaluation methodology . . . 85
4.3.2. Results using the LSI dataset . . . 88
4.3.3. Results in the OSU Database . . . 90
4.3.4. Latent-SVM . . . 94
4.4. Small pedestrians . . . 97
4.5. Scale Approximation . . . 100
4.6. Improving the performance . . . 103
4.6.1. Selection of Regions of Interest . . . 103
4.6.2. Part-based detection . . . 108
4.7. Conclusions and Discussion . . . 112

5. Tracking **117**
5.1. Kalman Filter Variables . . . 119
5.1.1. Time Update . . . 119
5.1.2. Measurement Update . . . 122
5.2. Detection Matching . . . 123
5.3. Experimental Results . . . 123
5.3.1. Non-occluded Pedestrians . . . 124
5.3.2. Occluded Pedestrians . . . 125
5.3.3. Motion Model . . . 128
5.4. Conclusions . . . 131

6. Conclusions and Future Work **135**

A. Introduction to the Kalman Filter and its derivates **141**
A.1. The Kalman Filter . . . 141
A.1.1. Constraints . . . 141
A.1.2. Principles . . . 142
A.1.3. Equations. . . . 145
A.1.4. Kalman Filter Variants . . . 147

A.2. Unscented Kalman Filter 148
A.2.1. Prediction. 151
A.2.2. Measurement Update 152

B. Vision System. **155**
B.1. Calibration of the camera parameters. 155
B.1.1. Intrinsic parameters. 156
B.1.2. Extrinsic parameters. 158
B.2. Projective Geometry of the World into the Image. 158
B.3. Projection of the points of the image into the world 159
B.4. Calibration of the gain curve of a microbolometer. 160

References **163**

List of Figures

1.1. Time of day in traffic involving pedestrians. Data from [NHTSA,2010]). . . 5

1.2. Mortality rate relative to the speed of collision [Rosen et al., 2011.] 8

1.3. Experimental vehicles used in the work presented in this thesis. 11

1.4. Visual part of the detection system by [Musleh et al. 2010]. 12

1.5. Matching of key-points of the ground-plane as presented in [Musleh et al. 2012] 12

1.6. Traffic sign detection of [Carrasco et al. 2009]. Candidates are filtered by segmenting the image based on hue. A neural network is use to recognize the specific kind of traffic sign. 12

1.7. Detection of road marks using the Hough transform as presented in [Collado et al. 2009]; (a) View from the camera perspective; (b) Inverse perspective; (c) Lines as points after applying the Hough transform. 13

1.8. Somnolence monitoring as presented in [Flores et al. 2009]. Face and eyes are detected in both day and night configurations. The rate of blinking accumulated over time constitutes an indicator or somnolence on the driver. 13

2.1. In [Bertozzii et al. 2003a] a video stabilizer of FIR images is proposed. (a) Original FIR images, (b) Horizontal edges, (c) Histogram correlation of two images. 18

2.2. Tetracular system used in [Krotosky and Trivedi, 2007a] made up of two VL and two FIR cameras. 20

2.3. Example of stereo processing in [Krotosky and Trivedi, 2007a]. a) VL image with bounding boxes surrounding obstacles with proper dimensions. b) FIR image. c) Disparity map computed from the VL pair, with ground plane removed. d) Disparity map from the FIR pair. 21

2.4. Image registration of the system used in [St-Laurent et al., 2007] 22

2.5. Examples of simple Haar-like filters. (a) to (g): Edge features. (h) to (n): line features 25

2.6. Flow chart of the FIR pedestrian detection system proposed in [Sun at al., 2011]. They use Haar-like features in an AdaBoost learning framework. In order to reduce the number of ROIs, they first extract points of interest in the image and search for pedestrians only in the neighborhood of the detected keypoints. 25

2.7. HOG descriptor. (a): Average gradient of the INRIA Pedestrian Database training set. (b): Maximum positive SVM weight of each descriptor cell. (c): Maximum negative SVM weight of each descriptor cell. (d): Cropped image sample. (e): Illustration of its HOG descriptor. (f): Positive weight normalized descriptor. (g): Negative weight normalized descriptor. Source: [Dalal and Triggs, 2005] 27

2.8. Representation of a pedestrian in a FIR image using multiblock-LBP [Xia at al., 2011]. 28

2.9. Integral Channel Features as described in [Dollár et al., 2009a]: Gradient histograms, gradient magnitude and LUV channels. All these features can be computed using integral images, thus having an efficient computation. . . 30

2.10. Regions of interest generated using a laser rangefinder, presented in [Premebida and Nunes, 2013]. The experiments where conducted in the ISRobotCar, in Coimbra, Pt. 34

3.1. Example cropped-images of the classification dataset. The two upper rows contains examples of pedestrians acquired under different temperatures and illumination conditions. The lower rows contain randomly selected windows from images containing no pedestrians. For visualization purposes the contrast has been enhanced. 41

3.2. Histograms of bounding boxes sizes and areas for positive and negative samples of the train dataset. 42

3.3. Centres of bounding boxes for positives of the train and test dataset on a logarithmic scale. 43

3.4. Infrared images under different illumination and temperature conditions. . . 44

3.5. Gray Level of three constant temperatures of the human body plotted against the sensor temperature. 45

3.6. Average value of thresholded pedestrian samples. 45

3.7. Detection Error Trade-off curve of the correlation with the probabilistic models. 46

3.8. Examples of phase congruency and gradient of an infrared image. 48

3.9. Real and imaginary parts of a Gabor filter in one dimension. 49

3.10. Scale and contrast invariance properties of phase congruency. Figures (a) and (b) are two synthetic signals with equal shape but different scale. Their phase congruency amplitude (h) is exactly the same. Figures (g) and (h) are two similarly shaped signals with different scale. The phase congruency amplitude (l) is almost exactly the same for both. 51

3.11. Four different orientations (θ) of filters for the same frequency λ. The filter rotates to captures image variations in different directions between $\theta = 0$ and $\theta = 2\pi$. 52

3.12. (a) Original image. (b) Magnitude of phase congruency. (c) Gradient orientation. (d) Representation of the descriptors packed into grids. 52

3.13. DET curves of the HOPE descriptor created with different numbers of Gabor scales. Legend states Miss Rate (MR) at 10^{-4} FPPW. 54

3.14. DET curves of the HOPE descriptor created with different numbers of Gabor orientations. Legend states Miss Rate (MR) at 10^{-4} FPPW. 54

3.15. DET curves of cell sizes of the HOPE descriptor. Legend states Miss Rate (MR) at 10^{-4} FPPW. 55

3.16. DET curves of the HOPE descriptor with different number of histogram bins . Legend states Miss Rate (MR) at 10^{-4} FPPW. 55

3.17. Orientations of a 4×4 spatial cell. (a) Signed orientation, in the $[-\pi, \pi]$ range: $O_{360^\circ} = O$. (b) Unsigned orientation: $O_{180^\circ} = O^- + \pi$. (c) Orientation can also be wrapped in the $[0, \pi]$ range by $O_{180^\circ} = |O|$. Subfigures (d), (e), (f) represent their respective histograms by splitting the orientation range into 12 bins. 56

3.18. DET curves of different cell normalizations of the HOPE descriptor. Legend states Miss Rate (MR) at 10^{-4} FPPW. 57

3.19. DET curves of of the HOPE descriptor trained with a different number of samples. Legend states Miss Rate (MR) at 10^{-4} FPPW. 58

3.20. DET curves of of the HOPE descriptor trained with different SVM kernels. Legend states Miss Rate (MR) at 10^{-4} FPPW. 59

3.21. DET curve of classifiers after iteratively searching for hard negatives and retraining. 59

3.22. DET curve of the HOPE SVM-Rbf classifier trained with an increasing number of negatives. Legend states Miss Rate (MR) at 10^{-4} FPPW. 60

3.23. Curve fitting of the relation between number of negatives on the training set and Miss Rate at 10^{-4} FPPW in the test set. Notice that, from $n = 20000$ onwards, adding more samples has little impact on classification performance. 60

3.24. Descriptors at different scales around the same keypoint 61

3.25. Comparison of the results with the single scale and the multi resolution HOPE descriptor. Legend states miss rate at 10^{-4} FPPW 61

3.26. First Row: Positive sample with Gaussian noise added. Second Row: Phase Congruency Magnitude Response. 62

3.27. Classification DET curves for different amounts of syntethic Gaussian Noise. Classification of noisy samples achieves an acceptable hit rate for Gaussian noise with variance $\sigma \leq 10^{-5}$. 63

3.28. First Row: Noisy samples reconstructed with a Median filter. Second Row: Phase Congruency Magnitude Response. 63

3.29. First Row: Noisy samples reconstructed with a Wiener filter. Second Row: Phase Congruency Magnitude Response 64
3.30. Classification DET curve for the database denoised with a Wiener filter. . . 65
3.31. Classification DET curve for the database denoised with a Median filter. . . 65
3.32. Descriptor computed using the integral image. 67
3.33. HOPE features. Each channel is a bin of the histograms 67
3.34. Histograms of gradient magnitude. Each channel is a bin of the histograms. 67
3.35. Integral Channels (Random Forest Classifier). Legend states Miss Rate (MR) at 10^{-4} FPPW. 68
3.36. Integral Channels (Random Forest Classifier). Legend states Miss Rate (MR) at 10^{-4} FPPW. 69
3.37. Integral Channels (Adaboost). Legend states Miss Rate (MR) at 10^{-4} FPPW. 70
3.38. DET curves of the random rectangular features. For these experiments the number of histogram bins is set to 6. Legend represent Miss Rate (MR) at 10^{-4} FPPW. 71
3.39. Representation of feature importance in the random rectangular descriptor. Each pixels is scaled between 0 (least important) and 1 (most important). The subfigures represent the following feature vector: a) Gray-level; b) PC; c) Gradient; d)-i) HOPE; j)-o) Hist 72
3.40. Two layered neural network with ten hidden and one output sigmoid neurons. The inputs $f_1 \cdots f_n$ are the histogram bins. 74
3.41. DET curves of the HOG descriptor. a) Normalized histograms; b) Unnormalized histograms. Legend states Miss Rate (MR) at 10^{-4} FPPW. 75
3.42. DET curves of the HOPE descriptor. Legend states Miss Rate (MR) at 10^{-4} FPPW. 75
3.43. Example of an LBP descriptor 76
3.44. DET curves of the LBP descriptor. Legend states Miss Rate (MR) at 10^{-4} FPPW. 76
3.45. First 5 eigenpedestrians 77
3.46. DET curves of the PCA descriptor. Legend states Miss Rate (MR) at 10^{-4} FPPW. 78
3.47. DET curves of the combination HOG and HOPE descriptors. Legend states Miss Rate (MR) at 10^{-4} FPPW. 78
3.48. DET curves of the combination of LBP with HOG and HOPE descriptors. Legend states Miss Rate (MR) at 10^{-4} FPPW. 78
3.49. DET curves of the combination of PCA with HOPE descriptors. Legend states Miss Rate (MR) at 10^{-4} FPPW. 79

3.50. Misclassified samples. (a) False negatives due to low resolution, motion blur and pose variation. (b) ExaFalse positives of areas with a high vertical symmetry. 80

4.1. Flow chart of a pedestrian detector in images. 84
4.2. Subset of the Pedestrian Detection Dataset. 86
4.3. Histogram of mean gray level of the images in the Train and Test Databases. 87
4.4. Sliding Window approach. 87
4.5. Overlapping area of Ground Truth and Detection. 87
4.6. Example of non maximum suppression of multiple detections for each pedestrian. 88
4.7. DetectionScale. 89
4.8. Miss rate at 10^{-1} FPPI for the detector in the *Small*, *Medium* and *Large* and *Very Large* Test Subsets. 89
4.9. Pascal Overlap Criteria. 90
4.10. Detection DET Curves after applying the PM NMS algorithm with an overlap threshold of $a_o = 0.5$. Legend states Miss Rate (MR) at 0.1 FPPI. 90
4.11. Processed images from the OSU Thermal Pedestrian Database. 91
4.12. DET curves for the 10 sequences of the OSU database, using the HOPE detector trained on the LSI database. 92
4.13. DET curves for the 10 sequences of the OSU database, using the HOG detector trained on the LSI database. 92
4.14. DET curves for the 10 sequences of the OSU database, using the HOPE detector trained on the OSU database. 93
4.15. DET curves for the 10 sequences of the OSU database, using the HOG detector trained on the OSU database. 93
4.16. Latent SVM filters using HOG features. 95
4.17. Latent SVM filters using HOPE features. 95
4.18. Detection DET Curves of Latent-SVM HOG after applying the PM NMS algorithm. Legend states Miss Rate (MR) at 0.1 FPPI. 96
4.19. Detection DET Curves of Latent-SVM HOPE after applying the PM NMS algorithm. Legend states Miss Rate (MR) at 0.1 FPPI. 96
4.20. Example of small pedestrian 98
4.21. PSNR of a range of values of the minimum wavelength for different scales. . 99
4.22. Feature approximation. 99
4.23. Scale approximation 101
4.24. Scale approximation. 101

4.25. Reference system of world and camera coordinates. 104
4.26. Selection of regions of interest. 106
4.27. Search of the pedestrian inside the previously calculated ROI. 107
4.28. Selection of ROIS by edge density. 110
4.29. Descriptor blocks of pedestrian parts. For each block an SVM classifier is trained. 110
4.30. Relative classification weight of pedestrian parts scaled in $w_p = \{0, 1\}$. . . . 111
4.31. Occlussion 112
4.32. Occluded area of the region of interest. 112
4.33. Roc curves for occluded pedestrians using a full model and part based detection based on logic inference. Legend states occlusion and area under the curve (auc). 113
4.34. Roc curve of full body classification for different percentages of occlusion. Legend states occlusion and area under the curve (auc). 114
4.35. Roc curves of the best performing part classifiers. Only parts classifiers with an area under the curve $a_{uc} > thr$ are plotted. 114

5.1. Detections 118
5.2. Predictions 118
5.3. Representation of the movement of the vehicle between two consecutive frames. The perspective of the object, represented as a circle, changes due to a roto-translation of the camera. 120
5.4. Samples of tracking test sequence # 1 124
5.5. Pedestrian detections projected on the ground plane (sequence # 1). 124
5.6. Pedestrian detections projected on the ground plane (sequence # 1). 125
5.7. Samples of tracking test sequence # 2 126
5.8. Pedestrian detections projected on the ground plane (sequence # 2). 126
5.9. Pedestrian detections projected on the ground plane (sequence #2). 127
5.10. Samples of tracking test sequence # 3 127
5.11. Pedestrian detections projected on the ground plane (sequence # 3). 128
5.12. Precision-Recall curves of filtered and unfiltered tracks (sequence # 3). . . . 128
5.13. Samples of tracking test sequence # 4 129
5.14. Pedestrian detections projected on the ground plane (sequence # 4). 129
5.15. Precision-Recall curves of filtered and unfiltered tracks (sequence # 4). . . . 130
5.16. Samples of tracking test sequence # 5 130

5.17. Pedestrian detections projected on the ground plane (sequence # 5). a) Unfiltered tracks; b) Filtered tracks (Static Model); c) Filtered tracks (Dynamic Model). 132

5.18. Pedestrian detections projected on the ground plane (sequence # 5). 132

5.19. Samples of tracking test sequence # 6 133

5.20. Pedestrian detections projected on the ground plane (sequence # 6). a) Unfiltered tracks; b) Filtered tracks (Static Model); c) Filtered tracks (Dynamic Model). 133

5.21. Pedestrian detections projected on the ground plane (sequence # 6). 134

A.1. Gaussian distributions, representing the measurements of two different sensors of the same variable. 142

A.2. The combination of two Gaussian distribution is also Gaussian. 144

A.3. Recursive Kalman filter algorithm. 145

A.4. Kalman filter equations. 146

A.5. Behaviour of the Kalman filter for large values of R. 147

A.6. Behaviour of the Kalman filter for small values of R. 147

A.7. In a linear system, if the input is Gaussian noise, the output will as well be. 147

A.8. In a non linear system, for a Gaussian input, the output distribution is not a Gaussian. 149

A.9. If the systems is fairly linear the propagated veriable can be aproximated to a Gaussian. 149

A.10. Linearization around the working point. 149

A.11. Visual representation of UKF, EKF and sampling approaches (Eric A. Wan and Rudolph van der Merwe). 150

B.1. Pinhole projective model. 155

B.2. Aluminum chessboard pattern for FIR cameras calibration. 157

B.3. Coordinate reference system of world and camera. 158

B.4. Gray level of three constant temperatures of the human body, against temperature of the sensor. 161

List of Tables

1.1. Pedestrians fatalities, by age and location (Data from [NHTSA,2010]) . . . 3
1.2. Nonmotorists fatalities between 1994 and 2010 (Data from [NHTSA,2010]) . 3
1.3. Pedestrians fatalities, by related factors (Data from [NHTSA,2010]) 4
1.4. Pedestrians fatalities, by time of day and day of week (Data from [NHTSA,2010]) 5
1.5. Pedestrian fatalities at 30 days in EU Countries (Data from European Commision [Care, 2011]) 6
1.6. Infrared sub-divisions 10

3.1. Pedestrian databases. The first 13 databases are built with images in the visible light (VL) spectrum. Their information is extracted from [Dollár et al., 2012] . The LSI and OSU databases contain images in the FIR spectrum. 40
3.2. Results of the McNemar's approximate significance test for every pair of classifiers. The value expressed in the table's fields is χ^2, as stated in equation 3.23 79

Abstract

Detection of people in images is a relatively new field of research, but has been widely accepted. The applications are multiple, such as self-labeling of large databases, security systems and pedestrian detection in intelligent transportation systems. Within the latter, the purpose of a pedestrian detector from a moving vehicle is to detect the presence of people in the path of the vehicle. The ultimate goal is to avoid a collision between the two. This thesis is framed with the advanced driver assistance systems, passive safety systems that warn the driver of conditions that may be adverse.

An advanced driving assistance system module, aimed to warn the driver about the presence of pedestrians, using computer vision in thermal images, is presented in this thesis. Such sensors are particularly useful under conditions of low illumination.The document is divided following the usual parts of a pedestrian detection system: development of descriptors that define the appearance of people in these kind of images, the application of these descriptors to full-sized images and temporal tracking of pedestrians found. As part of the work developed in this thesis, database of pedestrians in the far infrared spectrum is presented. This database has been used in developing an evaluation of pedestrian detection systems as well as for the development of new descriptors. These descriptors use techniques for the systematic description of the shape of the pedestrian as well as methods to achieve invariance to contrast, illumination or ambient temperature. The descriptors are analyzed and modified to improve their performance in a detection problem, where potential candidates are searched for in full size images. Finally, a method for tracking the detected pedestrians is proposed to reduce the number of miss-detections that occurred at earlier stages of the algorithm.

Resumen

La detección de personas en imágenes es un campo de investigación relativamente nuevo, pero que ha tenido una amplia acogida. Las aplicaciones son múltiples, tales como auto-etiquetado de grandes bases de datos, sistemas de seguridad y detección de peatones en sistemas inteligentes de transporte. Dentro de este último, la detección de peatones desde un vehículo móvil tiene como objetivo detectar la presencia de personas en la trayectoria del vehículo. EL fin último es evitar una colisión entre ambos. Esta tesis se enmarca en los sistemas avanzados de ayuda a la conducción; sistemas de seguridad pasivos, que advierten al conductor de condiciones que pueden ser adversas.

En esta tesis se presenta un módulo de ayuda a la conducción destinado a advertir de la presencia de peatones, mediante el uso de visión por computador en imágenes térmicas. Este tipo de sensores resultan especialmente útiles en condiciones de baja iluminación. El documento se divide siguiendo las partes habituales de una sistema de detección de peatones: desarrollo de descriptores que defina la apariencia de las personas en este tipo de imágenes, la aplicación de estos en imágenes de tamaño completo y el seguimiento temporal de los peatones encontrados. Como parte del trabajo desarrollado en esta tesis se presenta una base de datos de peatones en el espectro infrarrojo lejano. Esta base de datos ha sido utilizada para desarrollar una evaluación de sistemas de detección de peatones, así como para el desarrollo de nuevos descriptores. Estos integran técnicas para la descripción sistemática de la forma del peatón, así como métodos para la invariancia al contraste, la iluminación o la temperatura externa. Los descriptores son analizados y modificados para mejorar su rendimiento en un problema de detección, donde se buscan posibles candidatos en una imagen de tamaño completo. Finalmente, se propone una método de seguimiento de los peatones detectados para reducir el número de fallos que se hayan producido etapas anteriores del algoritmo.

Acknoledgements

Quisiera agradecer a las siguientes personas el apoyo prestado durante el tiempo de realización de esta tesis. En primer lugar quiero darle el agradecimiento más importante a Verónica, por compartir su vida conmigo y llenarme de esperanza, incluso cuando la vida se da la vuelta. A mi madre, sé que estaría orgullosa. A mi familia, presente, futura y pasada le agradezco su paciencia, apoyo y todo el amor que me han dado.

Quiero agradecer a mis directores de tesis José María Armingol y Arturo de la Escalera, sus consejos y orientación durante la realización de este trabajo. Pero sobre todo, quiero agradecerles el aspecto humano de nuestra relación.

A Basam Musleh, que ha sido mi compañero en la realización de esta tesis y me ha apoyado en momentos muy duros. Te estoy muy agradecido por todo.

Finalmente quisiera agradecer a las personas que tan bien me acogieron en mi estancia en la Universidad de Coimbra: Urbano Nunes, Cristiano Premebida, João Luís Ruivo Carvalho y Álvaro Arranz.

Daniel Olmeda Reino
Madrid, Noviembre del 2013.

1

Introduction

The understanding of traffic is an important concern nowadays. There is wide agreement that current trends in the number of vehicles will make traffic unsustainable at some point in the future. Two are the main concerns with the current traffic systems: economic and safety-wise. Traffic accidents are one of the main causes of deaths and permanent physical disabilities in every country with an important presence of vehicles [192]. A reliable traffic infrastructure is also an important factor in economies. The constant growth of the number of vehicles is pushing the current roads to their flow limit. From both points of view, the traffic architecture has to be improved. The scientific community is also participating from this interest, with very interesting ideas as where the future of traffic will be. A relatively new knowledge area, and an actively developed one, is the study of Intelligent Transportation Systems (ITS). These systems focus both in traffic reliability and safety. The solution proposed for both is to take over responsibilities from the human driver and relocate them to an automatic system. These systems can integrate the information of every vehicle on the road and synchronize their movements, obtaining a much more fluid traffic. And because these systems can have a much wider sensorial information than a single person can, the risk of an accident can be decreased.

Traffic safety is a factor whose importance has been increasing in recent years. The infrastructure has improved, drivers are now subject to continuous awareness campaigns and vehicles incorporate more safety measures. The result is a reduction in road accidents, even if the number of vehicles has been in a rising trend for some time. Vehicles are more secure and today the chances suffering an accident are smaller than once were, and even if the accident occurs, the statistics reflects that the damage suffered by the passengers are not as severe. This is due to the fact that these safety measures are aimed at protecting the passenger compartment.

Driving takes place in an unpredictable environment. Therefore, designing safety systems is easier if there is an understanding of what it is to be protected. The presence of obstacles, their shape, and trajectory is information that the driver is expected to acquire and use to prevent and accident. However, if there is a collision between the vehicle and an obstacle, there exists a huge number of configurations in which this can occur, and the safety of the obstacle is not guaranteed. By contrast, the inside of vehicle is a much more controlled environment in which the position where the occupants can be is known. It is then possible to study the damage that these people may suffer in an accident and try to mitigate it, for example, placing airbags specifically where most important impacts happen. As for outside-the-vehicle safety, the most fragile and least protected element are pedestrians.

Unlike occupants of the passenger compartment of the vehicle, which have its structure to partially absorb the energy of the impact, pedestrians do not have built-in safety systems. So the odds of being killed in a traffic accident are much higher for pedestrians.

Safety outside of the vehicle has not been developed as much. Vehicles move in environments that are unknown to the designers of safety systems. It is not possible to anticipate the driving conditions, nor it is possible to predict the presence of obstacles in the road, the curvature of the path, or surrounding traffic conditions at all times. Therefore, the new safety systems must incorporate environmental perception. This capability will allow a quicker reactions to unexpected events.

This thesis aims to provide a system for analysis of the driving environment capable of detecting the presence of pedestrians in low visibility conditions. An artificial perception system identifies pedestrians in front of the vehicle and determines whether there is any risk that endangers the integrity of pedestrians as well as passengers.

1.1. Motivation

Every year 400 000 pedestrians are killed worldwide [155], more than 6000 in the European Union only [35]. Reducing the reducing the number of mortal traffic accidents is a challenging task, which will require the integration of several technologies, yet to be fully developed. In this section, the most relevant circumstances under which accidents involving pedestrians happen are reviewed. As an advance of the conclusions drawn from the following argumentation it should be noted that pedestrian accidents usually involve healthy, capable people in low illumination conditions.

The emphasis on road safety is a tendency which is growing. Consequences of traffic accidents are, for example, the death of a pedestrian in traffic accidents per minute. They also cause serious injuries to 10 million people a year. Poor traffic management also causes significant economic losses. As an average, 10% of roads in Europe are affected by traffic jams, causing losses of 50 billion in logistics each year, a 0.5% of European GDP [192].

The NHTSA or National Highway Traffic Safety Administration is the agency responsible for the implementation of measures to improve traffic safety in the U.S.A. and, as such, publish each year a collection of statistics related to traffic accidents [158]. This section summarizes some of the conclusions reached on the basis of their data, as a generalization of the traffic conditions in an important part of the world. However, the mere fact that the U.S. collects so accurate statistics says that they have an advanced traffic infrastructure and, as such, this data can not be extended to the entire population of the planet. Countries with fewer resources have less traffic flow, but also have a worse infrastructure and an older fleet. The number of accidents per vehicle is higher, but exact figures are not known.

The types of traffic accidents involving pedestrian, ranked by age pedestrian and location are shown in table 1.1, distinguishing between those that have happened in intersections and elsewhere. For each category, the total number of casualties in listed in the first column, and the percentage in the second. It should be noted that approximately 75% of deadly accidents have occurred outside the areas designed for pedestrians crossing.

Age	Location						Total		Chart
	Intersection		Nonintersection		Unknown				
	#	%	#	%	#	%	#	%	
< 5	12	12.8	71	75.5	11	11.7	94	2.2	
$[5-9]$	8	11.8	56	82.4	4	5.9	68	1.58	
$[10-15]$	32	24.4	82	62.6	17	13.0	131	3.06	
$[16-20]$	42	14.9	194	68.8	46	16.3	282	6.58	
$[21-24]$	31	11.2	212	76.5	34	12.3	277	6.17	
$[25-34]$	77	12.9	453	75.6	69	11.5	599	14.0	
$[35-44]$	93	16.2	415	72.4	65	11.3	573	13.38	
$[45-54]$	149	18.7	563	70.6	86	10.8	798	18.64	
$[55-64]$	155	25.2	399	64.9	61	9.9	615	13.37	
$[65-74]$	120	33.2	202	56.0	39	10.8	361	8.43	
$[> 74]$	178	38.3	255	54.8	32	6.9	465	10.86	
Unknown	3	17.6	14	82.4	0	0.0	17	0.4	
Total	900	21.0	2916	68.1	464	10.8	4280	100.0	

Table 1.1: Pedestrians fatalities, by age and location (Data from [NHTSA,2010])

The introduction of safety measures in new vehicles, among other factors, is slowly reducing the number of pedestrians injured or killed each year in traffic accidents. As seen in table 1.2 there is a decline in the number of pedestrians fatalities each year in traffic accidents. This decrease is greater that it may seem, since the number of vehicles has had an historical upward trend. The cause of this reduction is, on the one hand, the improvements in population awareness of traffic safety measurements, and, on the other on the improvements in road infrastructure and safety systems of vehicles.

Year	Pedestrian	Pedalcyclist	Other	Total	Chart
1994	5489	802	107	6398	
1995	5584	833	109	6526	
1996	5449	765	154	6368	
1997	5321	814	153	6288	
1998	5228	760	131	6119	
1999	4939	754	149	5842	
2000	4763	693	141	5597	
2001	4901	732	123	5756	
2002	4851	665	114	5630	
2003	4774	629	140	5543	
2004	4675	727	130	5532	
2005	4892	786	186	5864	
2006	4795	772	185	5752	
2007	4699	701	158	5558	
2008	4414	718	188	5320	
2009	4109	628	151	4888	
2010	4280	618	182	5080	

Table 1.2: Nonmotorists fatalities between 1994 and 2010 (Data from [NHTSA,2010])

Pedestrians are vulnerable in road environments. The major causes of run over is a bad use of the road on the part of the pedestrian. Among other causes, accidents happen when the pedestrian is:

- Standing, lying, working, playing in the roadway.
- Under the influence of drugs.
- Crossing the road where it is not allowed.
- Darting or running into road
- Not visible (dark clothing, no lighting, etc.)
- Ignoring traffic signs, signals, or officer.

The main factors leading to an accident are behavioral. The statistics of the occurrence of those actions, where an accident is involved are listed in table 1.3.

Factors	Number	Percent	Chart
Failure to yield right of way	976	22.8	
In roadway improperly (standing, lying, working, playing)	793	18.5	
Under the influence of alcohol, drugs, or medication	788	18.4	
Darting or running into road	727	17.0	
Not visible (dark clothing, no lighting, etc.	585	13.7	
Improper crossing of roadway or intersection	557	13.0	
Failure to obey traffic signs, signals, or officer	141	3.3	
Physical impairment	99	2.3	
Inattentive (talking, eating, etc.)	89	2.1	
Entering/exiting parked/standing vehicle	49	1.1	
Wrong-way walking	47	1.1	
Emotional (e.g. depression, angry, disturbed)	47	1.1	
Traveling on Prohibited Trafficways	37	0.9	
Ill, blackout	17	0.4	
Non-Motorist pushing vehicle	10	0.2	
Asleep or fatigued	8	0.2	
Vision obscured (by rain, snow, parked vehicle, sign, etc.)	8	0.2	
Portable Electronic Devices	6	0.1	
Other factors	171	4.0	
None Reported	1,139	26.6	
Unknown	34	0.8	
Total	4,280	100.0	

Table 1.3: Pedestrians fatalities, by related factors (Data from [NHTSA,2010])

If we analyze the data in table 1.1, distribution of mortality by age, we see that most accidents involve pedestrian of ages between 25 and 65. Pedestrians in this age range largely retain their mental and physical abilities intact. The main cause of accidents is due to the fact that pedestrians cross the road at unauthorized places. Drivers do not expect the presence of pedestrians and takes them longer to react, or fail to perceive the danger until the accident has occurred.

Lighting conditions have a major influence on the number of traffic accidents. This figure is particularly significant in the case of accidents involving pedestrians. With less light it takes longer for a driver to perceive a pedestrian on the road. Another important factor is exhaustion on the part of the driver. In this case the reaction time to a stimulus is

much higher and chances of an causing an accident grow. Figure 1.1 divides fatal pedestrian violations depending on the light conditions: day, night, and dawn or dusk. This type of accident is more common in conditions of limited visibility, even though there are fewer pedestrians and vehicles than during the day.

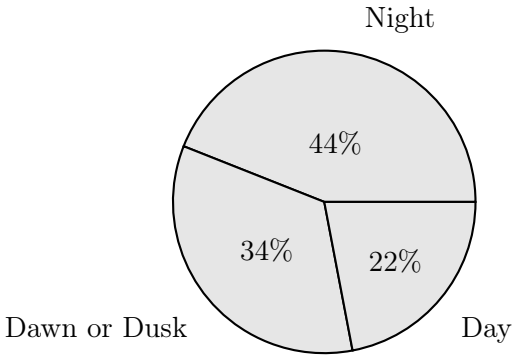


Figure 1.1: Time of day in traffic involving pedestrians. Data from [NHTSA,2010]).

Similarly, table 1.4 shows the number of fatal accidents categorized by time of day. In this case it is obvious that the number of accidents increases significantly at night, and has its minimum in the central hours of the day.

Time of Day	Day of Week						Total		Chart
	Weekday		Weekend		Unknown				
Midnight to 2:59 a.m.	204	8.2	352	19.6	0	0.0	556	13.0	
3 a.m. to 5:59 a.m.	189	7.6	237	13.2	0	0.0	426	10.0	
6 a.m. to 8:59 a.m.	326	13.2	73	4.1	0	0.0	399	9.3	
9 a.m. to 11:59 a.m.	193	7.8	56	3.1	0	0.0	249	5.8	
Noon to 2:59 p.m.	196	7.9	58	3.2	0	0.0	254	5.9	
3 p.m. to 5:59 p.m.	341	13.8	109	6.1	0	0.0	450	10.5	
6 p.m. to 8:59 p.m.	600	24.3	439	24.4	0	0.0	1,039	24.3	
9 p.m. to 11:59 p.m.	413	16.7	471	26.2	0	0.0	884	20.7	
Unknown	12	0.5	5	0.3	6	100.0	23	0.5	
Total	2,474	100.0	1,800	100.0	6	100.0	4,280	100.0	

Table 1.4: Pedestrians fatalities, by time of day and day of week (Data from [NHTSA,2010])

Data from the European Commission shows that pedestrians are specially vulnerable in urban environments, with very few accidents ocurring in highways or in rural environments. Table 1.5 shows the number of pedestrian casualties in European Countries.

From these statistics it is clear that pedestrians are especially vulnerable in urban environments, especially at night. Poor lighting and exhaustion are the main causes of traffic accidents involving pedestrians. Safety measures to prevent this kind of accidents should focus on early detection of dangerous elements on the road, to reduce braking distance. These safety measures are included in the Driver Assistance Systems, which are introduced in section 1.2.

Country	Year	Motorway	Rural	Urban	Total	Chart
Belgique/België	2011	4	34	73	111	
Bulgaria	2009	0	135	63	198	
Ceská Republika	2011	5	56	115	176	
Danmark	2010	1	14	29	44	
Deutschland	2011	32	154	428	614	
Eesti	2009	0	12	11	23	
Éire/Ireland	2010	4	21	19	44	
Elláda	2011	8	40	175	223	
España	2010	52	142	278	471	
France	2011	26	143	350	519	
Hrvatska	2011	0	4	67	71	
Italia	2010	18	112	484	614	
Kýpros - Kibris	2004	0	0	18	18	
Latvija	2011	0	34	26	60	
Luxembourg	2011	2	1	3	6	
Magyarország	2010	13	59	120	192	
Malta	2010	0	0	2	2	
Nederland	2019	5	12	46	63	
Österreich	2011	5	23	59	87	
Polska	2011	8	483	917	1408	
Portugal	2011	5	28	166	199	
România	2011	4	150	593	747	
Slovenija	2010	2	6	18	26	
Slovensko	2010	2	38	86	126	
Suomi/Finland	2011	1	12	28	41	
Sverige	2009	3	16	25	44	
United Kingdom(GB only)	2011	24	79	302	405	
Total	2011	224	1808	4501	6532	

Table 1.5: Pedestrian fatalities at 30 days in EU Countries (Data from European Commision [Care, 2011])

1.2. Advanced Driver Assistance Systems

Advanced Driver Assistance Systems (ADASs) are active safety measures onboard vehicles with a human driver. Such systems are known as active because, although at no time taking control of the vehicle, their function is to prevent the accident. To do this, these systems gather information from the environment, evaluating the possibility of occurrence of hazardous events. In contrast, a passive safety system is one that tries to minimize the damage while the accident is happening. The ADAS seek, among other information, signs of drowsiness or inattention on the driver, obstacles on the way, and monitor the correct position of the vehicle on the road. Drivers are responsible for controlling the vehicle, but can receive this type of information to complete their cognitive limitations.

The industry has adopted some of these technologies, and is integrating them in commercial products. Probably, the most commonly available are satellite navigation systems, which provide autonomous geo-spatial positioning, and usually also traffic information. Another technology which is now widely available are the Autonomous Cruise Control Systems. These systems automatically adjusts the speed of the vehicle in order to maintain a minimum safety distance with the vehicle ahead. In 1995 the Mitsubishi Diamante was the first vehicle to offer such a capability, enabled by a laser rangefinder. Lane departure warning systems have also a strong presence in commercial products. Dating to in 2000, the introduction of the Iteris lane departure system, integrated in the Mercedes Actros trucks, has led to this kind of ADAS to be commonly available nowadays. From 2008 manufacturers such as BMW, Opel and Mercedes-Benz began offering Traffic Sign Recognition (TSR) systems. To this date, several others have introduced these kind of ADAS, as Volkswagen, Saab or Volvo. Driver Monitoring Systems were first introduced in the market by Lexus in 2006. The systems monitors the driver's engagement in the task of driving by tracking the the eyes. Obstacle avoidance is introduced as part of an active safety system, and usually rely on laser or radar information. One of the first examples of these kind of systems was introduced in the Mercedes S-Class in 2006. The system relies on radar information to detect obstacles in the path of the vehicle. In case of imminent collision a partial automatic breaking system is activated. In what is called by the industry *Night Vision*, several car manufacturers are offering thermal imaging in their high end models. In 2000 the Cadillac Deville was the first vehicle to be sold with this system. In 2004 Honda introduced a Legend model equipped with a pedestrian detector based on temperature segmentation from a thermal camera. The system require an ambient temperature below 30 degrees celsius in order to properly function. More recently, Audi introduced an A8 model equipped with a similar system. Automatic driving is also growing. An specially relevant case is the issue of the first license for a self-driven car for in May 2012 to Google Inc. by the Nevada Department of Motor Vehicles.

Active safety systems, in which the vehicle seizes control by a brief moment, also benefits from the technology developed for ADAS. Specifically, pedestrian detection can be integrated into systems known as *Pre-Crash*. In this case, the system takes control of the vehicle, but only for a very limited time. They only begin to function when normal reaction time on the part of the driver has been exceeded. In the case of prevention of run over, these systems begin to operation when a collision is imminent or is already happening.

These systems usually are unable to avoid the accident, but make its consequences less severe.

Figure 1.2 shows the mortality rate relative to the speed of the collision based on the data collected in [2], [170] and [177]. Generally, it can be seen that mortal collisions are reduced sharply if the vehicle is moving with a speed of less than 40 km/h. A good *Pre-Crash* system would slow down the vehicle to this speed just before impact. There is a remarkable difference between a colision at 40 and 50 km/h as the probability of fatality goes from 30% to 85% in the worst case scenario.

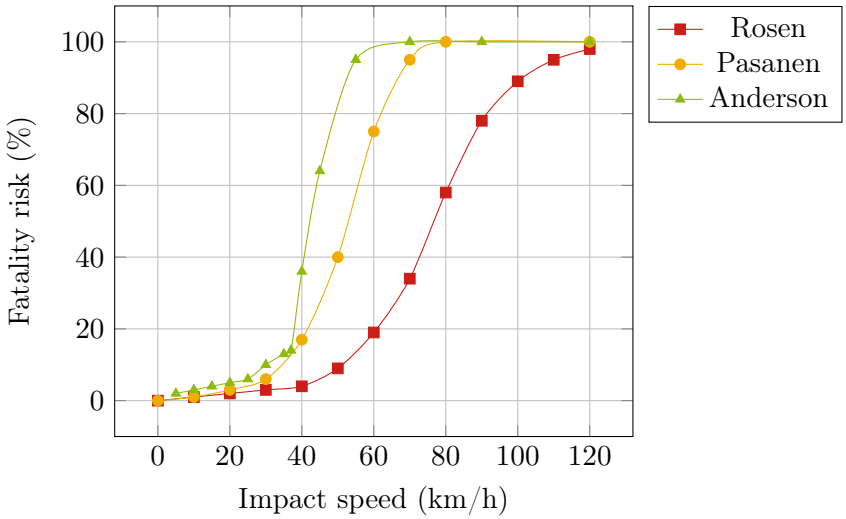


Figure 1.2: Mortality rate relative to the speed of collision [Rosen et al., 2011.]

1.3. People safety in Intelligent Transportation Systems

Object recognition in images has become a very important topic in the fields of traffic infrastructure and driving assistance system. Applications such as traffic signs recognition [45] and obstacle avoidance [63] have gotten the attention of the industry for some time now. The case of people detection is an exceptionally relevant case, as it leads to a number of important applications, some of which strive for saving lives. Pedestrian recognition in images is geared toward a variety of applications, which include safety focused road infrastructures [201], driver assistance systems [91] and autonomous robotic vehicles [173]. It is also useful in security, be it for automatic surveillance [111] or people counting [233] [130], including automatic recognition of people in low light conditions in unmanned aerial vehicles aimed to rescue missions [147].

Pedestrian recognition in Intelligent Transportation systems is usually geared to improving pedestrian safety. Pedestrian protection systems can be classified in three categories: infrastructure enhancements, passive and active detection. In the first one, infrastructure design, pedestrian detection is not mandatory, but rather infrastructures should be designed in way that minimizes the accidents by, for instance, limiting parking removing on-street parking in residential areas [175]. Other infrastructure enhancements, that rely

on pedestrian detection are, for instance flashing light warnings on pedestrian crosswalks. Active and passive pedestrian protection require having vehicle-mounted sensors. Another possibility is infrastructure-mounted sensors and a vehicle to infrastructure communication. For reference, in [83] an extensive review of pedestrian protection systems is presented.

Safely driving is a challenging task for an human driver. The environment is not fully controlled, so there is always an unknown probability of encountering an unexpected obstacle. Pedestrians are a special case among obstacles the driver might encounter. In urban scenarios, vehicles and pedestrians share the same ground so there is higher probability of a collision than in highway traffic. It is a particularly dangerous situation because pedestrians are much more likely to be hurt than the occupants of the vehicle, even at low speeds. ADAS provide drivers with additional information relevant to the driving task. These systems usually exploit on-board sensors that broaden what the driver is able to perceive. There are a number of reasons why these sensors exceed the driver capacity. The point of view of the driver, inside of the vehicle, may be incomplete due to occlusions of the driveway, or because several attention points may be needed at the same time at different locations. Another benefit of the use of ADAS is that the sensors can be designed to acquire information not available to the driver senses.

However, detection of pedestrians from a moving vehicle is not trivial, as they can appear with fairly different shapes, and in a random fashion. The use of computer vision to solve this situations is justified as other approaches, such as lidar scanners, although delivering very precise measurements of distance, doesn't provide enough information to discriminate between different types of obstacles. On top of that, vision is a non intrusive method. On the downside, the performance of a computer vision application is very dependent on the illumination conditions. There is a rich bibliography about pedestrian detection using cameras in the visible range light. As for night driving, there are two possibilities: to illuminate the scene with infrared leds and capture it with near infrared cameras [129], or the use of thermal cameras that captures the emission of objects in the far infrared spectrum.

Far infrared images have a very valuable advantage over the visible light ones. They do not depend on the illumination of the scene. The output of those cameras is a projection on the sensor plane of the emissions of heat of the objects, that is proportional to the temperature. Tracking can greatly simplify the task of pedestrian detection and cope with temporal occlusions or mis-detections. It can also be used to predict trajectory and time left for collision between pedestrian and vehicle. Yet, this step is usually neglected in papers describing far infrared pedestrian detector.

The use of far infrared cameras, besides all its advantages, is usually unable to cope with every scenario. Infrared cameras are unable to replace visible light cameras, as they present some disadvantages. As the outside temperature raises to high levels, the sensor's noise render the images useless for extracting distinctive features for pedestrian detection. Besides, direct sunlight, no matter what temperature, affects infrared images, as reflection on some surfaces make them appear hotter than they really are. The tendency is to integrate infrared vision cameras with other sensors (e.g. radar, visible light images) in a system that decides which information would be more useful under different circumstances.

1.4. Thermal Imaging

Pedestrian detection in low illumination conditions requires a sensor which is able to acquire information in the absence of illumination, but also are able to capture the shape of a person. Thermal imaging offers both advantages.

Any object with a temperature above the absolute zero emits radiation in the infrared range, as defined in the Plank's wavelength distribution function. The wavelengths of these radiations go from $3\mu m$ to $14\mu m$, and are usually refer to as Thermal Infrared (TIR) range. That range matches the emissions of object between $190K$ and $1000K$ [82]. The infrared range goes from $0.7\mu m$ to $1000\mu m$ and encompasses several other regions or subdivisions. The denominations of the infrared ranges varies depending on the field of study and the authors. Table 1.6 shows an infrared range division as defined in [214]. In it, the infrared spectra is divided based on the sensibility of common detectors.

Table 1.6: Infrared sub-divisions

Name	Abbreviation	λ (μm)	Sensor
Near	NIR	0.7 - 1.0	From human eye to Si
Short Wave	SWIR	1.0 - 3	InGaAs
Mid Wave	MWIR	3 - 5	InSb
Long Wave	LWIR	7 - 14	Microbolometers
Very Long Wave	VLWIR	12 - 30	Doped Silicon

In the field of ITS the term *Far Infrared* (FIR) is used indistinctly to refer to the *Long Wave Infrared* (LWIR) [198] [23] [231] [163] [148] and others. The same range is denoted as *Thermal Infrared* (TIR) in [91] and as *Long Wave Infrared* (LWIR) in [189]. In this work, the term FIR is used when referring to the sensitivity range of a microbolometer.

Generally, there are two kinds of thermal cameras [176]:

- Photon detectors are based on the *photoeffect*, which states that the the absorption of photons in a material results in the transition of electrons to a higher energy level and thus the generation of charge carriers. In the presence of an electric field these carriers move, producing and electric current. This current is proportional to the radiation absorbed by the material. These sensors are sensitivity to small variations of scene temperature, given that their own temperature is kept low. Photon detectors have to be refrigerated, if thermal noise is to be avoided.
- Thermal detectors measure a physical property of the sensor's material, related to its temperature. This property is electrical resistance, in the case of the microbolometers, or electrical polarization, in the case of ferroelectric detectors. The latter captures the changes of temperatures in the scene by measuring the ferroelectric phase transition in dielectric materials.

Microbolometers have several advantages over other thermal cameras. First of all, they do not usually require refrigeration, thus reducing their volume, price and maintenance

requirements. Moreover, their sensitivity is higher than the one a ferroelectric detector and have largest pixel densities. The camera used in this work is a microbolometer with a range sensitivity that goes from the $7\mu m$ to the $14\mu m$.

1.5. IVVI Research Platform

The Intelligent Vehicle based on Visual Information (IVVI) is a research vehicle for the development of advanced driver assistance systems based on computer vision. This platform enables the testing of the developed algorithms in real driving environments, both in urban and highway driving. The sensory system of the IVVI vehicle consists of a number of cameras compatible with the 1394 standard for road signs detection and monitoring of driver drowsiness, a sick® lidar for segmenting obstacles based on discrete distances. A stereo-based vision system is used for three-dimensional modeling of the driving environment (mainly to infer the position of the vehicle relative to the road) and to calculate the vehicle odometry using computer vision techniques. Detecting pedestrians in adverse lighting conditions is performed by using thermal information from an Indigo Omega® camera.

The developed pedestrian detection system is currently part of the IVVI$^{2.0}$ research platform (Fig. 1.3). Other assistance systems being developed on it are:

(a) IVVI (b) IVVI$^{2.0}$

Figure 1.3: Experimental vehicles used in the work presented in this thesis.

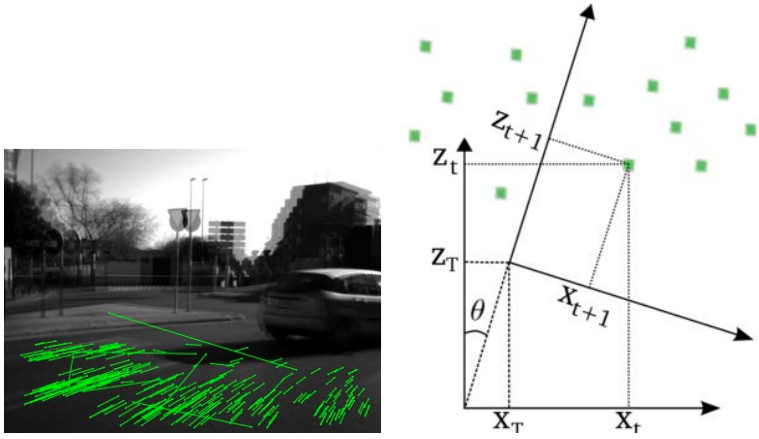
1. *Anti-Collision*: Detects and informs the driver of obstacles in the trajectory of the vehicle [153]. Figure 1.4 shows the obstacles segmented by the stereo-vision system. These results are merged with low-level distance information from a lidar scanner.

2. *Visual Odometry*: Infers vehicle movement by cross-correlation of key-points belonging to the ground plane [154]. Figure 1.5 shows the matching of ground keypoint found in two consecutive images.

3. *Speed Supervisor*: Detection and recognition of traffic signs. The system looks for speed traffic signs in its environment and alerts the driver if the speed of the vehicle is over the limit [37] [36]. The visualization of the traffic sign detection system is found in Fig. 1.6.

(a) (b) (c)

Figure 1.4: Visual part of the detection system by [Musleh et al. 2010].



(a) (b)

Figure 1.5: Matching of key-points of the ground-plane as presented in [Musleh et al. 2012]

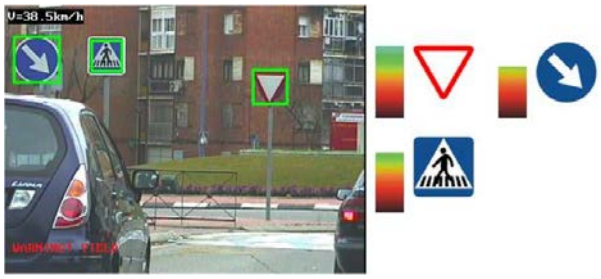
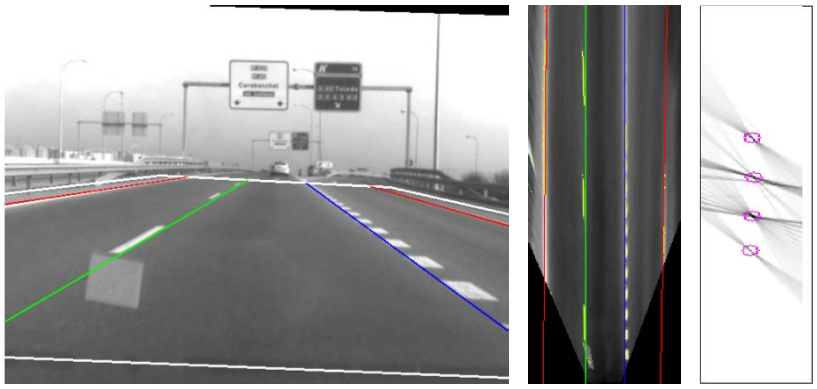


Figure 1.6: Traffic sign detection of [Carrasco et al. 2009]. Candidates are filtered by segmenting the image based on hue. A neural network is use to recognize the specific kind of traffic sign.

4. *Lane departure warning system*: Detection and classification of road markings. The driver is warned if the vehicle is about to cross a lane delimiter (Fig. 1.7). This systems also monitors the blind spot, looking for overtaking vehicles.[43]

5. *Drowssiness detection*: The system monitors driver behavior, looking for signs of fatigue or inattention. [78] [77]. Figure 1.8 shows a visualization of the night-time somnolence monitor algorithm.



(a) (b) (c)

Figure 1.7: Detection of road marks using the Hough transform as presented in [Collado et al. 2009]; (a) View from the camera perspective; (b) Inverse perspective; (c) Lines as points after applying the Hough transform.

Figure 1.8: Somnolence monitoring as presented in [Flores et al. 2009]. Face and eyes are detected in both day and night configurations. The rate of blinking accumulated over time constitutes an indicator or somnolence on the driver.

1.6. Objectives

This thesis is framed within the ADAS context. The developed system provides information to the driver about the presence of pedestrians on the road, analyzing information gathered with a camera sensitive to the far infrared (FIR) or thermal infrared (TIR) spectrum. Such information, is beyond what a human driver is able to perceive. The presented system, based on this kind of information, is able to detect pedestrians in conditions on which it would not be possible for the human vision. At night, a human driver has less visibility of the road ahead so, by the time the pedestrian is visible, reaction time has to be much quicker. The detection system proposed in this thesis has features that exceed the capabilities of the human vision. The use of a FIR camera allows to search for the heat that the pedestrians emit in conditions that, otherwise, would be unfit for exploiting a system based on visible light cameras, such as (and specially) night driving. As such, it does not require any kind of active illumination and can perceive objects at greater distances than it is possible with headlights. The system is also useful in adverse visibility conditions such

as rain or fog. The driving task is still a responsibility of the human driver. The ADAS monitor the exterior of the vehicle looking for situations that involve risk for people. These systems feed the driver with pertinent information, allowing for quicker reaction times.

The objectives of this thesis are the following:

- Understanding of the benefits and limitations of a FIR camera based on a low-resolution non-refrigerated microbolometer.
- Develop methods for selection regions of interest in the FIR images.
- Develop a descriptor that can benefit from the exclusive characteristics of FIR imagery.
- Asses the benefits of a tracking step to the overall detection performance.
- Establish a database of pedestrians in FIR images that can be used as a benchmark for detection algorithm in ITS applications.

1.7. Outline of the dissertation

This thesis is structured as follows. Chapter 2 reviews the relevant state of the art in pedestrian detection. The content of this chapter focus on the different steps of a pedestrian detection algorithm, i.e. preprocessing, selection of regions of interest, pedestrian descriptors, classification methods and tracking algorithms. The most relevant methods used in VL images are reviewed, along with methods applied exclusively to FIR imagery. In chapter 3 pedestrian recognition is treated as a classification problem. In it the characteristics of a FIR image-based pedestrian dataset are discussed, including the methodology of acquisition and sample selection. This chapter also focus on the features and methods used for classification performance assessment. Two new descriptors for pedestrian recognition in FIR images is also presented in this chapter. Chapter 4 focus on the problem of finding pedestrians in full-sized images. In it, a through evaluation of classification methods applied in an sliding window approach is presented. Benefits and drawbacks of this approach are commented, based on the results of the evaluation. An approach to handle small pedestrians and a scale approximation of features are presented. This chapter also covers initial research on two topics of pedestrian detection: selection of ROIs and occlusion handling. Chapter 5 focus on algorithms for pedestrian tracking. The benefits of using a tracking step in the pedestrian detection algorithm are further discussed. Conclusions and future work are presented in chapter 6.

2

State of the art

2.1. Overview

The main objective of a pedestrian detection system in ADAS is to avoid a traffic accident and eventually, in case the accident is inevitable, to reduce the possible damages to the passengers and pedestrians. Therefore, they require three phases: acquiring information, processing it, and communicating relevant information to the driver. The system does not take control of the vehicle, except in exceptional circumstances and for very small time intervals. Driving continues to be a human responsibility. Therefore, ADAS can be considered as an artificial copilot.

First, these systems need to acquire environment information. The different classes of objects that the system is aiming to detect define the approach to follow to solve the problem. Computer vision is usually the preferred technique, as these sensors are able to provide much more information than others, such as laser or radar scanners. These approaches may be extended by incorporating a set of priors provided by intelligent road networks, based on communication between infrastructure and vehicles. This methodology is intended to implement the foundations for a fully automatic driving. For an overview of important topics in automatic driving applications refer to [8].

Driving assistance systems based on visual information are being well received for several reasons. First, there is great potential information contained in images. From the analysis of an image sequence the global state of surrounding traffic in complex situations can be extracted. It also allows for obstacle detection by three-dimensional analysis of the scene. Also, these obstacles can be classified, differentiating between cars [192], bikers and pedestrians. Moreover, economic investment is much lower than other methods, such as radar. The evolution of the cameras and processors, allow analysis of images with increased resolution at lower prices.

Pedestrian detection is an extremely active research topic and new advanced techniques are being presented every so often. Much of the research presented recently rely on computer vision [83]. Visual information is rich in detail and allows not only to detect generic obstacles but to recognize the object type. In order to efficiently recognize objects in challenging scenarios, algorithms based on computer vision are evolving into more complex computational models and usually are divided into several stages. This section will summarize what are the most common and effective techniques. The methodology followed in the design of a pedestrian detection system can be divided into six steps:

- Pre-processing: depending on the kind of sensor used, some systems require an optional step of data pre-processing.
- Search of regions of interest in the image: at this stage, the system looks for areas in the image with high probability of containing a pedestrian. The aim is to reduce the computational complexity in the later stages of the algorithm. This step is optional and not all methods generate regions of interest, instead performing dense searches on the entire image. Some methods apply a pre-classification within this step. This approach removes from future steps of the algorithm regions unlikely to hold any relevant information. Pre-classification methods are usually simple and fast, for instance, filtering by symmetry or spatial location.
- Description of regions: patches of images need to be encoded into a descriptor that captures the information necessary to tell apart a pedestrian from any other object. At the same time this information must allow generalization, as each pedestrian present slight unique variations.
- Classification: from the information contained in the extracted regions, a pattern recognition algorithm makes the decision about whether it contains a pedestrian or not.
- Refinement: In this optional step, the resulting detections are re-evaluated, and false positives discarded.
- Tracking: from a set of images captured in sequence, the pedestrians future trajectory can be anticipated. Tracking can also be used to improve the detection performance by filtering isolated false detections and inferring the presence of a pedestrian when the detector fails or because the pedestrian is momentarily occluded.

In this chapter, a review of pedestrian detection in images is presented. In it, the most representative methods using Visible Light (VL) imaging are enumerated, while at the same time, focusing in relevant techniques applied to Far Infrared (FIR) images.

2.2. Preprocessing

Preprocessing an image before applying a pedestrian detection algorithm may ease subsequent steps of the algorithm. This step is optional but, when applied, is usually within the following topics.

Histogram equalization This step attempts to enhance the information contained within the dynamic range of the camera. Histogram equalization is applied to every computer vision algorithm, at least at a very low level and within the camera hardware by applying a pixel intensity transformation that approximates the sensibility curve of the sensor to something more appealing to the human eye.

FIR pedestrian detection algorithms that rely on segmentation of hotspots usually apply an intensity transformation to the pixel information. This is due to the rapid changes in dynamic range between images, so common in this kind of equipment. In [225], the authors apply uniform distribution equalization followed by a clipping of darkest and brightest pixels. Hotspots are later segmented from this pre-enhanced images. In [162] the authors address the problem of dynamic range variation due to shifts in sensor temperature in non-refrigerated microbolometers. A thermal calibration of the sensor simplifies that latter step of hot-spot segmentation.

Camera calibration and pose estimation Calibration of intrinsic and extrinsic parameters is an essential step in any pedestrian detection algorithm [30], [51], especially those depending of target tracking. Intrinsic parameters calibration involves an optimization process to fit a set of coplanar visual features to their projections in the image. In [26] an overview of the calibration process may be found. A variation of the previous method for FIR imaging devices is presented in [162].

A common assumption is to consider that the ground in front of the vehicle is flat. If that assumption holds, the relation between the floor plane and the image plane becomes an homography. Monocular algorithms may infer an approximate value of the distance between camera and pedestrian, also assuming that the pedestrian is standing on the ground plane. As the vehicle moves the homography parameters has to be updated. To this end, three methods have been proposed: inertial measurement units, monocular visual features and v-disparity.

Estimating the pose of a monocular setup is challenging, as there is no depth information available. Because of it, pose estimation is simplified by only considering variation in the pitch and yaw angles. The roll angle is supposed to be negligible at any time as is the steepness of the road. In [12] and [29] the pitch angle variations are calculated based on the position of the horizon. Other methods, based on matching of visual features or holistic image correlations between two consecutive images [24] make the assumption that the whole scene captured by the camera lies within a single plane. In [100] the authors demonstrate that 3D geometry can be estimated from monocular images by modeling the interdependence of objects, surface orientations, and camera viewpoint.

Stereo-based systems provide a much more rich information of a scene. The flat-world assumption is no longer needed as depth of objects may be accurately calculated. However, by keeping that assumption, the slope of the road may now be easily calculated using the v-disparity algorithm [120]. In it, Laybarade states that the slope may be calculated by assuming that the ground is contained in a plane and that any other plane in the scene is smaller. By horizontally projecting the disparity, the ground plane becomes a line, which is detected using the Hough transform. This approach can be generalized to model complex road surfaces, such as high order polynomial curves. A review of stereo preprocessing methods used in pedestrian detection application is presented in [132].

Video Stabilization Video stabilization is another useful preprocessing step, specially on systems that rely on tracking to detect pedestrians. In [24] the authors present an

evaluation of monocular image stabilization techniques applied in automotive applications. The existing algorithms are grouped into three different approaches:

- Based on signature: a signature of each image in the video sequence is used to estimate image shifts caused by vehicle pitch. Signatures are generated using the horizontal edges histogram in [29]. In their work Broggi et al. states that this approach is especially effective in FIR videos with high contrast.

- Feature tracking: features are small regions in an image that may be uniquely identified after a pose change of the camera. The features extracted from an image in a sequence are cross-correlated with the ones extracted from the next frame. In the most simple form, and assuming the whole scene is contained in a coplanar surface, the pose change between two consecutive frames can be approximated by a simple rototranslation transformation. If depth cannot be disregarded, a stereo approach may solve the shortcomings of monocular approaches. An early review on fast features for automotive applications may be found in [195].

- Correlation tracking: the computational complexity of feature calculation and matching has led to some authors to follow a holistic correlation approach [139]. Assuming only a vertical and a horizontal shift of the camera between two consecutive frames, a set of image correlations are computed by shifting the latter image in the u and v axis a number of pixels between 1 and a maximum shift value s. The shifted image that produces a better correlation is selected as the new frame of the stabilized video. Though proven to be a robust approach in surveillance applications, the results in automotive applications degrade due to both dynamic objects in the scene and the ego-motion of the vehicle.

  In [12] vehicle oscillations due to uneven pavement are compensated in FIR images (Fig. 2.1). Four types of movements in the images are considered: perspective movements, horizontal translation, vertical translation and vertical oscillations. The stabilizer addresses the latter by evaluating the motion of horizontal edges. Any abrupt oscillation in the position of those edges forces the stabilizer to shift the image in the opposite direction.

(a) (b) (c)

Figure 2.1: In [Bertozzii et al. 2003a] a video stabilizer of FIR images is proposed. (a) Original FIR images, (b) Horizontal edges, (c) Histogram correlation of two images.

2.3. Selection of regions of interest

Object detection in cluttered scenes is a challenging task. The view contains a huge amount of information, much of it irrelevant to the task at hand. Successive steps of the detection algorithm specialize in telling apart to classes of objects: pedestrians and everything else. These classifiers are usually processor-time-demanding, and they usually take the same time to process a sample whether it contains a pedestrian or background. However, by looking at the whole scene, it is obvious that there are parts of the image that does not hold any useful information. These areas are easily discarded using a *fast classifier*. There are many features on which the *fast classifier* can rely to discard some part of an image, such as depth, motion, and, in the case of FIR imagery, temperature or radiance. This results in a small number of regions a interest (ROI) in the image, thus the latter classification step may run faster.

2.3.1. Stereo-based segmentation.

Stereoscopic vision,or simply stereo, involves combining two images captured from coplanar sensors. By finding a correspondence between points in the two images, and the distance between the sensors is possible to find the three-dimensional position of those points.

The difference between the distance to the center of the image in each projection is called disparity. In the case of two identical and parallel cameras this value is proportional to the distance the object, and it can be calculated as $Z = \frac{f \cdot T}{x^l - x^r}$. Where Z is the distance of the object; f is the focal length; T is the distance between the optical center of both sensors; and $d = x^l - x^r$ is the disparity, calculated as the difference between the x coordinate of the projected point in the left and right images.

Stereo vision techniques are fairly common in visible-light computer vision algorithms [237] [31]. Recently, Llorca et al. review the current state of the art on ROI selection using stero vision in [132].

Some authors have incorporated stereoscopic vision techniques to FIR computer vision [95]. Bertozzi et al. use an stereo pair of FIR cameras for pedestrian detection in [20] and [11]. The proposed method is based on estimation of depth of warm clusters in the scene. The procedure results in a set of ROIs that follows certain geometric restrictions. In [17] and [13] this approach is extended by using a two stereo pair system, merging information from visible and thermal infrared imagery.

An experimental analysis on FIR and VL approaches to pedestrian detection using stereo vision is presented in [119] and [118]. The two stereo pairs used in this application are shown in Fig. 2.2. Their candidate bounding-box algorithm involves several steps. For both stereo pairs they perform a dense-stereo matching, processing two distinct disparity images. Then, from each disparity image they process their corresponding u and v disparity, which are histograms that bin the disparity values for each column or row in the image, respectively. This approach allows to easily segment the largest plane in the image, supposedly the ground plane, and thus any other object is subject to be an obstacle. Those obstacles

that adhere to some geometrical restrictions constitute the list of ROIs to be further process. An example of obstacles found in both kinds of images is shown in Fig. 2.3. The experimental methodology followed in this work test unimodal detectors (using only information from either VL or FIR images) and multimodal detections. The first set of experiments demonstrate that stereo-based detection using unimodal imagery achieves high detection rates both for VL and FIR images. They used histograms of orientation of FIR, color and disparity. A second set of experiments prove that detection performance can be significantly improved by combining color, disparity and infrared features.

Figure 2.2: Tetracular system used in [Krotosky and Trivedi, 2007a] made up of two VL and two FIR cameras.

Motion-based segmentation Motion is a feature that can indicate the presence of a pedestrian in a sequence of images [72] [97]. It is simple to implement, but ignores pedestrians that are stopped, and its effectivity may be compromised by other moving objects in the scene, such as tree branches. This method is often used in video surveillance applications, in which the camera remains static. In the case of mobile applications, such as ADAS, the motion of the vehicle makes detection of other moving objects a much greater challenge.

- Subtraction of images: this simple technique involves comparing two images taken consecutively in a short space of time. If that time is short enough it can be assumed that the only difference that will exist in the two images is because something in the scene has moved. Differences are also expected due to the noise of the sensors.

- Feature points: In this approach, point correspondences are searched very specific features in both images. Those points located in different zones shall correspond to images of moving objects. Normally, is usually applied stage textit clustering to group the points and determine which belong solid body. In the case of pedestrians, it is necessary to incorporate information from the dynamics of their anatomy, since it can not be considered a rigid body.

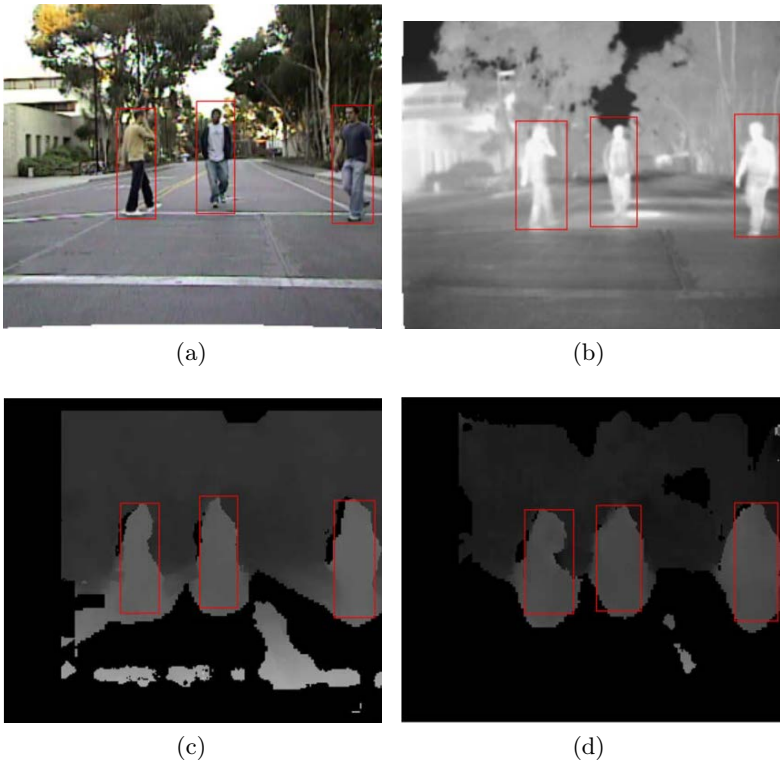


Figure 2.3: Example of stereo processing in [Krotosky and Trivedi, 2007a]. a) VL image with bounding boxes surrounding obstacles with proper dimensions. b) FIR image. c) Disparity map computed from the VL pair, with ground plane removed. d) Disparity map from the FIR pair.

2.3.2. Far Infrared Spectrum

There are two possibilities for nighttime pedestrian detection using computer vision. Since there is not enough light to use algorithms based on visual information in the visible range, an option is to illuminate the scene with infrared LEDs and capture images with near-infrared sensitive cameras [89]. Or, to make use of thermal cameras that capture the heat emission in the far infrared range. There are advantages to using far-infrared cameras versus conventional ones. First, they do not depend on an external light source, instead they project onto the sensor plane the heat emission of the objects so that the image obtained is proportional to the temperature distribution in the scene. Most of the systems developed make use of this feature by selecting regions of interest based on the presence of hot spots [18] [19] [23] [156]. Another important feature is the presence of sharp edges between the background and these hot items. In early work by Fang et al. an study on FIR pedestrian segmentation is presented [70] [71]. In it, the authors review available features for VL images and its application to FIR images, i.e. symmetry of vertical projection and intensity histograms. Besides using thresholding techniques and edge detection Meis et al. filter false positives based on symmetry, by calculating the local direction of gradients [141].

2.3.3. Integration of visible-infrared imagery

Pedestrians present different properties depending on the type of camera used. Some authors combine the information obtained from several vision systems that simultaneously record the same scene. In the most widespread methodology, regions of interest found in the far infrared images are studied in the visible range images [196] [189] [121]. In figure 2.4 the image registration of the system used in [189] is shown. In low light applications, in which external light is insufficient for normal camera usage, near-infrared information is combined with thermal infrared [198].

(a) VL image (b) FIR image (c) Combination of VL and FIR images

Figure 2.4: Image registration of the system used in [St-Laurent et al., 2007]

The most straight-forward technique for object segmentation in FIR images is to binarize the image based on a grey-level threshold. Classic thresholding techniques [165] [113], make way to more complex algorithms. In [39] a multi-level thresholding technique for segmentation specifically targeted at thermal images is presented. The survey in [38] covers entropy thresholding techniques.

2.3.4. Sliding Window approach

The sliding window approach for candidate selection consists on an exhaustive search of the image. Thus, methods following this approach do not rely on previous segmentation steps, instead selecting all possible candidates in an image. The window selection does not depend on the image content but rather on geometrical restrictions. First, the window shape is selected based on the kind of object to detect. For pedestrian detection, these windows usually have a rectangular shape, with a constant width-height ratio. For other problems, such as face detection, a square window shape is used instead. The second parameter is window density, or spacing between windows. Finally, multi-resolution detection is achieved by scaling the sliding window, or the original image.

In [168] the authors introduce one of the first sliding window detectors, by exhaustively selecting small patches of the image and encoding them into Haar-wavelet descriptors. Those descriptors are then fed to a Support Vector Machine for classification. On later work by Viola and Jones [202], the main ideas of this approach are extended, focusing on computing speed gains, and applied to a face detection problem. They demonstrate that discrete Haar wavelets may be efficiently computed using an integral representation of an

image. Instead of using a large set of features, they use Adaboost to automatically select only best performing features. Their detector uses a cascade structure to quickly discard regions.

Sliding window detector were popularized with the introduction of the Histograms of Orientations detector [49]. In it, the authors propose encoding each candidate window into a dense array of histograms of gradient-weighted orientations, an idea that had proved to produce robust discrete features [133].

Following the introduction of the HOG descriptor it was established in [238] that large speed gains could be achieved by pre-computing integral histograms [172]. This method tries to address one of the main drawbacks of the sliding window approach, that is, the slow computation times due to the vast number of potential candidates. A number of methods have been introduced since then to the same end [215], [236]. The number of candidates can be drastically reduced by applying a segmentation algorithm, however sliding windows tend to outperform segmentation [93] or keypoint based algorithms [123], [181] for small pedestrians.

2.4. Silhouette Matching

In its simplest form, silhouette matching involves correlating a binary shape model with a pre-computed template. In [31], the authors propose matching an edge map with an upper-body binary pattern by simple correlation. This pattern is obtained from averaging a number of sample shapes, and is scaled to three different sizes.

Gavrila et al. propose in several articles using the distance transformation of the edge image and computing a pairwise similarity measure with a set of shape examples, in a coarse-to-fine manner. [87], [85], [88], [84], [86].

Edgelets and Shapelets Edgelets features encode local shape as a set of silhouette oriented features. These consist of small connected chains of edges [218] This approach was extended in [220] and [221] to handle multiple viewpoints. Wu and Nevatia also use edgelets as local shape features in FIR images in a pedestrian detection problem [235].

Shapelets are shape descriptors discriminatively learned from gradients in local patches. In [178] propose to use AdaBoost to model. Later, in [59], boosting was again used to combine multiple shapelets.

In [131] a hierarchical multi-feature detector, called granularity-tunable gradients partition (GGP), is proposed. Their descriptor properties range from deterministic description (edgelet) to statistical representation (histogram of orientations).

Snakes Active contours, also known as *Snakes*, are deformable curves which can evolve on an image to delineate the boundaries of an object. There are different methods to fold a *snake*, but generally requires defining an energy related to the position of the edges in the

image. The *snake* try to evolve seeking positions of lower energy. Restrictions may also be added to make it more or less rigid and not be segmented.

Applied to pedestrian detection, once the snake has attached to the contour of the object, it may be determined whether the shape is similar enough to a person. This technique is relatively old, finding one of its first applications in the article by Kass et al. 1988 [112]. Subsequently, other authors have used active contours for pedestrian classification [213] [6] [98].

FIR Silhouette Templates Recognition of the silhouette in far infrared images usually depends on the temperature distribution of the human body. Such systems rely on non-deformable models, which include sufficient information to adapt to the many shapes of the pedestrian class. It is a simple, but has proven to be very robust in comparison with other approaches. One of the first examples that uses a recognition method based in the shape of the silhouette can be found in [156], with a similar development in [33] and [19]. In [137] a hierarchical template-based classifier for FIR pedestrians is proposed.

2.5. Pedestrian Descriptors

Pedestrian descriptors are projections of an image sample containing a pedestrian in a feature-space. These descriptors are used in a subsequent classification step to determine if the sample belongs to one of the two following classes: pedestrians or non-pedestrians.

2.5.1. Holistic Methods

Haar-like Features One of the first successful pedestrian descriptors, Haar wavelets, were introduced in [164] and used in a pedestrian detection problem. Later, Papageorgiou and Poggio apply the same approach in a general detection problem in [168]. An extended set of Haar-like features were later introduced in [125]. A representation of some the Haar-like filter is depicted in Fig. 2.5. Viola and Jones use them in their detector [203] [204], which achieves higher framerates because of the use of integral images and its rejection-cascade structure. Jones apply a similar approach in a pedestrian detector for surveillance applications [107]. Haar features were also used in combination with distance-transform in [137] on FIR imagery. In [191] Haar-like features are used to detect pedestrians in FIR images from an automotive platforms, employing an implementation that focus on real-time performance. The flow chart of the system is shown in Fig. 2.6.

Discrete Feature Points Feature points are small spatial areas in an image that are persistent in different views of the same scene. Descriptors are parameterizations of those feature points, so that each can be uniquely defined, being clearly distinguishable from other similar points. The simplest kind of descriptor are corners, image areas with high values for the second order derivatives. A comprehensive study on the use of corners as descriptors may be found in the work of Harris [96].

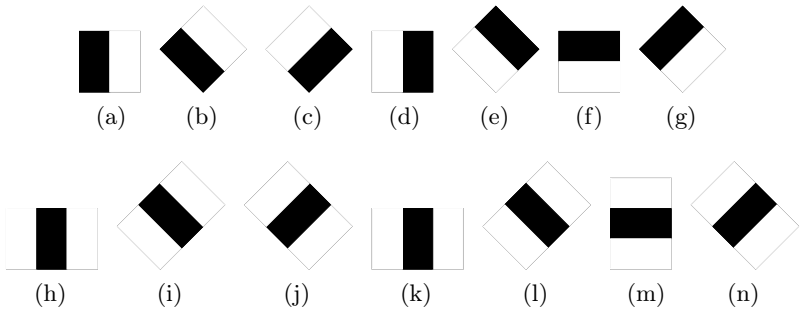


Figure 2.5: Examples of simple Haar-like filters. (a) to (g): Edge features. (h) to (n): line features

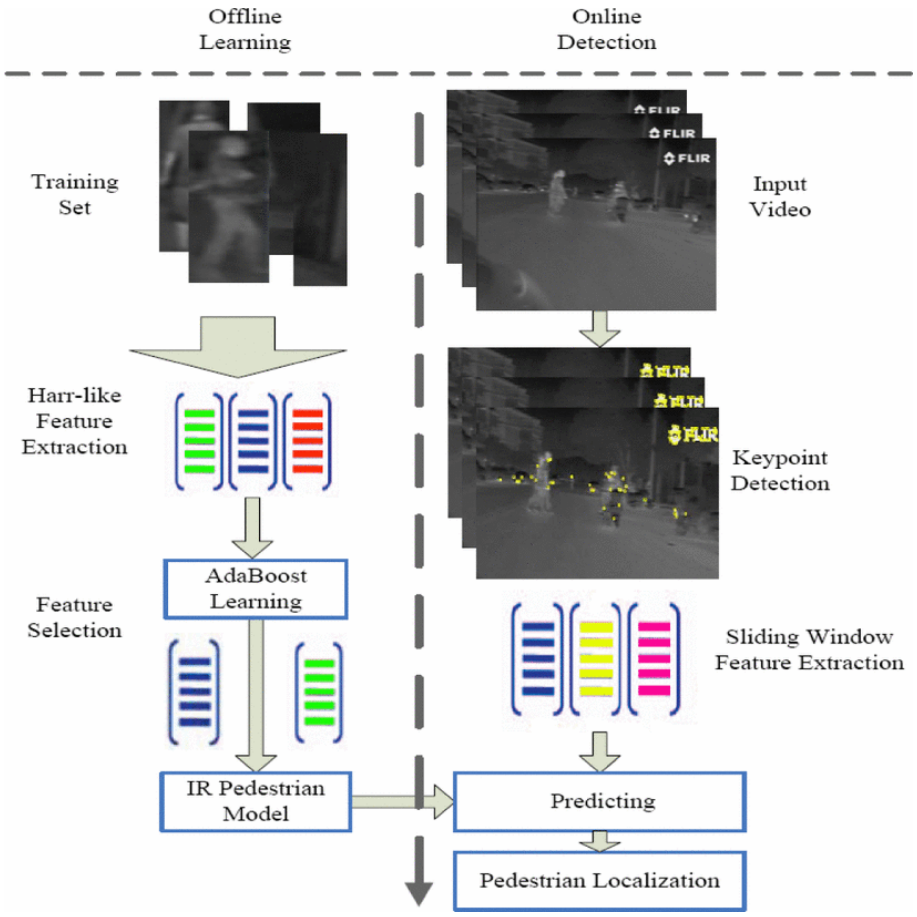


Figure 2.6: Flow chart of the FIR pedestrian detection system proposed in [Sun at al., 2011]. They use Haar-like features in an AdaBoost learning framework. In order to reduce the number of ROIs, they first extract points of interest in the image and search for pedestrians only in the neighborhood of the detected keypoints.

By themselves, corners are not good descriptors. They can not uniquely define a point in the image, since all corners resemble one another. Building upon Harris work, other authors have developed different types of descriptors. Lowe's SIFT descriptor [135] [133] [134] is a specially relevant one. The main insight of this descriptor is the following: a

maxima or minima appearing in the same image area in different *scale spaces* should be repeatable in different scales. Orientation invariance is achieved by encoding the gradient around each feature point in a histogram of orientations. This descriptor is meant to be invariant to the scale, rotation, perspective and illumination. The main idea of the SIFT descriptor has been further developed by other authors, resulting in new descriptors, such as SURF [7]. In earlier work, Shashua et al. [182] proposed a similar representation for characterizing spatially localized parts for modelling pedestrians.

Besbes et al. also use SURF features, this time in FIR images from a camera mounted in a moving vehicle. They use an SVM as classifier and a hierarchical codebook of scale and rotation-invariant SURF as the discriminative feature [22]. Their implementation prove to be partially immune to difficult recognition situations, such as occlusions.

Histograms of Orientations The HOG descriptor, as introduced by Dalal and Triggs in [49], is generic in nature and can be used to classify any type of object, but it is in the people detection topic where it has found a widespread use. Its operation is inspired by SIFT [133], defining the shape of an object as a dense grid of histograms of orientation, instead of using them as discrete descriptors around a feature point. Insofar as the descriptor resembles a trained model, it is decided whether or not the image contains a pedestrian. In the original implementation, it uses a support vector machine (SVM) for linearly separating pedestrian and non-pedestrian classes. Figure 2.7 illustrates a HOG descriptor of a cropped sample image containing a pedestrian, as well as the SVM weights of the trained model.

This approach has had an influence on many descriptors since. In [238] the authors propose using HOG features with an Adaboost learning algorithm, for faster detection rates. In [50] the authors propose a spatial selective method that removes less important information out of the HOG feature vector. They report achieving slightly better performance than the original detector by adding to the feature vector multi-level information. In [207] HOG dimensionality is reduced by applyin a locality preserving projection. In [226] the R-HOG feature is proposed, which creates binary patterns from the HOG features extracted from two local regions, thus reducing memory requirements.

Some authors propose encoding pedestrian contours as histograms of orientation. A geometric active contour model is used in [211] to track the silhouette of a pedestrian, which is encoded as HOG features, extracted on a set of points located on a narrow band around the contour. In [229] histograms of orientations are computed by analyzing the diffusion tensor fields of the suggestive contour extracted from different viewpoints of a 3D model.

Computation speed is a relevant factor in the intelligent vehicles field. There have been some efforts to implement optimized versions of the HOG detector that can run in real time [215]. In [234] and [215] two different HOG GPU implementations are presented. A more recent GPU implementation of the HOG descriptor is described in [227].

HOG features have been used in many object detection applications, such as cyclists [109], traffic signs [233] [166] [232], general purpose object detection for service robotics [61], gesture recognition [110] and even fabric defect inspection [183].

The HOG descriptor has been successfully tested in pedestrian detection in infrared

images [190] [235] [14] [235] [144]. Ambient temperature has a big impact on pedestrian appearance, specially if they wear clothing with varying degrees of thermal insulation. In [163] a preprocessing step is applied to the candidates, before computing the HOG descriptor, which compensates for variations in clothing temperature using vertically-biased morphological closing.

However, and regardless the popularity of HOG features, its performance on very large, general purpose, databases is proving to be limited. In [194] the authors explore the representation capabilities of the HOG descriptor, concluding that an impostor image can be morphed into an image sharing the same HOG representation as the target object, while retaining the initial visual appearance.

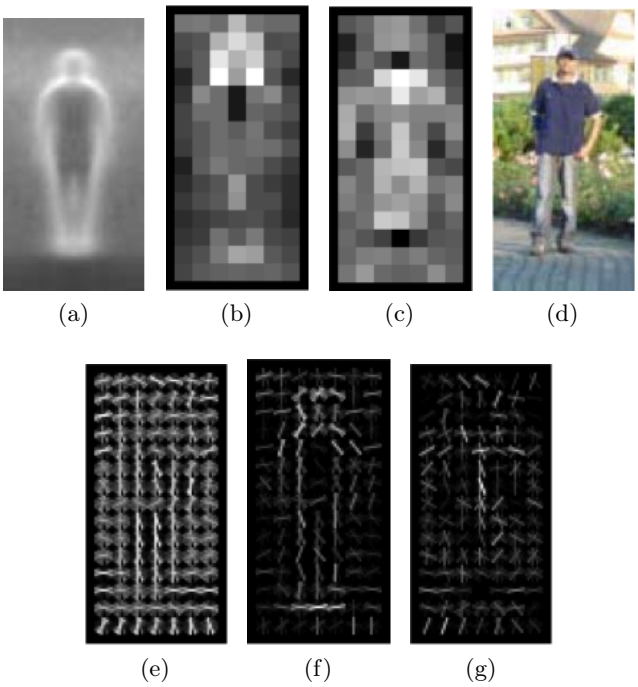


Figure 2.7: HOG descriptor. (a): Average gradient of the INRIA Pedestrian Database training set. (b): Maximum positive SVM weight of each descriptor cell. (c): Maximum negative SVM weight of each descriptor cell. (d): Cropped image sample. (e): Illustration of its HOG descriptor. (f): Positive weight normalized descriptor. (g): Negative weight normalized descriptor. Source: [Dalal and Triggs, 2005]

Other descriptors There are a number of works describing pedestrian descriptors in images. Notably, in [181], the authors evaluate the performance of the descriptors *Shape Context* and *LocalChamfer*. Other methods do not take into account any real-time operation restriction. Among them is the approach adopted in [167], which classifies objects according to an optimization of a swarm of particles. In this case, the descriptors are densely calculated and processing speed is lower than real time systems.

Local Binary Patterns (LBP) are a fast and straightforward descriptor that need little resources for its computation. Some authors, focusing on real-time detection performance,

use LBP as a stand-alone descriptor. In [223] LBP is used in a pedestrian-detection problem in FIR images. Representation of a pedestrian in a FIR image using multiblock-LBP is shown in Fig. 2.8.

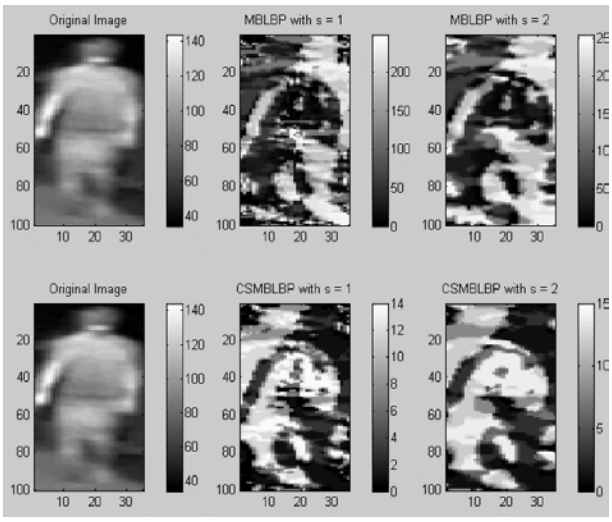


Figure 2.8: Representation of a pedestrian in a FIR image using multiblock-LBP [Xia at al., 2011].

2.5.2. Part-based descriptors

Part-based detectors are based on many descriptors, for the different parts of the human body, as opposed to holistic methods. The simplest part-based methods, rely on manually annotating different *parts of interest*. In [149] four independent detectors are trained, for head, legs and both upper limbs. On a second stage the four parts are combined in a linear SVM. Similarly, in [182], a higher number fixed parts is proposed. Each part classifier is treated as weak classifier in an AdaBoost approach. There are many examples of part-based detection that follows the same paradigm [218], [145], [65]. Wu and Nevatia [217] propose using edgelets as features of four body parts (full body, head-shoulder, torso, and legs) and three view categories (front/rear, left profile, and right profile) in an Adaboost learning method.

In previous methods, the positions of pedestrian parts are manually annotated. Intuitively, parts containing limbs or heads should contain relevant information. In [210] objects are represented as flexible constellations of rigid parts. Parts are computed and selected in an unsupervised manner. This work is extended in by learning scale-invariant object models using entropy-based features

Methods that automatically select parts, such as latent SVM [74] [75] prove regions other than head and limbs may also contain discriminative information. Felzenszwalb et al. propose data-mining hard negative samples, combined with an iterative learning method, that optimize the position of pedestrian parts. The detector is based on a root filter and many part filters, at double the resolution. This imposes a limit on the minimum size a pedestrian has to have in order to be detected. In [169] the authors propose using a similar

part-based detector for large pedestrians, switching to a holistic approach for small ones. Felzenszwalb et al. use HOG features as the pedestrian descriptor, while also stating that their method should work with other kinds of features. This was explored in [57] by using Haar-like features.

In [197] human pose is estimated discriminatively using structure learning. After the most likely pose is identified, the authors use as the classifier an SVM fed with local histograms of oriented gradients and local PCA of gradient. They state that pose estimation significantly improves the accuracy of the detector for people in configurations that are very uncommon, such as riding a bicycle.

Wu et al. address the problem of detecting partially occluded pedestrians by using an assemble of part detector, learned by boosting a number of weak classifiers which are based on edgelet features [217]. Possible occlusions are integrated into a joint probability model based on the responses of detectors parties.

In [1] the authors address the issue of pedestrian detection in cluttered images by using a subtractive clustering attention mechanism based on stereo vision. Candidates are selected based on a nondense 3-D geometrical representation. Using a parts-based approach the authors claim that the detector is able to deal with variability in pose, illumination, occlusions, and rotations.

While local part-based detectors are able to handle occlusions, holistic methods achieve better results in normal conditions. A combination of both approaches is presented in [128], where the authors combine local parts templates with a global template-based scheme, using a Bayesian optimization scheme.

2.5.3. Multi-feature Methods

Pedestrian descriptors can be combined to include in a single feature vector complementary information. There many examples of this approach in the literature. In [222] the authors combine HOG, edgelet and covariance features, achieving better results that any of any those descriptors on their own. Local Binary Patterns (LBP) [159] were combined with HOG in [208]. A variation of LBP, local tertiary patterns, are used with HOG in [103]. Walk et al. [205] add to the multi-feature vector colour self-similarity and motion features.

Another approach that focus on combining different kinds of information may be found in [59]. In their work a channel is defined as a registered map of the original image, where the output pixels are computed from corresponding patches of input pixels by applying a linear or non-linear transformation to the original image. As example channels, they propose using LUV color channels, grayscale, gradient magnitude and histograms of orientations, as shown in Fig. 2.9. All selected channels can be computed using the *integral image* paradigm. This approach is extended in the *Fastest Pedestrian Detector in the West* detector [58], which approximates features at nearby scales for efficient multi-scale detection in full-sized images.

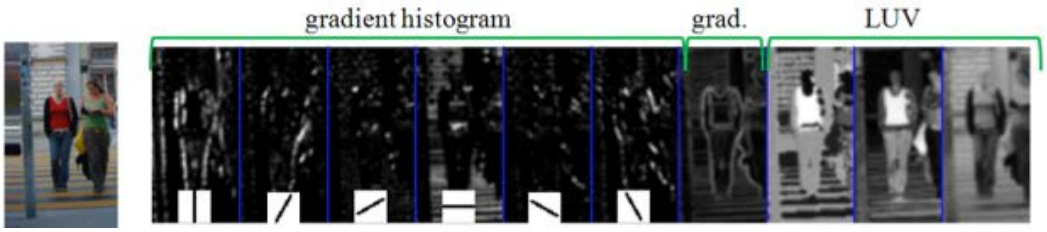


Figure 2.9: Integral Channel Features as described in [Dollár et al., 2009a]: Gradient histograms, gradient magnitude and LUV channels. All these features can be computed using integral images, thus having an efficient computation.

2.6. Classification

There are a number of classification methods that have been applied to the problem of pedestrian classification using local spatial features. Among them boosting and support vector machines should be highlighted.

Boosting The principle of *Boosting* is that it is possible to learn a good classifier from many *weak classifiers*, where week a classifier is defined a classifier with a level of accuracy only slightly better than chance. In the boosting family of classifiers each weak classifier or a subset of the total, is evaluated iteratively, assigning a weight depending on how many times they appear in the training set and how well they classify by themselves. The main advantage of these algorithms is usually its fast performance while classifying a new sample. This approach needs only a limited number of descriptors to discard an image. At any stage, after evaluating only a subset of weak classifiers, a negative result terminates the algorithm, so it is not necessary to calculate the remaining features. In [81] a good introduction to *Boosting* algorithms may be found. Viola and Jones apply a boosting algorithm called *Adaboost* to learn objects from a large set of weak classifiers. In their application, Haar-like features are used as the weak classifiers [203]. Tuzel et al. [199] [200] modify the boosting framework to work on Riemannian manifolds. They use local covariance as features, which are vector space independent.

Wang et al. use the HOG feature in [209] but, instead of using a linear kernel SVM, they threshold the histogram of orientations of each cell and use the results as weak classifiers. They state that the resulting detector has similar performance as the original Dalal-Triggs implementation in the INRIA database, while being much faster.

Support Vector Machines SVM is a classification algorithm which separates the feature space using a hyperplane in higher dimensional space. The projection of higher dimensional data often make them more easily separable. The algorithm selects a plane such that the spacing between classes is maximized. This plane can be defined just with the elements of each class that are closer to this, known as support vectors.

The use of SVM for feature-based pedestrian detector was popularized by the success of the HOG descriptor [49]. In their implementation they use a linear kernel. However, other SVM kernels have been proposed and used in feature-based object detection. In [62] the

authors review SVM kernels for detecting objects based on an unordered set of discrete local descriptors. They test different combinations of kernels (Matching kernel, Bhattacharyya kernel, Kernel Principal Angles) and descriptors (SIFT, JET [180], Image Patch). In [138] the HIKSVM is introduced, which uses an approximation to the histogram intersection kernel. The computational efficiency of this approach allows for the use of complex kernels in almost real-time. Owechko et al. propose an efficient search mechanism of features based on swarm intelligence [167]. In their implementation they demonstrate the use of a particle swarm optimization algorithm to this end, and apply it to a FIR pedestrian database.

2.7. Verification and Refinement

Some systems use an additional refinement step to disregard false positives using a method complementary to the classification step. The techniques used tend to look for simple cues of pedestrian geometry. For instance, a vertical symmetry check is performed in [31], where the authors use vertical edges to discard detections that are non-symmetric around the central vertical axis. Regarding pedestrian refinement in FIR images, some authors use a 2D [12] or 3D model matching [15], [32] as well as vertical symmetry [13] to verify detections. Another validation method involves crosschecking results from independent detections in two images of a stereo pair. In [88], [86] the authors verify detections by cross correlating the silhouette extracted in both stereo images.

Verification is sometimes used after the tracking step. Temporal integration of the detections is used in [79], where the same authors extends the work in [86] by analysing the gait pattern of pedestrians walking perpendicular to the movement of the vehicle. In [99] gait analysis is performed in FIR images, this time by applying a markov network that can discriminate open and close legs of a pedestrian being tracked. Verification methods involving tracking assume that the pedestrian is not occluded in most the frames. If gait pattern recognition is applied, the algorithm also needs to accurately detect the legs. Other multi-frame refinement methods use even more cues other than gait pattern recognition, such as motion tracking [182].

Classification methods relying on the sliding window methods usually show multiple detections for the same pedestrian. Detection windows neighbouring the ground truth window usually output a high score classification. In this case, the refinement step involves clustering detections into just one, in a process known as non-maximum suppression (NMS). Mean shift [44] is used in [48] to select just one of the detections of the multi-scale clusters around pedestrians. Another NMS algorithm commonly applied to pedestrian detection is Pairwise Max (PM) [74]. It involves rejecting detections that overlaps with any other with a higher score in the classification step

2.8. Tracking

Tracking can greatly simplify the detection of pedestrians. If the detection algorithm has a low failure rate it is usually more productive to add a tracking stage to the algorithm,

rather than trying to find a method that provides perfect hit rates. Tracking has other advantages, as being able to predict the pedestrian future trajectory, locating the pedestrian in the case of temporary occlusion, and selecting ROIs based on those predictions. However, there is a lack of literature that addresses the subject of tracking infrared pedestrians.

2.8.1. Kalman

The most common solution in VL systems is to use a Kalman filter to determine the position, as applied in [114] and [224]. In [80], the authors propose using two Kalman filters, separating lateral and longitudinal motion. Another approach followed in [16] and [23] uses a Kalman filter to track the ROI position in the image, in the second case using an Inertial Measurement Unit (IMU) to include the egomotion of the vehicle into the filter. Other authors have used variations of this method as the *Extended Kalman Filter* (EKF) [108] [206] and *Unscented Kalman Filter* (UKF) [142].

2.8.2. Particle Filters

Particle filters, also known as Sequential Monte Carlo (SMC) methods, estimate posterior density of the state-space by implementing the Bayesian recursion equations. Introduced in [92], particle filters provide a solution for estimating non-linear non-Gaussian transformations, which do not rely on local linearization, as the Extended Kalman Filter does. The main drawback of this techniques is its computational complexity.

Particle filters have become a popular technique for tracking pedestrians in images. Many different cues have been proposed to that end. In [23], the authors propose tracking silhouette, stereo and texture of pedestrians in a three-dimensional space, using a particle filter.

The Conditional Density Propagation algorithm (Condensation), as introduced in [171], detects and tracks the contour of objects in a cluttered background. It is an application of the Sequential Importance Resampling algorithm (SIR) proposed in the original article by Gordon et al. The Condensation algorithm is used in [171] to track a silhouette model of pedestrians, which consists of Euclidean transformation and deformation parameters. In [53] this approach is extended to work from a moving vehicle.

In [4] the authors feed the classifier only with regions of interest generated by the particle filter tracker. Colour cues are used in [122] to track in a space-time volume the trajectory of the object. Chateau et al. propose using using statistical learning algorithms [40] as a likelihood observation function of a particle filter. This approach is able to simultaneously detect and track objects. Their work include two demonstration of said idea: one using an SVM as the classifier, the other using Adaboost.

Particle filters are computational expensive. Some attempts have been made to speed up computation by, for instance, parallelizing the execution in graphics processing units [136].

2.8.3. Other techniques

Surveillance systems, where the camera is in fixed position, often rely on image differences or optical flow [104] to track objects. A graph matching-based pedestrian tracking algorithm is presented in [47], aimed to pedestrian surveillance application in FIR images. A similar approach for pedestrian tracking in FIR images is suggested in [46] . First images are segmented based on motion, using a generalized expectation-generalization algorithm. Then pedestrian tracking is formulated as a matching problem on weighted bipartite graphs. For a classic review of object tracking in images refer to [228].

Tracking pedestrians from a moving vehicle is more challenging and more advanced techniques has been developed. Some authors track discrete descriptors belonging to a pedestrian, using a recursive algorithm such as *Mean Shift*. Within a search window the centroid of the contained points is calculated. Then the center of search window is placed over the previously calculated centroid, and the new centroid is calculated. The process ends when the difference between the new and the old centroid is below a threshold. Xu et al. use *Mean Shift* in their article *Pedestrian detection and tracking with night vision* [224]. Swarm intelligence, a family of biological-inspired algorithms, has been used to track pedestrians in [157]. In it the authors describe a Bacterial Foraging Optimization algorithm used to track a part-based pedestrian model.

Occlusions and camera movement are major challenges in pedestrian tracking. A pedestrian may be occluded by objects in the scene of by other pedestrian. The research presented in [219] address this issue by proposing a part-based tracking technique. The pedestrian model is a joint representation of four body parts and a full body descriptor. Detections are matched between frames, so that two detections are matched if the detection response is similar. If there isn't a correspondence detection are tracked using the meanshift algorithm.

Finally, detectors that use active contours to identify pedestrians [185] allow the snakes calculated on the previous images to evolve in the new ones. The same authors propose in [184] the use of multiple tracking algorithms. Their tracking system first use a head detector to initialize the trackers. After that, pedestrian are tracked using an active shape tracker and a region tracker, which splits and merges multiple hypothesis.

2.9. Other Important Issues

2.9.1. Sensors and Fusion

It may be intuitively appreciated that images contain a large amount of information. Features in images that can be used to differentiate objects include texture, two-dimensional geometry, colour and motion in the case of image sequences. It also includes contextual information about the size and geometry of known objects. Other visual systems also allow obtaining distance information, in the case of stereo pairs, or temperature, in the case of FIR cameras. However, the same richness of information that makes them so useful for the task of detecting pedestrians makes it particularly challenging.

The disadvantages of using visual information include, among others, the problems associated with illumination, background complexity and target complexity. The latter disadvantage is particularly true in the case of dynamic objects. People are highly variable in appearance due to the degrees of freedom of their anatomy and the variation of texture and colour of clothing. In pictures FIR temperature fluctuations also adversely affect the performance of detection algorithms.

The information extracted from different sensors can complement each other. Laser scanners can provide an accurate distance map and are widely used in scene segmentation and obstacle detection. Though less reliable than laser scanners, radar is also used in ITS applications, specially under difficult weather conditions [42]. Sensor fusion usually follows one of the following approaches: low level, on which ROIs are generated by combining the sensors information, or high level, on which each sensor independently generates ROIs, which are later combined.

Premebida and Nunes propose a system encompassing three sources of information: a laser scanner is used to cluster and track objects, an sliding window pedestrian detector in VL images validates the clusters generated by the laser scanner [174]. This approach is expanded in [173] by adding contextual information obtained from a semantic map of the roads. An example of the output of their pedestrian detector system, and the experimental platform used in their experiments can be seen in Fig. 2.10. In [72] a laser scanner is combined with a FIR camera, using a Kalman Filter to fuse detections.

(a) Regions of interest (b) ISRobotCar

Figure 2.10: Regions of interest generated using a laser rangefinder, presented in [Premebida and Nunes, 2013]. The experiments where conducted in the ISRobotCar, in Coimbra, Pt.

Radar is fused with VL and FIR images in [146] . In their two-step method, a set of ROIs are generated by the radar sensor and later checked by the vision systems.

In [119] a stereo VL and a monocular FIR camera are used together. Regions of interest are extracted from a disparity map and evaluated using VL, FIR and disparity information. In [17] pedestrians are independently detected in two stereo system. The resulting detections of the VL and FIR stereo systems are then fused together based on the percentage of

overlapping and their distance in world coordinates. A positive detection then requires that both systems generate a ROI for the same pedestrian. Other authors propose the generation of ROIs based on temperature segmentation in FIR images and the analysis of said ROIs in visible images [196] [189] [121]. In [41] FIR images are segmented by seeded region groing of warm areas. Then VL and FIR images are fused together. Detection results prove to be better than any of the FIR and VL detectors on their own.

2.9.2. Applications

There are two main applications of pedestrian detectors in ITS: ADAS and automatic driving. In both cases, a complete application would feed location and trajectory of the detected pedestrians to a decision module that determines if any action is necessary. In the ADAS case, the final step of the system has to promptly communicate the driver any dangerous situation in an unobtrusive way. In the case of automatic driving, the speed and trajectory of the vehicle has to be updated based on the presence of pedestrians in the path.

For an overview of how detectors are incorporated into full automotive systems that utilize stereo, scene geometry, tracking, or other imaging modalities (e.g. [86], [1], [5], [67], [216]), we refer readers to [83], [54] and [91].

2.10. Other surveys in pedestrian detection and recognition

Pedestrian detection in images is a very broad topic. This review of the state of the art tries to highlight the most representative research in each of the described sections. However, readers may want to refer to the following surveys. Each covers the state of the art on pedestrian detection, and other topics, from different perspectives and up to the date of publication.

- Vision-based intelligent vehicles: State of the art and perspectives [21].
- Pedestrian detection for driving assistance systems: Single-frame classification and system level performance [182].
- Vision Technologies for Intelligent Vehicles [10].
- Pedestrian protection systems: Issues, survey, and challenges [83].
- Pedestrian detection: A benchmark [60].
- Monocular pedestrian detection: a survey [66].
- Study on pedestrian detection and tracking with monocular vision [94].
- The Applications and Methods of Pedestrian Automated Detection [101].
- Survey on Pedestrian Detection for Advanced Driver Assistance Systems [91].

- Pedestrian detection: An evaluation of the state of the art [56].
- Backgroundless detection of pedestrians in cluttered conditions based on monocular images: a review [186].
- Thermal cameras and applications: a survey [82].

3

Classification

3.1. Introduction

In this chapter the classification of pedestrians in FIR images is addressed. Classification is defined as the decision of wether a a fixed set of cropped images belongs to the pedestrian or the background classes. Classification differs from detection in that, in the latter, the pedestrian have to found in full-size images. Other aspects, such as the procedure for the selection of regions of interest and non-maximum suppression methods must be considered. Any detection algorithm must have a classification algorithm to keep or discard the selected windows. By improving the classification methods, the overall detection process is also improved.

Most of the recent research in pedestrian pattern recognition is based on visible light (VL) images. FIR images share some key characteristics with their VL images counterparts. They both are 2D representations of a scene captured by redirecting electromagnetic waves by means of a lens, light in the first case and infrared radiation, which is proportional to the objects temperature, in the second. Some of the key ideas on pedestrian classification in VL images can be extended to work on FIR images, exploiting common characteristics of both, or adapt them to take benefit of the different kinds of information provided by FIR images. Mobile vision applications, such as ADAS, relying on microbolometer sensors have some intrinsic difficulties, for instance their sensitivity curve of an uncooled microbolometer sensor changes very quickly with minimum changes of its temperature [73].

There are several examples in the literature of systems that exploit FIR images to detect pedestrians in night driving. Most of them rely on temperature segmentation. The main objective of this work is to develop a classification algorithm that can achieve high recall rates and low miss rates in a temperature un-biased database. That is, the database has no information about the temperature of the sensor, nor of the environment. The gray-level of the FIR images used represent the relative temperature of the objects, but absolute temperature is not known.

The results in this chapter are based on a new public dataset of cropped pedestrian images, captured with a low resolution, uncalibrated, non-refrigerated microbolometer sensor from a static or moving vehicle. Those images where captured under a different illumination and temperature conditions, including warm summer.

3.1.1. Chapter Structure

This chapter is structured as follows.

- Characteristics of the FIR image-based pedestrian dataset are discussed in sections 3.2, including the methodology of acquisition and sample selection, as well as useful statistics.
- In section 3.4 a classification scheme based on the histogram of oriented phase congruency to detect pedestrians in infrared images is presented. The phase congruency theory, on which our Histogram of Oriented Phase Energy (HOPE) descriptor is based is fully explained in section 3.4.1. The procedure for the descriptor extraction and the classification procedure are further described in sections 3.4.2. The impact on the classification performance of the different descriptor parameters is discussed on section 3.4.3. The classifier parameters are evaluated in section 3.4.4. In section 3.4.5 the impact of noise on the classification performance is evaluated. This section also presents results on a multi-scale version of the descriptor.
- In 3.5 the Int-HOPE descriptor is presented. This descriptor integrates different sources of information in a Random Forest Classifier. The selected features can be computed using the *integral image* paradigm.
- In section 3.6 an analysis of several well known VL pedestrian classifiers applied to FIR images is presented: Principal Component Analysis (PCA), Local Binary Patterns (LBP) [159], and Histogram of Oriented Gradients (HOG) [49]. All descriptors are tested using a number of pattern recognition methods. Comparative results are further discussed in this section.
- Finally, conclusions and future work are presented in section 3.7.

3.2. Classification Dataset

One of the contributions of this work is our pedestrian classification dataset, which consists of FIR images collected from a vehicle driven in outdoors urban scenarios. The dataset was recorded in Leganés, Spain and Coimbra, Portugal. Images were acquired with an Indigo Omega imager, with a resolution of 164×129 pixels, a grey-level scale of 14 bits. The camera was mounted on the exterior of the vehicle, to avoid infrared filtering of the windshield.

3.2.1. Pedestrian Datasets

The availability of publicly released datasets for pedestrian classification has been a key element that helped advances in the ITS area. It provides a way for researchers to test and benchmark new classification algorithms in a way that can be directly compared with other works. It is also useful for replicating experiments performed by other research groups.

Regarding pedestrian classification in VL images, there exists a reasonable number of benchmark datasets publicly available, such as: MIT [168], CVC [90], TUD-det [3], INRIA [49], DC [151], ETH [68] and Caltech [60]. For an overview of recent work on pedestrian classification on these datasets, we refer to [91], [64] and [56]. In the case of FIR images, there is a lack of a complete pedestrian dataset that could serve as a tool to benchmark new features and methods.

In this domain, datasets are usually divided into two types: classification and detection datasets. In the first one, a fixed set of cropped windows containing pedestrians and background is provided, while detection datasets consist on full images with annotated locations of pedestrians. Usually, a subset of full-frames, with no positives (pedestrians), are provided for negative examples extraction. The method for background sample extraction varies from one author to the other, so the classifiers are not really trained on the same data.

A classification dataset is useful for approaches based on the sliding window paradigm. This detection technique consists on analyzing an image by shifting a fixed sized window in the horizontal and vertical axis. This approach can be extended to a multi resolution search by incrementally resizing the original image. Each window analysis becomes independent from all the others and, as such, the detection turns into a classification problem. Improving the classifier performance would also improve detection performance. The classification performance is usually expressed in terms of miss rate vs. false negative rate per window, while per frame is more suitable for detection performance.

In [151] introduced the DC classification dataset is presented. It consists of 4000 up-right pedestrian and 25000 background samples captured in outdoor urban environments. All of them are resized to 18×36 pixels. In their work, Munder et al. evaluate Haar [168], PCA and LRF [105] in combination with neural networks and Support Vector Machine (SVM) classifiers. From their results it can be concluded that the size of the dataset is a key element in improving the classification performance. For the extraction of a large number of background images they apply bootstrapping [193] techniques. The dataset is split into 3 train and 2 test subsets, for cross-validation purposes. In [49] Dalal et al. presented the INRIA dataset, which is still widely used nowadays. It consists on 2478 128×64 cropped images of people for training, and 566 for testing, along with full images for negative extraction. The images were selected from a collection of photographs acquired in urban and rural scenes, and not initially thought to serve as a dataset for driving assistance systems. More recently, Dollar et. al introduced in [60] the Caltech Detection Dataset, as well as a benchmark of several pedestrian detection algorithms. Their results were further extended in [56]. This dataset contains approximately $250k$ labelled pedestrians within several video sequences acquired from a moving vehicle in urban traffic.

Based on the methodology followed by the mentioned datasets, a new FIR pedestrian dataset has been created and made publicly available[1]. The results derived from this study are based on this dataset. The LSI FIR pedestrian dataset is divided in two parts, classification and detection. The Classification Dataset contains a preset of cropped images of positives (pedestrians) and negatives (background), rescaled to the same dimensions.

[1] *http://www.uc3m.es/islab/repository*

The Detection Dataset contains full size images and labels indicating the position and dimensions of each pedestrian. Table 3.1 synthesize some important characteristics of the mentioned pedestrians database in VL as well as the OSU thermal pedestrian database and the LSI FIR pedestrian dataset.

Table 3.1: Pedestrian databases. The first 13 databases are built with images in the visible light (VL) spectrum. Their information is extracted from [Dollár et al., 2012] . The LSI and OSU databases contain images in the FIR spectrum.

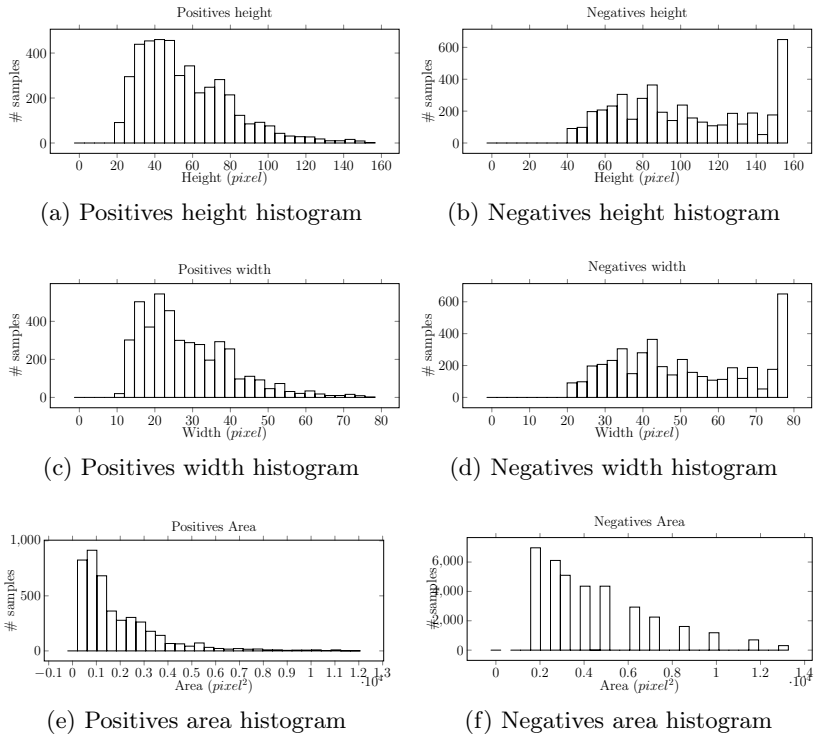
		Training			Testing			Height				
Database	Spectrum	#pedestrians	#neg.images	#pos.images	#pedestrians	#neg.images	#pos.images	10% quartile	median	90% quartile	video	year
MIT	VL	924	-	-	-	-	-	128	128	128		2000
USC-A	VL	-	-	-	313	-	205	70	98	133		2005
USC-B	VL	-	-	-	271	-	54	63	90	126		2005
USC-C	VL	-	-	-	232	-	100	74	108	145		2007
CVC	VL	1000	6175	-	-	-	-	46	83	164		2007
TUD-det	VL	400	-	400	311	-	250	133	218	278		2008
Daimler-CB	VL	$2.4k$	$15k$	-	$1.6k$	$10k$	-	36	36	36		2006
NICTA	VL	$18.7k$	$5.2k$	-	$6.9k$	$50k$	-	72	72	72		2008
INRIA	VL	1208	1218	614	566	453	288	139	279	456		2005
ETH	VL	2388	-	499	$12k$	-	1804	50	90	189	✓	2007
TUD-Brussels	VL	1776	218	1092	1498	-	508	40	66	112		2009
Daimler-DB	VL	$15.6k$	$6.7k$	-	$56.5k$	-	21.8	21	47	84	✓	2009
Caltech	VL	$192k$	$61k$	$67k$	$155k$	$56k$	$65k$	27	48	97	✓	2009
OSU	FIR	984	-	284	-	-	-	30	35	40	✓	2005
LSI	FIR	$10.2k$	$1.6k$	$4.5k$	$6k$	$4.8k$	$4.2k$	30	60	120	✓	2013

Recorded images were manually annotated, were each pedestrian is labelled as a bounding box. Fig. 3.1 shows some cropped-image examples of positives and negatives of the classification dataset. Original images with annotations are also available, so cropped samples can be generated with any padding around the bounding boxes.

Number of samples The dataset comprises 81592 14 bit one channel images, divided in 16152 positives and 65440 negatives. The train set contains 10208 positives and 43390 negatives, while the test set contains 5944 positives and 22050 negatives. The train and test sets were independently recorded on different dates. Full-size images are also available, in case the training algorithm requires context information, or a hard-negative bootstrapping technique is needed.

Aspect ratio Out of the annotated images, the bounding boxes are resized to a constant aspect ratio $(w/h) = 0.5$ by changing their width (w) appropriately. Figure 3.2 contains histograms for height, widths and areas of positive and negative bounding boxes. The height of positive bounding boxes has a maximum between of 40 and 80 pixels. Those bounding boxes refer to pedestrians standing between $10m$ and $20m$ from the camera. Pedestrians appear up to $50m$. Any bounding box below 10 pixels in height is ignored. The remaining bounding boxes are resized to 64×32 pixels using bilinear interpolation.

Figure 3.1: Example cropped-images of the classification dataset. The two upper rows contains examples of pedestrians acquired under different temperatures and illumination conditions. The lower rows contain randomly selected windows from images containing no pedestrians. For visualization purposes the contrast has been enhanced.



(a) Positives height histogram (b) Negatives height histogram

(c) Positives width histogram (d) Negatives width histogram

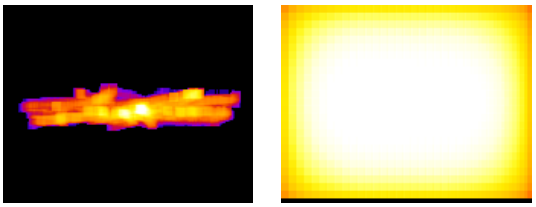
(e) Positives area histogram (f) Negatives area histogram

Figure 3.2: Histograms of bounding boxes sizes and areas for positive and negative samples of the train dataset.

Density Images were acquired from the usual point of view of the driver. As such, pedestrians appear more often in the centre of the image as shown in Fig 3.3, which represents the logarithmic density of the centres of the bounding boxes. In the case of negative samples, the bounding boxes are randomly selected, so the centres appear all over the image, with less density near the borders.

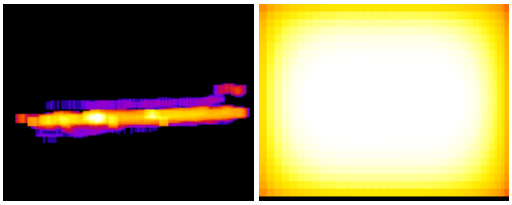
3.3. Probabilistic models

This work has its foundation in the research presented in [99] about FIR pedestrian detection. It was established in [156] that pedestrians can be detected in images by correlating them with a precomputed model. That model is the result of averaging the gray level values of a set of thresholded images, or of a derived function. The main insight of this idea is that, under certain conditions of temperature, pedestrians have a higher gray-level value than their background. Also, the color of clothing is not a factor to consider, contrary to what happens in VL images. Variations of this approach have been explored in [18], [23], [19] and others. Hilario propose computing the probabilistic models by averaging a binarized version of the pedestrian. The threshold selected is based on the histogram distribution of a set of background samples. By setting this threshold to $T = \mu + 3\sigma$, where μ is the mean gray-level value and σ its standard deviation, brighter (hotter) pixels of the image are set to one and darker (cooler) pixels are set to zero.



(a) Positives on Train Dataset

(b) Negatives on Train Dataset



(c) Positives on Test Dataset

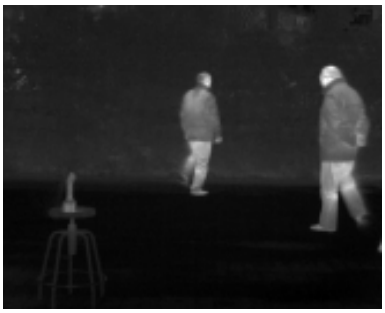
(d) Negatives on Test Dataset

Figure 3.3: Centres of bounding boxes for positives of the train and test dataset on a logarithmic scale.

The work in this section, propose a variation of those previous works, where the discriminant factor is not the gray level of the image, but the temperature of the object. By thresholding the image based only on gray-level information, a probabilistic model approach cannot adapt to every possible scenario. For instance, on a hot summer day, under direct daylight, pedestrian appear darker than the background. Figure 3.4 shows infrared images under different illumination and temperature conditions. If this is the case, the thresholding process would not correctly segment the pedestrians. Another issue with non-refrigerated microbolometers is that the gray level of its images, while being proportional to the temperature of the object, also relies on the temperature of the sensor. As such, in order to measure temperature with this kind of sensor, a radiometric calibration is necessary.

The gray level of each pixel of infrared images represents the amount of heat that the sensor captures at that point. The camera sensibility to external radiancies changes in a way that is also function of the flux of radiance coming from inside the camera, as a result of its temperature. The output of the camera is function of the sensor's and the object's temperatures, among other parameters. As the sensor heats up the apparent temperature of an object also rises. The segmentation based on temperature relies then in a calibration process that relates sensor temperature with object temperature. Since the system only looks for pedestrians, the sensor have been calibrated focusing on a good detection of the lower and upper temperatures of the human body, and also for the average temperature of the head. Since the temperature of the sensor is a known value it is possible to calibrate the sensor sensitivity, relating the temperature of a gray body with gray levels on the image. The resulting curve is an approximation that, as mentioned before, only takes into consideration the temperature of the target object and the temperature of the sensor. The gray level value of the pixels of the sensor also depends on the distance of the object and

(a) Reflection of direct sunlight against a brick wall (28°C).



(b) Cold winter scene (5°C).

(c) Scene under heavy clouds (12°C).

(d) Summer night scene (15°C).

Figure 3.4: Infrared images under different illumination and temperature conditions.

the absorption factor of the atmosphere. However, these parameters can be considered very small for short distances such as the range of pedestrian detection. Another factor, that have not been considered is the gain of the sensor itself. The camera will be more sensitive to a particular wavelength.

Figure 3.5 represents the overall sensibility curve obtained. Three sensibility curves are precomputed for the higher and lower temperatures of the head and the lower temperature of the body. Within the work temperature of the camera the sensibility can be approximated to a cubic function of the sensors temperature.

In this approach images are thresholded and only objects within normal pedestrian temperatures are taken into account. The image can contain objects with a temperature higher than the human body, such a heated parts of a vehicle. Therefore, warm areas of the image are extracted based on their apparent temperature, neglecting objects with temperatures that doesn't match those of the human body.

Pedestrians in FIR images presents a particular distribution of the body temperature. Usually the pedestrians head and legs are the parts of the body that emits more heat, being their apparent temperature barely lower than their real one. Chest and arms are more often covered by thicker clothes, specially in winter, therefore their apparent temperature is usually only a little higher than that of the background. The border's definition is higher if the difference between the pedestrians and the backgrounds temperature is significant.

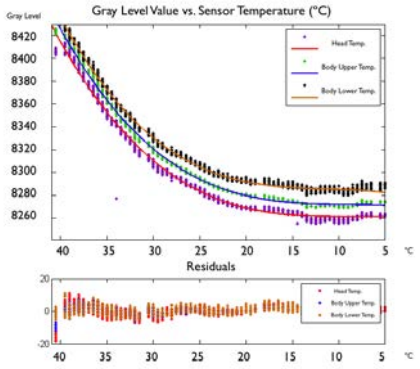


Figure 3.5: Gray Level of three constant temperatures of the human body plotted against the sensor temperature.

Driving at night, pedestrians in images of a far infrared camera present very pronounced edges, and the distribution of their pixels intensities can easily be separated from that of the background. For daylight scenarios the temperature-distribution approach is less effective as the difference in temperature between pedestrian and background is smaller.

The model shown in Fig. 3.6 was calculated from a set of samples with temperature information, which have been thresholded using equation 3.1.

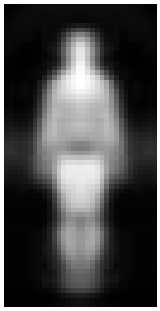


Figure 3.6: Average value of thresholded pedestrian samples.

The classification score of the test samples is the value of the gray scale correlation with some precomputed models.

$$B(x,y) = \begin{cases} 1, & \text{if } \phi_{t_2}(I(x,y),t_s) \geq I(x,y) \geq \phi_{t_1}(I(x,y),t_s) \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

Where ϕ_t is the calibration curve of the bolometer for temperature t, t_s is the temperature of the sensor, t_1 is the lower body temperature of the pedestrian and t_2 is the upper temperature. The models are created by computing the mean value of each pixel of the binarized train subset. Equation (3.2) returns the value of each pixel $M(x,y)$, being

$B_{tr}(x, y)$ the selected samples.

$$M(x, y) = \sum_{i=0}^{N} \frac{B_{tr}(x, y)}{N} \tag{3.2}$$

where N is the number of samples in the train subset.

The score of a sample is calculated by means of a normalized correlation (eq. (3.3)).

$$c = \frac{\sum_{x=1}^{m} \sum_{y=1}^{n} (I(x, y) - \bar{I})(M(x, y) - \bar{M})}{\sqrt{\left(\sum_{x=1}^{m} \sum_{y=1} (I(x, y) - \bar{I})^2\right) \left(\sum_{x=1}^{m} \sum_{y=1} (M(x, y) - \bar{M})^2\right)}} \tag{3.3}$$

where $I_{m,n}$ is each pixel of candidate ROI, M_{mn} is each pixel of the model; $\bar{I}$ and $\bar{M}$ are the mean value of the sample and the model, respectively.

The results of the correlation with the test subset of the FIR pedestrian database is plotted in Fig. 3.7. This results suggest that this kind of approach is effective in FIR images. However this approach also presents some drawbacks. As mentioned before, as the temperature of the camera rises, the gray-level image of the same scene changes. As this temperature gets higher the error of the calibration curve also grows, which would degrade classification results.

The conclusions drawn from this work led to the proposal of a descriptor that is invariant to contrast and illumination.

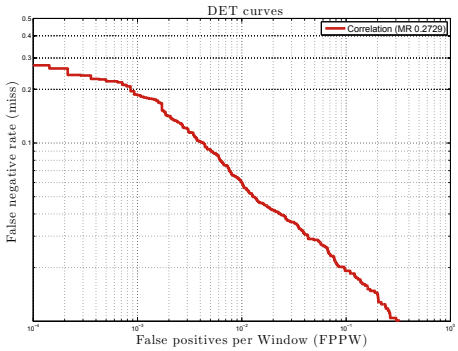


Figure 3.7: Detection Error Trade-off curve of the correlation with the probabilistic models.

3.4. Histograms of Oriented Phase Energy

The images from microbolometer-based cameras reproduce the magnitude of heat emission by the scene objects that hit the sensor plane. The main advantage over visible light cameras is that there is no need of any illumination in the scene, so they can be used in total darkness or, as in this case, while driving at night. In any case, FIR images contain relevant information even in sunny and hot conditions. The work presented in this section

aims to provide a classification method for pedestrians in FIR images that is independent from ambient temperature.

This kind of cameras usually are sensible in a much wider spectrum than their VL counterparts, and this sensibility is greatly determined by the sensor's own temperature. The variation of this temperature shifts the image histogram in a way that is both non-linear and dependent of the specific sensor being used. The quality of far infrared images can easily degrade as the external temperature rises. In the case of pedestrian detection the challenge is even greater, as there is a wide range of appearances a pedestrian can have due to the different kinds of clothes worn throughout the year.

Usually, pedestrian classification algorithms are based on edges information. During the development of this Thesis it has been found that simple gradients in far infrared images are not enough to satisfactorily define the shape of pedestrians. This is due to the much wider infrared spectrum, compared with visible light. Another difficulty is that the sensitivity curve of an uncooled microbolometer sensor changes very quickly with minimum changes of its temperature. To overcome these challenges, a contrast invariant descriptor for object detection is proposed.

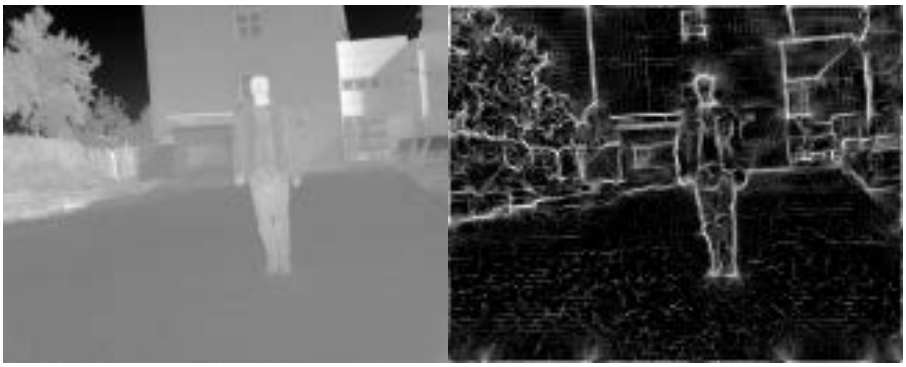
The features should be invariant to illumination, scale and contrast. The theory of phase congruency in signal analysis provides such an invariance. The resulting features are proportional to the local symmetry in a way that does not depend on the image contrast. As such, the resulting edges are not biased by the temperature difference between them and the background. Because these features do not depend on the contrast or the object temperature the resulting magnitude is also invariant to the temperature of the sensor.

Fig. 3.8 is an example comparing the resulting magnitude image of points with high phase congruency and the gradient of the image, both applied to a far infrared image. The most prominent edge in fig. 3.8c is that between the buildings and the sky. Intensity gradients also depend on magnification, making it difficult to identify small objects. Local normalization is applied in Fig. 3.8d, where there can be appreciated some information loss, compared with the results of phase congruency, in figure 3.8b, where symmetric areas have the same importance, despite of their contrast.

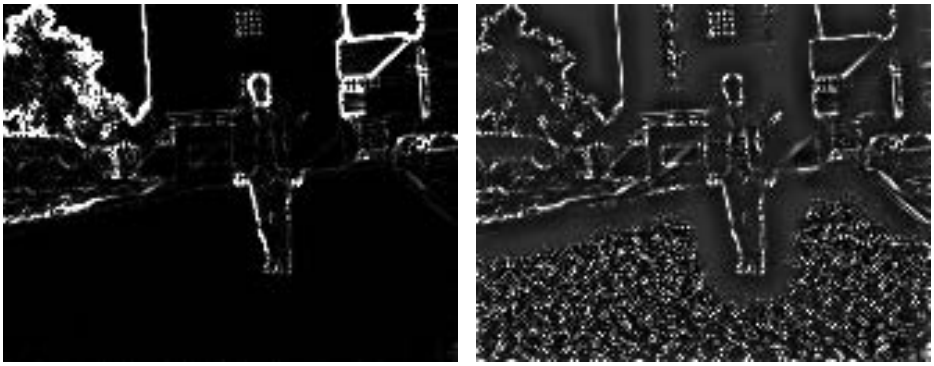
In this chapter, a new descriptor for pedestrian classification in FIR images is introduced, one that encodes the image as blocks of local histograms of phase congruency.

3.4.1. Phase Congruency

Phase congruency was first proposed as a biologically inspired vision model of mammalians in [150]. The main insight is the notion that points in a waveform representing lines or edges are those where the Fourier components are in phase with each other. A point i a signal with all of its Fourier components in phase will achieve a phase congruency score of one, while it none of them are in phase, the score would be zero. This measure is independent of the magnitude of the signal so it is invariant to changes in illumination and/or contrast. The sample shape of an object in two different images with different illumination will produce closely the same results. Phase congruency is used as a feature in some works. Kovesi proves in [116] that phase congruency can be use to extract line



(a) Original Image. (b) Phase Congruency.



(c) Gradient. (d) Locally Normalized Gradient.

Figure 3.8: Examples of phase congruency and gradient of an infrared image.

and point features from images. In his work, Kovesi extends the original definition to 2D signals. In [230] phase congruency is used as a feature vector the iris of the human eye. The discriminative function is the Euclidian distance of the phase congruency response.

In this section the properties of phase congruency are reviewed and its application in a pedestrian detector in FIR images is justified.

Features of high phase congruency values are those in which a wide range of their Fourier components are in phase. Those features are invariant to variations in image illumination, as will be illustrated in this section. In a one-dimensional signal those are points in the signal with a high slope or at peaks. Decomposition of smooth areas has its frequencies spread over a wider range, thus being its phase congruency score lower.

In order to calculate the phase congruency of a signal, a set of frequencies are extracted from it using a set of filters with the same amplitude spectrum, but shifted in the phase spectrum. Each of these filter extract the information at a narrow range of frequencies. Because the filters have to be used over a complex signal, they have to be complex too. In this case, a set of Gabor filters. An example of a one-dimensional Gabor filter is represented in Fig. 3.9. The even part of the filter is a sine curve and the odd part a cosine. Both signals are convoluted with a Gaussian of the same variance.

The real and imaginary parts of the one-dimensional Gabor filter are given by equations

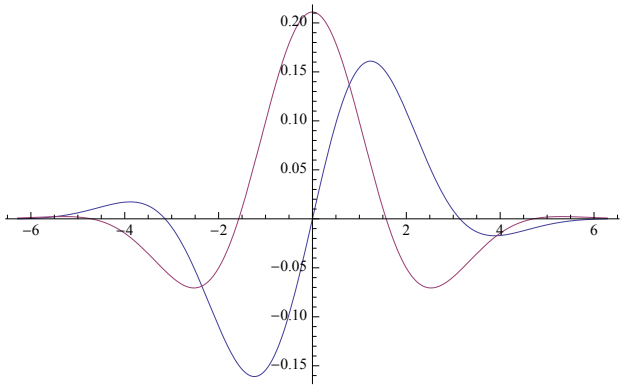


Figure 3.9: Real and imaginary parts of a Gabor filter in one dimension.

3.4 and 3.5.

$$O = \sin(2\pi w_0 x) \cdot \frac{1}{\sigma \cdot \sqrt{2\pi}} e^{\phi} \tag{3.4}$$

$$E = \cos(2\pi w_0 x) \cdot \frac{1}{\sigma \cdot \sqrt{2\pi}} e^{\phi} \tag{3.5}$$

Where w_o defines the center frequency, μ is the mean, σ is the standard deviation and ϕ is:

$$\phi = \frac{-(x-\mu)^2}{2\sigma^2} \tag{3.6}$$

The amplitude of the signal at the frequency of the filter is the square mean of the convolution of the signal with the odd and even filter (equation 3.7) ,

$$A_n(x) = \sqrt{(S(x) * O_n)^2 + (S(x) * E_n)^2} \tag{3.7}$$

where $S(x)$ is the signal at point x, E_n the even Gabor filter and O_n the odd one, n is the index of the frequency to be extracted.

The phase of the signal is given by equation 3.8.

$$\phi_n(x) = \arctan(S(x) * O_n, S(x) * E_n) \tag{3.8}$$

Because a convolution in the spatial domain is a product in frequency, the filters can be applied to the signal once it is transformed to its Fourier decomposition. After applying all the filters the weighted mean of phase for each point in the signal is calculated. This value maximizes equation 3.9 and determines the phase congruency score, as defined in [117],

$$PC(x) = \frac{W(x) \sum_{n=1}^{N} A_n(\cos(\phi_n(x) - \bar{\phi}) - |\sin(\phi_n(x) - \bar{\phi})|)}{\sum_n A_n} \tag{3.9}$$

Where N is the number of frequencies to be extracted and $(\phi_n - \bar{\phi})$ is the the deviation of each phase component from the mean, and $W(x)$ is a sigmoid function that penalizes low frequency spreads. In order to avoid noise, small local energy values are set to zero.

Fig. 3.10 shows the scale and contrast invariance properties of phase congruency. Figures 3.10a and 3.10b are two synthetic signals with equal shape but different scale. Specifically, the second signal has an amplitude 1000 times greater than the first one. Their phase congruency amplitude (Fig. 3.10f) is exactly the same. Figures 3.10g and 3.10h are two similarly shaped signals with different scale. The difference in shape simulates the response of the same edge with different levels of gain and contrast. Notice that the phase congruency amplitude (Fig. 3.10l) is almost exactly the same for both.

The analysis of images extends the signal processing to two dimensions. The one-dimensional filters described previously can be extended into two dimensions by simply applying a Gaussian spreading function across the filter perpendicular to its orientation. The resulting signal has exactly the same phase as the original as the transfer function of a Gaussian is also a Gaussian. As before, local information of frequencies is extracted by applying symmetric and antisymmetric Gabor filter to the Fourier transformed image.

This filter is formed by simply applying a Gaussian perpendicular to the sine and cosine parts of the one-dimensional signal. The Gaussian function doesn't affect the phase of the signal, only its amplitude. In order to minimize the spatial extent of the filter in the images log-Gabor filters are used in this work, as described in [76].

The main difference with one-dimensional filters is that each of these two-dimensional filters only extracts a fixed orientation of the image features. The solution is to convolve the image with a set of filters with different orientations for each frequency. Fig. 3.11 contains an example of rotated filters. The upper row represent the real part of five filters for the same frequency, each rotated $\theta = \{\frac{\pi}{6}, \frac{\pi}{3}, \frac{\pi}{2}, \frac{2\pi}{3}, \frac{5\pi}{6}, \pi\}$. The lower rows represent the imaginary part of the same filters.

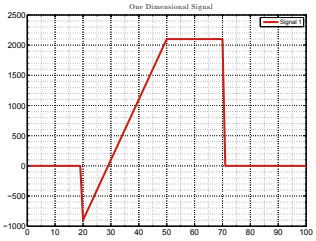
Each orientation contributes to the result at the given frequency proportionally to its energy. The result is a weighted sum that includes information at a wide range of possible orientations. As with the one-dimensional filter, the amplitude for each orientation is the square mean of the odd and even filtered images.

From the set of orientation images, phase congruency is calculated as indicated in equation 3.9.

3.4.2. Descriptor specifications

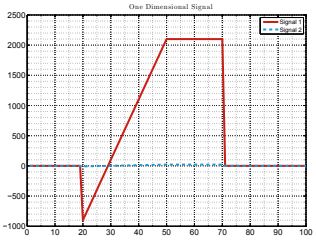
The pedestrian descriptor here presented follows the approach of encoding the shape of an object as a packed grid of SIFT-like blocks. The local information is extracted by dividing the image in sets of small spatial regions, called cells. Each cell contains a number of contiguous pixels of an image. For each cell an histogram is extracted. The combination of all of the histogram forms the feature vector of the image.

Fig. 3.12 represents the magnitude and orientation of the phase congruency of an image containing a pedestrian. In fig. 3.12c the orientation of each pixel, from 0 to 2π radians is

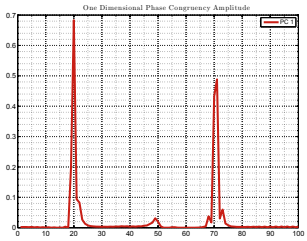


(a) Signal 1

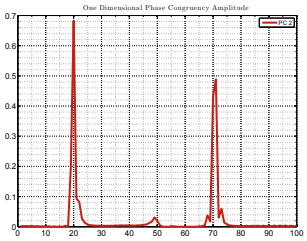
(b) Signal 2



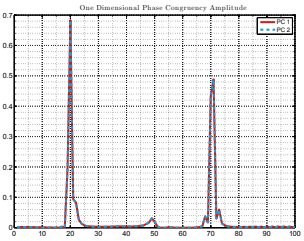
(c) Signal 1 and 2 to scale



(d) Phase congruency of signal 1 (a)



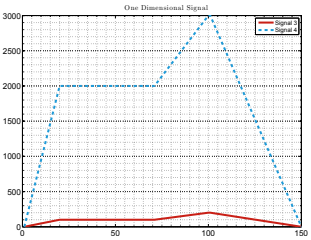
(e) Phase congruency of signal 2 (b)



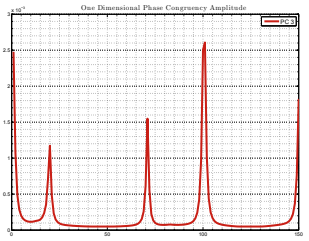
(f) PC of signals 1 and 2 to scale

(g) Signal 3

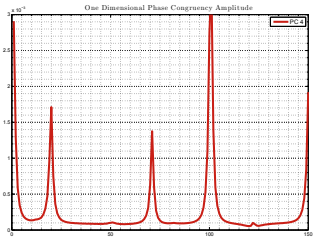
(h) Signal 4



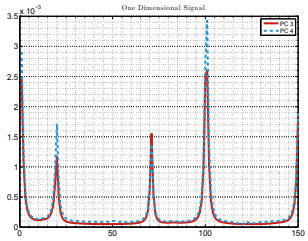
(i) Signal 3 and 4 to scale



(j) Phase congruency of signal 3 (g)



(k) Phase congruency of signal 4 (h)



(l) PC of signals 1 and 2 to scale

Figure 3.10: Scale and contrast invariance properties of phase congruency. Figures (a) and (b) are two synthetic signals with equal shape but different scale. Their phase congruency amplitude (h) is exactly the same. Figures (g) and (h) are two similarly shaped signals with different scale. The phase congruency amplitude (l) is almost exactly the same for both.

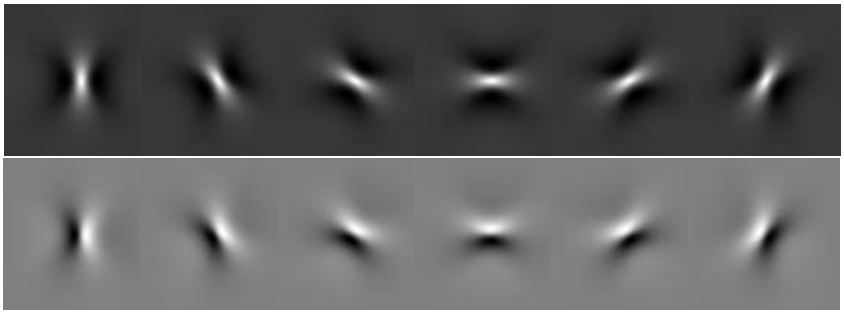
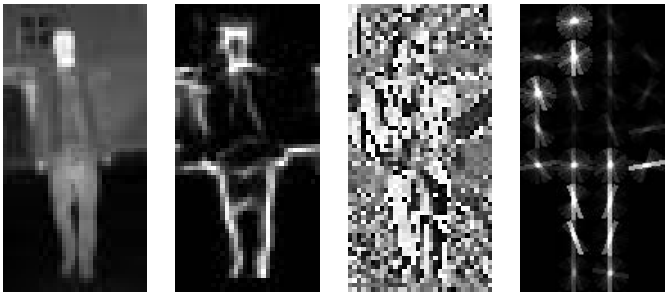


Figure 3.11: Four different orientations (θ) of filters for the same frequency λ. The filter rotates to captures image variations in different directions between $\theta = 0$ and $\theta = 2\pi$.

scaled from black (0) to white (1).



(a) (b) (c) (d)

Figure 3.12: (a) Original image. (b) Magnitude of phase congruency. (c) Gradient orientation. (d) Representation of the descriptors packed into grids.

The local feature representation allows more flexibility for small variations in the shape of changing objects, such as pedestrians. The fact that these features are also invariant to image contrast means that they can produce satisfactory results in a wider range of temperatures.

The histogram of each cell is the concatenation of the summatories of magnitudes for each orientation, as defined by equation 3.10.

$$h_c = \Big\|_{k=0}^{n} \sum_{x=0}^{cs} \sum_{y=0}^{cs} M(i,j) \cdot O_k(i,j) \tag{3.10}$$

where the concatenation operator $\|$ denotes concatenation of cells bins, O_k is the thresholded orientation image with pixels values equal to 1 if $O(i,j) = k$, zero otherwise, and M is the maximum covariance moment (eq. 3.11).

$$M = \frac{\sum_{i=1}^{n} PC_i^2 + \sqrt{\left(\sum_{i=1}^{n} PC_i^2 \cdot \sin(2\theta_i)\right)^2 + \left(\sum_{i=1}^{n} PC_i^2 \cdot \cos(2\theta_i)\right)^2}}{n} \tag{3.11}$$

Where PC_i is the phase congruency for the Gabor filter with rotation θ_i, and n is the

number of rotations.

The orientation image O is calculated as:

$$O = \arctan^{(*)}\left(\frac{\partial I}{\partial y}, \frac{\partial I}{\partial x}\right) \tag{3.12}$$

Where $\partial I/\partial y$ and $\partial I/\partial x$ are the vertical and horizontal gradients using $\begin{bmatrix}1 & -1\end{bmatrix}^T$ and $\begin{bmatrix}1 & -1\end{bmatrix}$ as filters and the function $\arctan^{(*)}(a, b)$ is equivalent to $\arctan\left(\frac{a}{b}\right)$, but preserving the orientation between $-\pi$ and π. The range of the orientation image O is discussed in section 3.4.3

The orientation image O is indexed into n values, corresponding to the division into equal angle ranges from 0 to π (eq. 3.13), where n is the number of bins in the orientation histogram, and $\{k \in \mathbb{Z} | 1, n\}$.

$$O(x, y) = i \forall \left((k-1) \cdot \frac{\pi}{n+1} < O(i, j) < k \cdot \frac{\pi}{n+1}\right) \tag{3.13}$$

The descriptor is the concatenation of the $h \times w$ cell histograms, $d = \|_{c=0}^{w \cdot h} h_c$.

3.4.3. Evaluation of Descriptor Parameters

The final descriptor depends on the parameters selected to create the phase congruency magnitude, the cell size, the number of bins of the histogram and its range. The parameters are selected based on the classification performance of a Support Vector Machine (SVM). Best parameters where selected by training one classifier for each combination of parameters within a range. Results are plotted as Detection Error Tradeoff (DET) curves, which plot the influence of a particular parameter while all the others are fixed to their standartd value. The classifiers are evaluated based on the miss rate at 10^{-4} False Positives Per Window (FPPW). The default parameters are:

- Scales = 4
- Orientations = 5
- Cell Size = $\{5 \times 5\}$
- Histogram bins = 9
- Crop Size = $\{64 \times 32\}$
- Radial Basis Function SVM kernel.

Number of scales Indicates the number of frequency segments to be extracted by the Gabor filters. For each scale a pair of Gabor filters with spread $\lambda_{min} \cdot \delta\lambda^n$ is created, where $n \in [1, N_s]$, N_s is the number of scales selected. The minimum wavelength is heuristically set to $\lambda_{min} = 2.0$ and the step to $\delta\lambda = 2.05$. Figure 3.13 shows performance of the classifiers for $N_s = \{2, 3, 4, 5, 6, 7, 8\}$ scales. Performance peaks at $N = 4$ scales. In order to prevent border artifacts, samples are cropped from the full-size images with a padding of, at least, $p = \lambda_{min} \cdot \delta\lambda^{N_s}$. For the standard number of scales used, $n = 4$, this padding is set to $p = 20$ pixels.

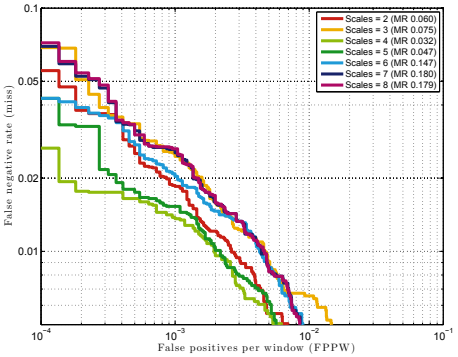


Figure 3.13: DET curves of the HOPE descriptor created with different numbers of Gabor scales. Legend states Miss Rate (MR) at 10^{-4} FPPW.

Number of orientations For each scale the filters are rotated to a number of angles to extract information at different rotations. Figure 3.14 shows performance of the classifiers for $N_o = \{2, 3, 4, 5, 6, 7, 8\}$ orientations. The peak performance is at 5 orientations. In that case the Gabor filters are rotated by $\theta = \{0, \frac{2\pi}{5}, \frac{4\pi}{5}, \frac{6\pi}{5}, \frac{8\pi}{5}\}$ radians. Interestingly, results are only slightly better than the worst performance at 2 orientations. Those two orientations are normal to each other and seem to comprise most of the information.

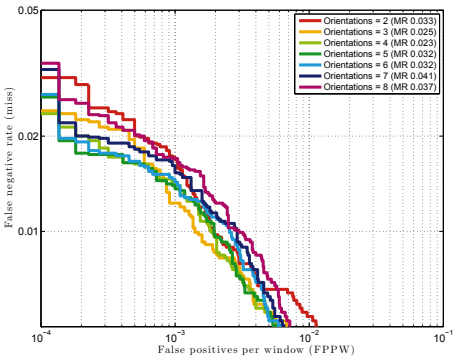


Figure 3.14: DET curves of the HOPE descriptor created with different numbers of Gabor orientations. Legend states Miss Rate (MR) at 10^{-4} FPPW.

Cell Size The samples are split into square, non-overlapping regions called *cells*. On each cell a histogram is calculated. In Fig. 3.15 the DET curves show the performance of the classifier for for different square cells sizes of $s = \{3, 4, 5, 6, 7, 8, 9, 10\}$ pixels. Small cells capture very fine details, however it is difficult to generalize that kind of information. Furthermore, small cells result in long feature vectors and longer training and evaluation times. Large cells encode the shape in a coarse manner, disregarding some information. We found that a good compromise is a cell size of $s = \{5 \times 5\}$ for $\{64 \times 32\}$ pedestrians.

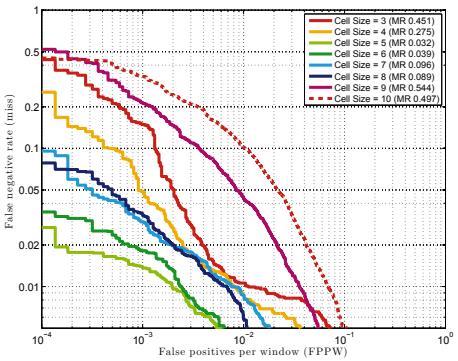


Figure 3.15: DET curves of cell sizes of the HOPE descriptor. Legend states Miss Rate (MR) at 10^{-4} FPPW.

Histogram orientations binning The cells are quantized into weighted histogram where the bin index is determined by the gradient orientations and weighted by the phase congruency magnitude. The minimum number of bins of the histogram of orientations for each cell has an important impact on classification performance, as shown in Fig. 3.16, where DET curves for $B = \{3, 6, 9, 12, 14\}$ histogram bins are plotted. Twelve bins over the default nine bins improves marginally the classification, just 0.39% at 10^{-4} FPPW, but this means a longer descriptor and more memory allocation.

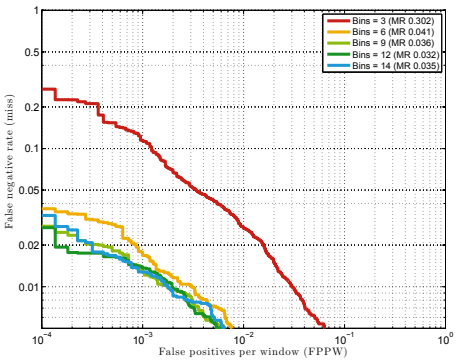
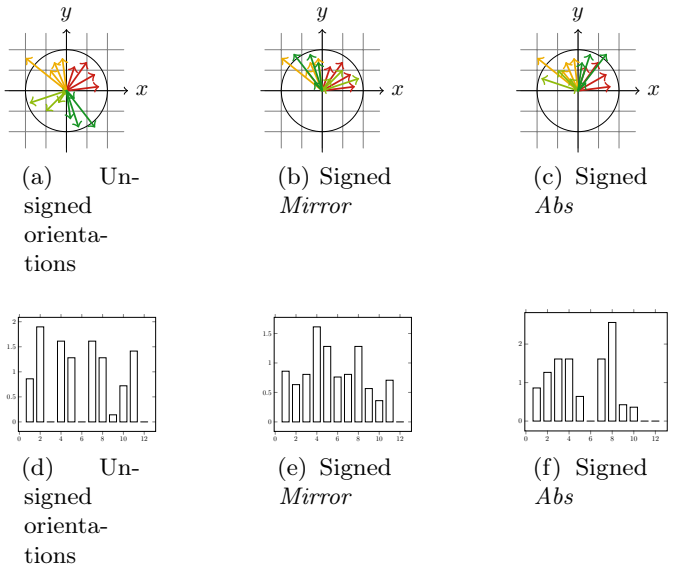


Figure 3.16: DET curves of the HOPE descriptor with different number of histogram bins . Legend states Miss Rate (MR) at 10^{-4} FPPW.

Range of histogram Orientation quantization splits the orientation range into B bins, between the minimum and maximum value. The selection of this range has an impact on classification performance. Three ranges of orientations are evaluated: signed orientation, in the $[-\pi, \pi]$ range $O_{360^\circ} = O$, unsigned *mirror* orientation: $O_{180^\circ} = O^- + \pi$, unsigned *absolute* orientation $O_{180^\circ} = |O|$. Figure 3.17 shows a representation of folding a 12 bin histogram into the three mentioned ranges. Notice that the histograms of those folds are quite different from one another. The unsigned orientation methods preserve the original orientation information. The mirror method wraps the orientation in the $[0, \pi]$ range by shifting the orientations with a negative value by π radians. The *abs* methods, wraps the orientations in the $[0, \pi]$ range by taking the absolute value of the orientations.

The results of the above mentioned methods are very similar to each other. Pedestrians in FIR images are darker or lighter that their background depending on the ambient temperature and the kind of clothing wearing. As images in the database were acquired in very different scenarios, binning the histogram with O_{360° only slightly improve classification, in this case.



(a) Unsigned orientations (b) Signed *Mirror* (c) Signed *Abs*

(d) Unsigned orientations (e) Signed *Mirror* (f) Signed *Abs*

Figure 3.17: Orientations of a 4×4 spatial cell. (a) Signed orientation, in the $[-\pi, \pi]$ range: $O_{360^\circ} = O$. (b) Unsigned orientation: $O_{180^\circ} = O^- + \pi$. (c) Orientation can also be wrapped in the $[0, \pi]$ range by $O_{180^\circ} = |O|$. Subfigures (d), (e), (f) represent their respective histograms by splitting the orientation range into 12 bins.

Normalization of blocks of adjacent cells The effect of normalization is shown in Fig. 3.18. The DET curves represent the classification performance for two normalization approaches: no normalization and L^2 normalization, as described in equation 3.14, where c if the histogram of a center cell and c_b is the concatenation of the histograms of four of its surrounding neighbors. In order to normalize each cell against all of their 8 neighbors, the L^2 normalization processed is performed on four blocks of 4×4 cells, resulting in four

normalization per cell. This process also results in a descriptor four times longer. It is worth remarking that cell normalization does not have a significant impact on performance. This is because the phase congruency map is in itself a normalized magnitude, as opposed to gradient. This property is specially useful as shown in section 3.5, where the Int-HOPE descriptor is presented.

$$c' = \frac{c}{\sqrt{\sum_{b=1}^{b=B} c_b}} \tag{3.14}$$

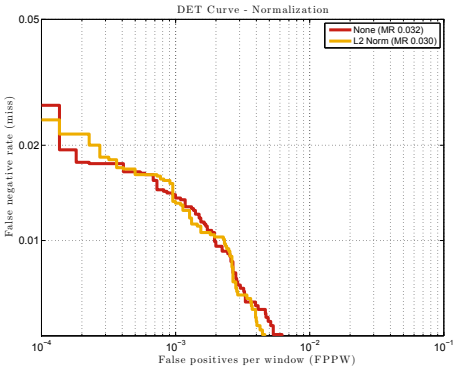


Figure 3.18: DET curves of different cell normalizations of the HOPE descriptor. Legend states Miss Rate (MR) at 10^{-4} FPPW.

3.4.4. Evaluation of the Classifier Parameters

In this section the classification method used to select the descriptor parameters is explained.

SVM classification calculates the boundary between two classes by searching the hyperplane that maximally separates the training set in a high-dimensional space. The boundary is defined by a subset of the training sample called the Support Vectors. The training set x_k is mapped in a high dimensional space defined by a function ϕ. The decision function in equation 4.22 is optimized so that $y(x)$ maximizes the distance between the nearest point (x_i) and the hyperplane.

$$y(x) = w^T \cdot \Phi(x) + b = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \sum_{l=1}^{m} w_{ijl} \Phi_l(c_{ij}) + b \tag{3.15}$$

Where w is normal to the hyperplane, b is the bias and $\frac{b}{||w||}$ is the perpendicular distance from the hyperplane to the origin. $\Phi(d_{ij}) \in \mathbb{R}^m$ is the corresponding histogram of gradients in the high-dimensional space of the descriptor, with m cell bins, at pixel d_{ij}. Φ is the kernel function that is used to project the samples. In this evaluation three different kernels have been tested: linear, quadratic and radial basis function. The sample is assigned to

one of the two classes by thresholding the decision function, where a sample with an score $y(x) > b$ is classified as a pedestrian, and as background otherwise.

Number of positive samples in the train dataset The training data set contains N samples $x\{k\} = (x_1, \cdots, x_N)$, manually classified and assigned a binary label $l = \{-1, 1\}$. Each one of this vector samples is a concatenation of the histograms of all the bins in the cropped image. The number N of samples can affect the performance of the classifier as too many input vector can over-fit the decision plane over the training set, becoming less effective with a wider representation on pedestrians. Figure 3.19 shows the impact on performance of incrementally adding more positive samples to the classifier.

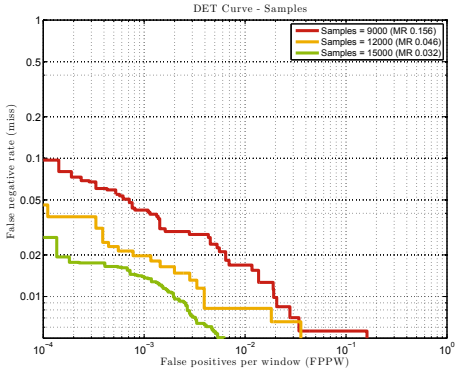


Figure 3.19: DET curves of of the HOPE descriptor trained with a different number of samples. Legend states Miss Rate (MR) at 10^{-4} FPPW.

SVM kernel The classifier performance will also rely on the initialization of the SVM. We first consider, as our base classifier, a one-norm, soft-margin support vector machine with a Gaussian Radial Basis (RBF) kernel function. Over the best sets of parameters, the best classifier overall classifier is searched for, by varying the kernel function.

The SVM kernels that have been tested with the database are linear, quadratic and RBF. The best results have been achieved using a RBF, though a simple quadratic kernel performs almost as good and needs less time to compute. The results are shown in Fig. 3.20.

Search of Hard Negatives Dataset images not containing pedestrians where densely scanned looking for false positives, known as hard negatives. In an iterative procedure, any detection classified as pedestrian by the SVM is added to the negative train dataset. A new classifier is then trained. This procedure is repeated several times until the classification performance degrades.

Fig. 3.21 shows that classification gets slightly better after two rounds of retrain with hard negatives. After that, some saturation can be observed.

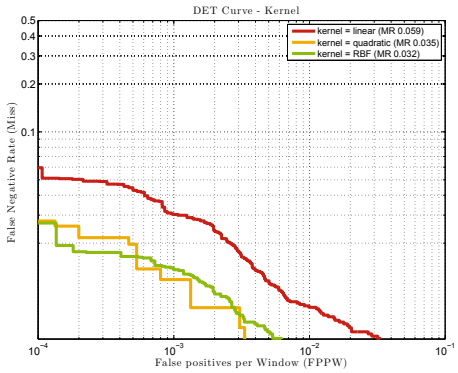


Figure 3.20: DET curves of of the HOPE descriptor trained with different SVM kernels. Legend states Miss Rate (MR) at 10^{-4} FPPW.

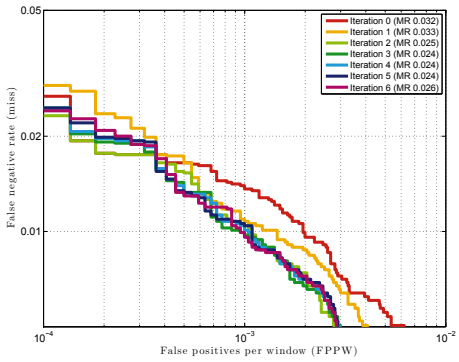


Figure 3.21: DET curve of classifiers after iteratively searching for hard negatives and retraining.

Number of Negatives Additionally, the impact on classification performance of naively varying the number of negative examples on the train set is assessed. Figure 3.22 shows that, for the HOPE SVM-Rbf classifier, the performance gets significantly better by increasing the number of train negatives. Figure 3.23 shows the relation between the number of the number of negatives on the training set and Miss Rate at 10^{-4} FPPW in the test set. That relation follows an exponential curve. Though it is expected that adding more negatives will results on even better classification performance, a very big number of samples used for training may lead to over-fitting. It should be also considered that using a large number of samples for training increase significantly the training time.

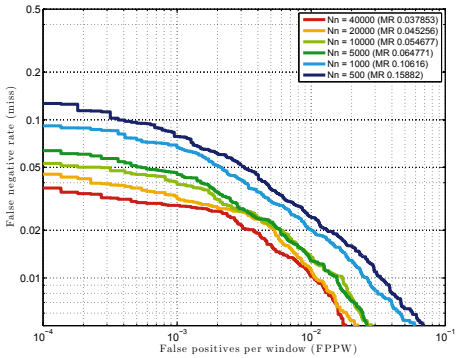


Figure 3.22: DET curve of the HOPE SVM-Rbf classifier trained with an increasing number of negatives. Legend states Miss Rate (MR) at 10^{-4} FPPW.

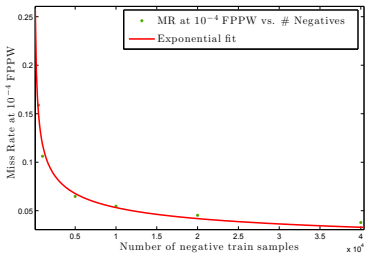


Figure 3.23: Curve fitting of the relation between number of negatives on the training set and Miss Rate at 10^{-4} FPPW in the test set. Notice that, from $n = 20000$ onwards, adding more samples has little impact on classification performance.

3.4.5. Other considerations

This section concludes the evaluation of the HOPE feature. In it, a multi-resolution feature approach, Pyramidal HOPE, is covered. Finally, the impact of noise is assessed.

Pyramidal HOPE [2] This descriptor can be extended by including multi-resolution features, as the Pyramid HOG descriptor does in [25]. The image is divided into increasingly finer spatial grids, where for each original cell $c_w \in \mathbb{R}^{w \times w}$, with $w = 5$, a new multi-resolution descriptor is calculated (equation 3.16).

$$c_m = \frac{c_{w1}}{w_1} \parallel \frac{c_{w2}}{w_2} \parallel \frac{c_{w_3}}{w_3} \tag{3.16}$$

The new descriptor is a concatenation of the original with the upper and lower scales. In this implementation $w_1 \in \mathbb{R}^{3\times3}$, $w_2 \in \mathbb{R}^{5\times5}$ and $w_3 \in \mathbb{R}^{7\times7}$. Fig. 3.24 is an example visualization of a three-scale descriptor around a keypoint.

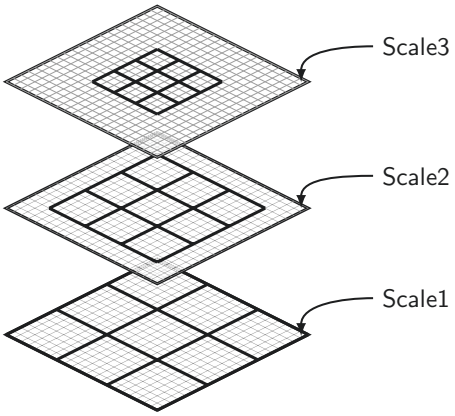


Figure 3.24: Descriptors at different scales around the same keypoint

Fig. 3.25 show the DET curves for single scale and multi resolution HOPE. An small improvement can be observed.

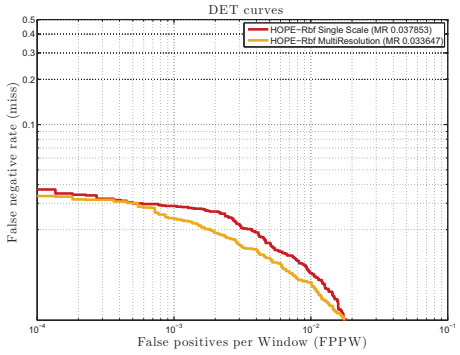
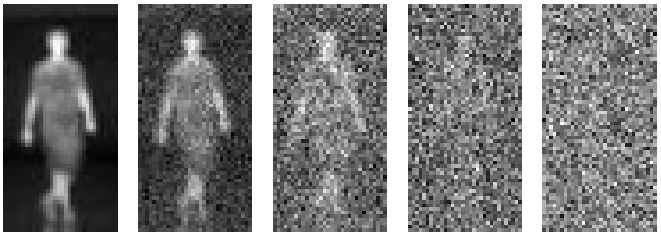


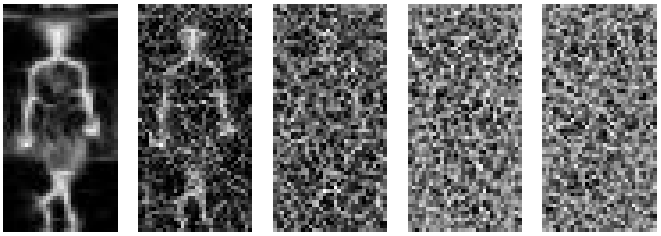
Figure 3.25: Comparison of the results with the single scale and the multi resolution HOPE descriptor. Legend states miss rate at 10^{-4} FPPW

[2] Unless otherwise stated the results presented in following sections of this work represent only single-scale features.

Impact of noise Phase congruency is known to be especially sensitive to noise level in the image. The impact of noise on the classification task is evaluated by adding a synthetic Gaussian noise to the dataset. The added noise follows a normal distribution $\mathcal{N}(\mu, \sigma)$, with mean $\mu = 0$ and variance $\sigma = \{10^{-9}, \cdots 10^{-3}\}$. Fig. 3.26 shows a noisy positive sample with increasing values of Gaussian noise variance added, along with its phase congruency magnitude response. Subsequently, classification results after applying a de-noising preprocessing are discussed.



(a) $\sigma = 10^{-9}$ (b) $\sigma = 10^{-6}$ (c) $\sigma = 10^{-5}$ (d) $\sigma = 10^{-4}$ (e) $\sigma = 10^{-3}$



(f) $\sigma = 10^{-9}$ (g) $\sigma = 10^{-6}$ (h) $\sigma = 10^{-5}$ (i) $\sigma = 10^{-4}$ (j) $\sigma = 10^{-3}$

Figure 3.26: First Row: Positive sample with Gaussian noise added. Second Row: Phase Congruency Magnitude Response.

The HOPE-Rbf classifier is trained again for each noise level added to the dataset. Its results are plotted in a DET curve, as shown in Fig. 3.27. Noise variance levels up to $\sigma = 10^{-6}$ seems to have hardly any impact on the classification task. For $\sigma = 10^{-5}$ the classifier degrades to an acceptable miss rate of 23.3% at 10^{-4} FPPW. From this point onward results deteriorates quickly, with a miss rate of 48.86% at 10^{-4} FPPW for $\sigma = 10^{-4}$.

A denoising preprocessing step improves classification performance in cases where the Gaussian noise variance is high. Two different approaches has been used to denoise the images: a median filter and a Wiener filter. A median filter, with a 3 pixel neighborhood, is applied to the noisy samples as shown in Fig. 3.28.

The noisy samples, after applying a Wiener filter [126] with a $\{5 \times 5\}$ pixel neighborhood, are shown in Fig. 3.29.

The Wiener filter estimates the local mean (equation 3.17) and variance (equation 3.18) around each pixel, where u is the horizontal coordinate of the pixels in the $\{5 \times 5\}$ pixel

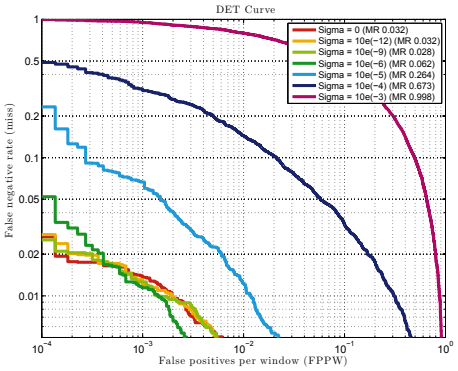
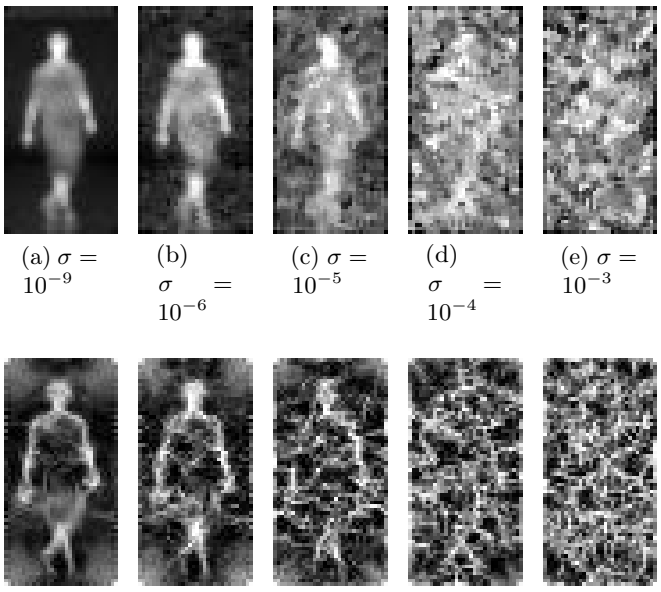


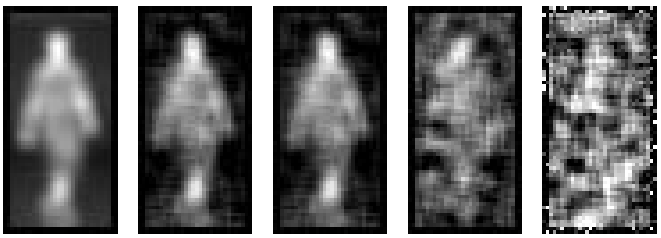
Figure 3.27: Classification DET curves for different amounts of syntethic Gaussian Noise. Classification of noisy samples achieves an acceptable hit rate for Gaussian noise with variance $\sigma \leq 10^{-5}$.



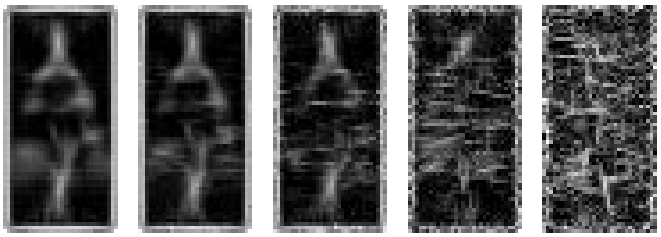
(a) $\sigma = 10^{-9}$ (b) $\sigma = 10^{-6}$ (c) $\sigma = 10^{-5}$ (d) $\sigma = 10^{-4}$ (e) $\sigma = 10^{-3}$

(f) $\sigma = 10^{-9}$ (g) $\sigma = 10^{-6}$ (h) $\sigma = 10^{-5}$ (i) $\sigma = 10^{-4}$ (j) $\sigma = 10^{-3}$

Figure 3.28: First Row: Noisy samples reconstructed with a Median filter. Second Row: Phase Congruency Magnitude Response.



(a) $\sigma = 10^{-9}$ (b) $\sigma = 10^{-6}$ (c) $\sigma = 10^{-5}$ (d) $\sigma = 10^{-4}$ (e) $\sigma = 10^{-3}$



(f) $\sigma = 10^{-9}$ (g) $\sigma = 10^{-6}$ (h) $\sigma = 10^{-5}$ (i) $\sigma = 10^{-4}$ (j) $\sigma = 10^{-3}$

Figure 3.29: First Row: Noisy samples reconstructed with a Wiener filter. Second Row: Phase Congruency Magnitude Response

neighborhood, and v the vertical coordinate.

$$\mu = \frac{1}{N \cdot M} \cdot \sum_{u=1}^{u=5} \sum_{v=1}^{v=5} \alpha(u, v) \tag{3.17}$$

$$\sigma^2 = \frac{1}{N \cdot M} \cdot \sum_{u=1}^{u=5} \sum_{v=1}^{v=5} \alpha^2(u, v) - \mu^2 \tag{3.18}$$

The Wiener filter is expressed as:

$$b(n_1, n_2) = \mu + \frac{\sigma^2 - \nu^2}{\sigma^2}(\alpha(u, v) - \mu) \tag{3.19}$$

Where μ^2 is the average noise variance of all pixels in the image, and $\{N, M\}$ is the local neighborhood of each pixel.

The results after applying the Wiener filter (Fig. 3.30) and the median filter (Fig. 3.31) suggest that the classification task has a generally better performance after denoising with the median filter for samples with a high noise variance. Classification seems to degrade slightly for low-noise samples, specially for the Wiener filter. For extreme values of noise, the preprocessing step has no effect. In these cases, a human expert achieves no better results.

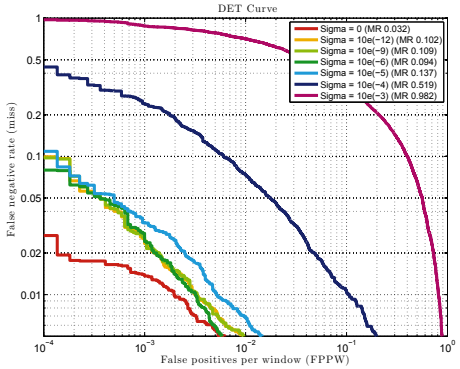


Figure 3.30: Classification DET curve for the database denoised with a Wiener filter.

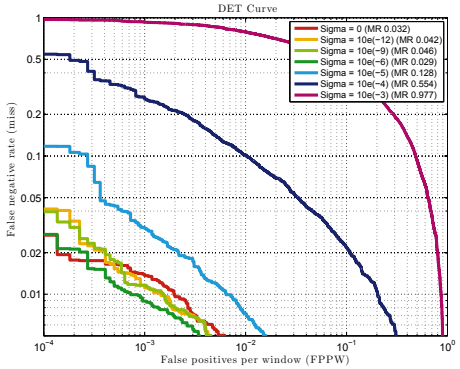


Figure 3.31: Classification DET curve for the database denoised with a Median filter.

3.5. Integral Features

There are a number of issues concerning the representation of an object as grid of local features. In the first place, not all of those features have the same importance. Some of them do not add any relevant information. However, a dense feature representation requires that all of them have to computed. This means that samples take the same time to compute, notwithstanding the fact that some could be easily discarded based on some key local features. The second point is that the descriptor may benefit of complementary information contained in the original image. This is solved by creating multi-feature descriptors. Finally, an efficient implementation is necessary to achieve high frame-rates, a required condition in ADAS.

This section addresses all those three issues. In it, a new descriptor is proposed, which uses different kinds of information in a multi-feature manner. All the features can be efficiently computed using the *integral image* paradigm [203]. The classification is sped up by using a *forest of decisions* approach. This classifier will be denoted as *Int-HOPE*.

Integral Features Every pixel in an integral image is the summation of the pixels above and to the left of it in the original image. This allows to rapidly calculate summations over square regions of the image. The integral image approach has been widely used for the purpose of feature calculations. In the original paper, Viola and Jones use the integral image to compute sums of small regions of the image, which are then compared with a set of Haar-like filters. A method for computing local histograms, based on an integral image, is discussed in [172].

The Int-HOPE descriptor can also be computed using integral images, because it is based on a grid of local histograms, that do not need to be normalized. Descriptor relying on a normalization step, as HOG does, need to compute the local feature and its neighbor before applying the normalization, defeating the purpose of using an integral image.

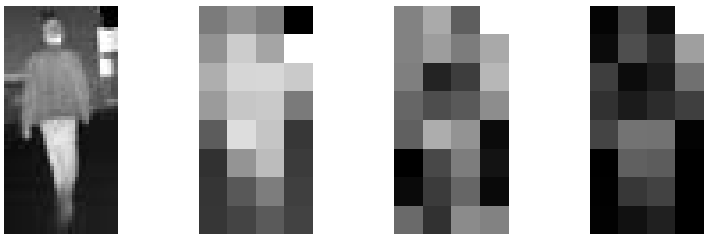
Feature Combination It is well stablished in the literature that combining features of complementary information may lead to a more robust classifier. Dollar et al. present an evaluation of the performance of combining several sources of information that can benefit from the integral image approach in [59]. Among them, they propose using integral histograms of not-normalized gradients.

The features subject to evaluation in this section are the following. The gray-level of an image contains all information contained in it. The raw information is hard to generalize, however, in combination with other features it may lead to a better classification performance. Figure 3.32a shows a sample of the positive train dataset. After rendering its integral image, a sampled-down version of it is computed, as shown in Fig. 3.32b. The representation of the features for the example image is shown in Fig, 3.32c, where each pixel of the resampled image is a feature. The dimensions of the *gray channel* is d/s, where $d = (w, h)$ is the dimension of the original image, and s is the size of the cell. The second

feature is the phase congruency magnitude as calculated in equation 3.11. The third feature is the gradient G of the image (equation 3.20).

$$G = \sqrt{\left(\frac{\partial I}{\partial y}\right)^2 + \left(\frac{\partial I}{\partial x}\right)^2} \tag{3.20}$$

The two remainder sets of features are the histograms of oriented phase energy (Fig. 3.33) and the histogram of oriented gradients (Fig. 3.34). In both cases, each bin of the histograms is a feature.



(a) Original (b) Gray (c) PC (d) Gradient

Figure 3.32: Descriptor computed using the integral image.

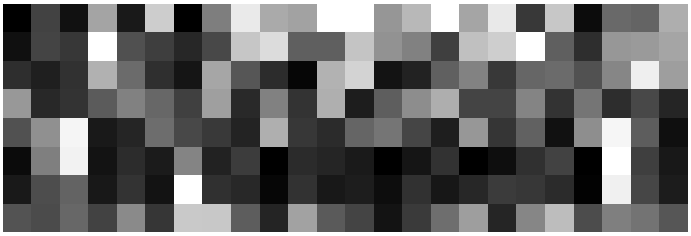


Figure 3.33: HOPE features. Each channel is a bin of the histograms

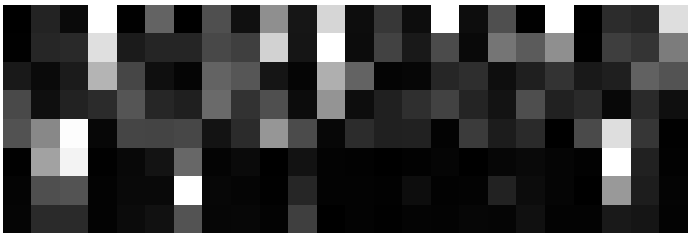


Figure 3.34: Histograms of gradient magnitude. Each channel is a bin of the histograms.

Random Forests The concept of Random Forests (RF), as introduced in [28], propose constructing a classifier from multiple decision trees. Each decision tree is a *weak learner*, that exploit a random set of samples from the training set. The number of samples is N, the same as the training set has, but selected with replacement, that is, a percentage of

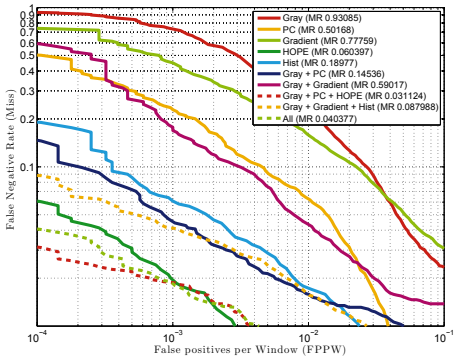
them are not selected and others are selected multiple times. At the root of each tree, a small set of $m < N$ features are randomly selected. The one that provides the best split is selected. This procedure is repeated in successive nodes, until the maximum tree depth is reached. While testing, a decision tree outputs the class of the samples, in this case, pedestrian or background.

A random tree tends to overfit data to the training set. A collection of trees can better generalize the data by using a *majority voting* strategy. A set of trees are trained using the same procedure as the original one. Their results are combined to form a *strong learner* by summing the outputs of each individual tree. The class that is awarded with more votes is selected. The confidence of this output is the percentage of votes that the class has got.

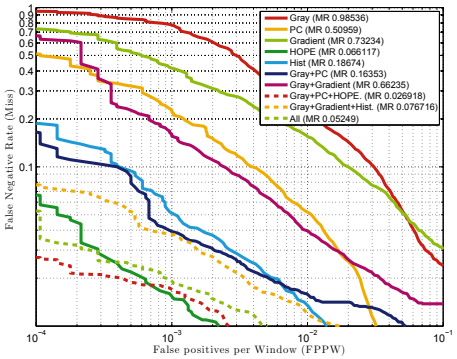
The Random Forest Classifier used in this section is trained with 200 trees, each of which uses a maximum of 425 samples. The number of features used at the node split decision level varies depending on the length of the feature vector. In general, that number is $\sqrt{n_f}$, where n_f is the length of the feature vector.

3.5.1. Square, non-overlapping features

The evaluation of several combinations of the cited sets of features is presented in this subsection. Ten features, or combinations of features, are tested using a Random Forest Classifier and an AdaBoost Classifier. Results are plotted as DET curves in Figs. 3.35 and 3.35.



(a) Six orientation bins, square cells (8 pixels)

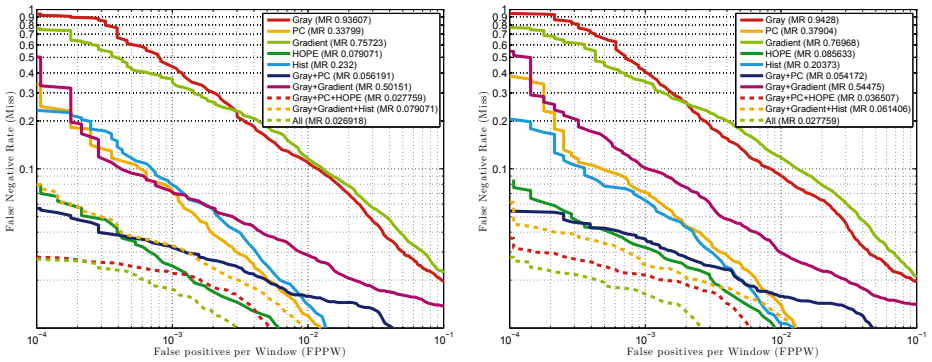


(b) Nine orientation bins, square cells (8 pixels)

Figure 3.35: Integral Channels (Random Forest Classifier). Legend states Miss Rate (MR) at 10^{-4} FPPW.

The sets of features have been:

- Gray
- PC: Phase Congruency Magnitude
- Gradient
- HOPE: Histograms of Oriented Phase Energy
- Hist: Histograms of Orientation
- Gray+PC
- Gray+Gradient
- Gray+PC+HOPE: Int-HOPE
- Gray+Gradient+Hist
- All: Gray+PC+Gradient+HOPE+Hist



(a) Six orientation bins, square cells (4 pixels)

(b) Nine orientation bins, square cells (4 pixels)

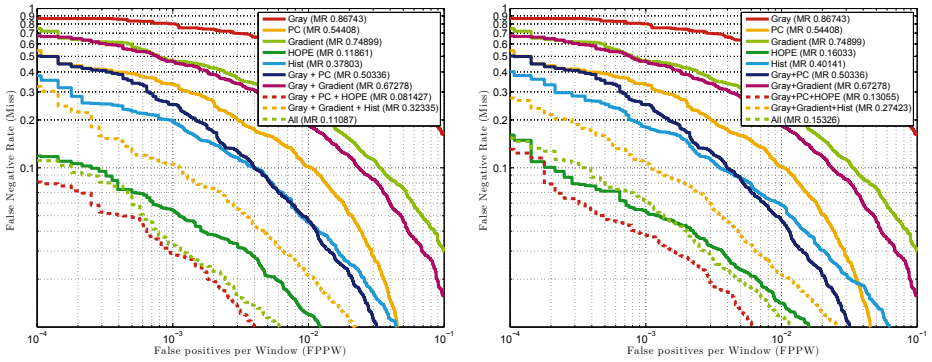
Figure 3.36: Integral Channels (Random Forest Classifier). Legend states Miss Rate (MR) at 10^{-4} FPPW.

Feature Performance As expected, the worst performing feature has been the gray level, with a miss rate of 93% at 10^{-4} FPPW (MR 0.93) for the 8×8 cells descriptor. Combining the PC feature with the gray level yields similar results as the Hist feature, with MR 0.14 and MR0.19 respectively, while the Gray+Gradient feature achieves MR 0.6, a performance similar to the PC alone (MR 0.5). Using only the gradient, the performance degrades, with a MR 0.78. Comparing the two approaches based on integral histograms, using 6 histogram bins, it is clear that HOPE features achieves better results than Hist, with MR 0.06 and MR 0.19 respectively. As expected, combining the gray level, PC and HOPE features produce a better classifier than the Gray+Gradient+Hist, as the individual features perform better. The Gray+PC+HOPE achieves MR 0.03, a similar result as the one obtained training the dense HOPE feature with an Rbf kernel SVM. The Gray+Gradient+Hist achieves MR 0.087. Interestingly, combining all the features degrades miss rate at 10^{-4} FPPW by 1%, when compared with the Gray+PC+HOPE feature. The latter descriptor will be refereed to from now on as Int-HOPE.

Cell size and orientation binning The election of the cell size have a significant impact for some of the descriptors. The results of the classification using 8×8 cells are plotted in Fig. 3.35. When compared with the results of the 4×4 descriptors (Fig. 3.36), it is clear that reducing the cell size benefits features based on gray-level, gradient and phase congruency. In the latter case, the miss rate at 10^{-4} FPPW goes from MR 0.55 to MR 0.33. A less important improvement is observed in the case of gradient. For small cells, the gray-level feature achieves similar results as the gradient feature. However, features based on histograms degrade slightly for small cells. Nevertheless, the combination of all features get significantly better for 4×4 cells, going from MR 0.04 to MR 0.026.

Packing the histograms into 9 bins instead of 6 does not have a relevant impact on performance. Features based on histograms have similar results with those histogram sizes, both for eight and four pixel cells. Other features behave the same way in both cases, though the results differ slightly due to the random sampling of features.

Adaboost The proposed collection of features have been also tested using an AdaBoost approach. Real AdaBoost is used, as described in [179], with a maximum of 50 iterations. Figure 3.37 shows the DET curves of the classifiers, using AdaBoost. The results seem to correlate with the ones obtained using Random Forests, but results are not as good. For the best (Gray + PC + HOPE) the miss rate of the AdaBoost classifier at 10^{-4} FPPW is a 5% higher.



(a) Six orientation bins, square cells (8 pixels)

(b) Nine orientation bins, square cells (8 pixels)

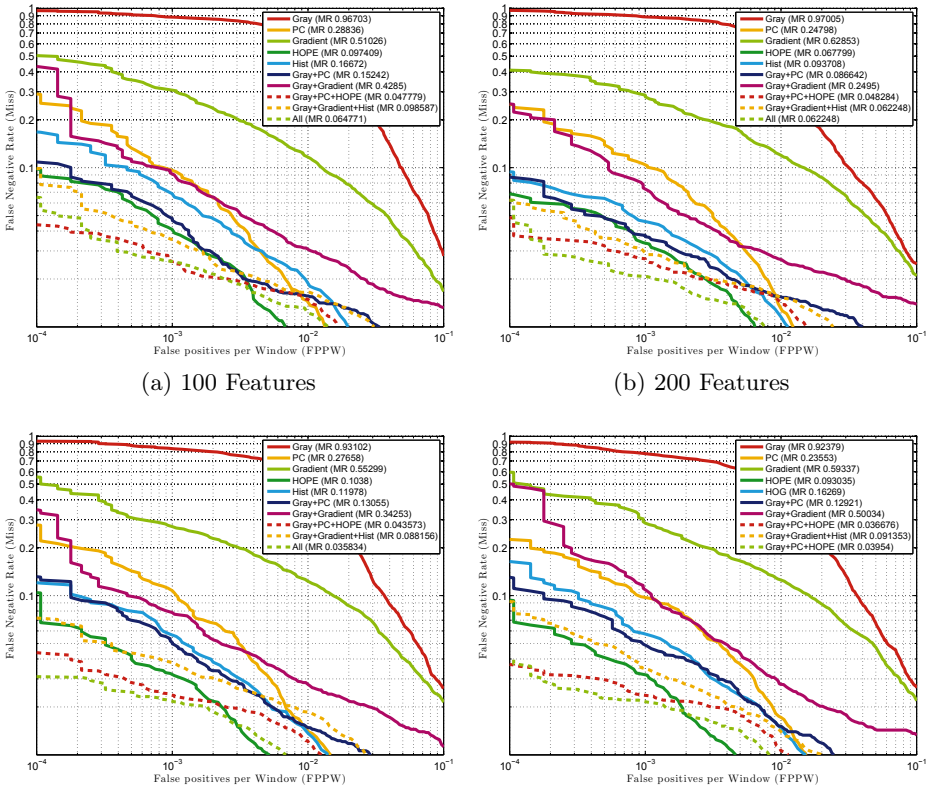
Figure 3.37: Integral Channels (Adaboost). Legend states Miss Rate (MR) at 10^{-4} FPPW.

3.5.2. Rectangular, overlapping features

Features based on Integral Images allow to quickly compute a response at any rectangular location of an image, with any size, with the same computational demands as computing a square feature. Descriptors based on grids of non-overlapping cells define accurately the shape of the object. Each one of those cells include information of a specific part of the object. However, the choice of these parts is based on intuition and not on the certainty that they are the most representative ones. Other parts of the samples may contain more relevant information, which is missed in the feature-grid approach. In this section this notion is put to test by randomly selecting a large number of rectangular features from the samples. At each decision node of the decision trees, the most representative features are selected by cross-validating them with the subset of unused samples.

Results are plotted in Fig. 3.38, where each subfigure represent the results with a different number of features per channel. The set of features have width and height randomly chosen between 5 and 10 pixels. The number of histogram bins is set to 6. The range of feature sizes allows to capture both fine details and coarse shape. The results are congruent with the ones in the previous section, as features not based on histograms have similar performance as in the fine-grid experiments (Fig. 3.36). However, features based on histograms do not get in par with the grid-based approach until the number of samples per channel is 1000. For the Int-HOPE descriptor, this means a bag of 8000 features.

The importance of each feature, as selected by the decision trees, is represented in Fig. 3.39. Lighter areas correlate with regions of greater importance. For plotting those figures,



(a) 100 Features (b) 200 Features

(c) 500 Features (d) 1000 Features

Figure 3.38: DET curves of the random rectangular features. For these experiments the number of histogram bins is set to 6. Legend represent Miss Rate (MR) at 10^{-4} FPPW.

each time a feature is used by any node of the forest of decision that corresponding area is incremented by one gray-level. For visualization purposes, the range of the figures have been rescaled between the minimum number of hits (darkest areas) and the value of the most used feature (lightest area). From inspecting these figures, it seems that the most relevant single feature is the phase congruency, while the least used seem to be the gradient. It should be noted that the HOPE and Hist features are made up from 6 orientation bins. As such, the importance of the overall descriptor is the sum of their bin images.

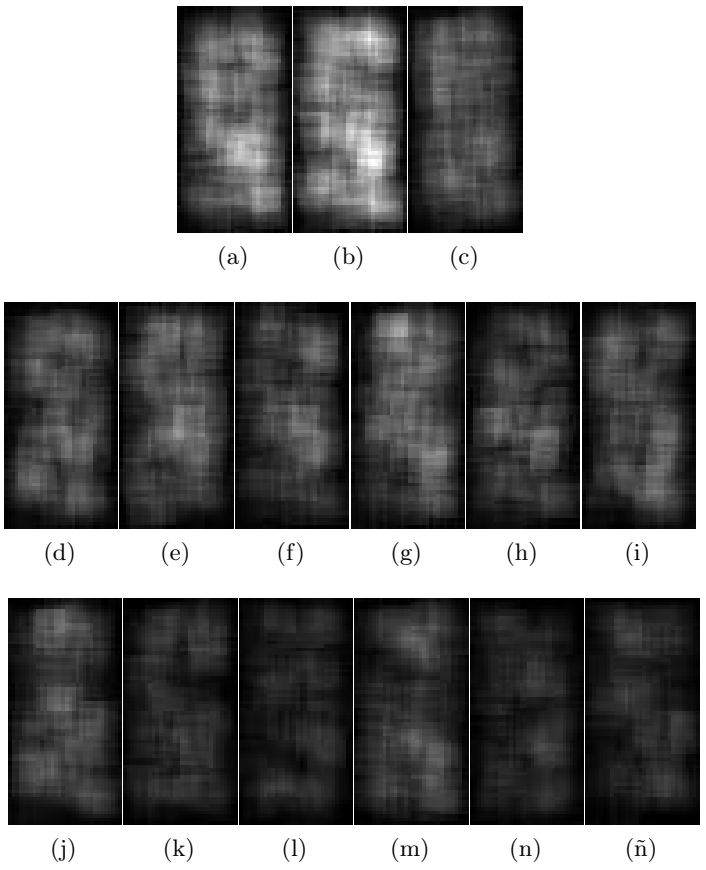


Figure 3.39: Representation of feature importance in the random rectangular descriptor. Each pixels is scaled between 0 (least important) and 1 (most important). The subfigures represent the following feature vector: a) Gray-level; b) PC; c) Gradient; d)-i) HOPE; j)-o) Hist

3.6. Comparative Results

In this section, other descriptors are tested with the LSI Far Infrared Pedestrian Dataset, in order to compare their results with the descriptors previously presented.

3.6.1. Classification Methods

Five kinds of classification methods have been used: SVM, Naïve Bayes Classifier (NBC), Quadratic Discriminant Analysis (QDA), Neural Networks (NN) and Adaboost. The parameters selected for the different classifiers are discussed in the sequel.

Support Vector Machines Concerning SVM [34], two different kernels were used for benchmarking: a linear classifier, hereafter called SVM-Lin, and a radial basis function (RBF) kernel, designated by SVM-Rbf. In this implementation the radial Gaussian function kernel $K(x, y) = e^{-\gamma\|x-y\|^2}$ has a scale parameter $\gamma = 1$.

Naïve Bayes Classifier NBC [106] is designed for use when features are independent of one another within each class, but it appears to work well in practice in other circumstances. Naive Bayes classification is based on estimating the conditional probability of the feature vector given the class.

Discriminant Analysis Linear Discriminant Analysis [140] is used as a linear classification model in terms of dimensionality reduction. Considering a two class separation problem the D-dimensional input vector x can be projected down to one dimension as $y = \mathrm{w}^T\mathrm{x}$, where w is the components weight vector. Selecting appropriate weights, the feature space can be projected over the dimension that maximally separates both classes. Over this projection a threshold w_o is selected, where values $y \leq -w_0$ are classified as pedestrians, whereas values $y > -w_0$ are classified as background. In this implementation the coefficient matrix of the boundary equation is quadratic thus, the discriminant analysis takes a quadratic form, designated QDA. The pedestrian and background classes are assumed to be normally distributed. The multivariate normal densities are fitted with covariance estimates stratified by group.

Neural Network A NN pattern recognition scheme [102] is used with a two-layer feed-forward network, with ten hidden and one output sigmoid neurons (Fig. 3.40). The network is trained with scaled conjugate gradient backpropagation. The overall network function follows equation 3.21, where σ is the sigmoid function, rnk is the output ranking, w are the feature weights, N is the number of inputs and M is the maximum number of linear combinations of the N inputs.

$$rnk(x, w) = \sigma\left(\sum_{j=1}^{M} w_j^{(2)} h\left(\sum_{i=1}^{N} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_0^{(2)}\right) \tag{3.21}$$

Adaboost Real AdaBoost is used, as described in [179]. The key idea is that the combined response of a set of weak classifiers can build a strong one, improving the performance that a complex classifier alone would have. Iteratively, Adaboost selects a threshold that

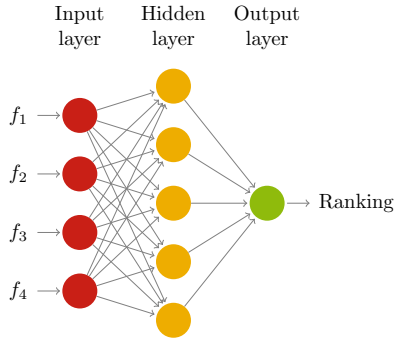


Figure 3.40: Two layered neural network with ten hidden and one output sigmoid neurons. The inputs $f_1 \cdots f_n$ are the histogram bins.

best separates each feature set x_i in one of the classes y_i, applying a higher weight to misclassified samples. In this implementation the maximum number of iterations is set to 50. The final ranking of each feature vector is $\text{rnk} = \sum_{i=1}^{N} x_i(f_i)$. In the case of HOG and HOPE, each bin in the orientation histograms is treated as a weak feature.

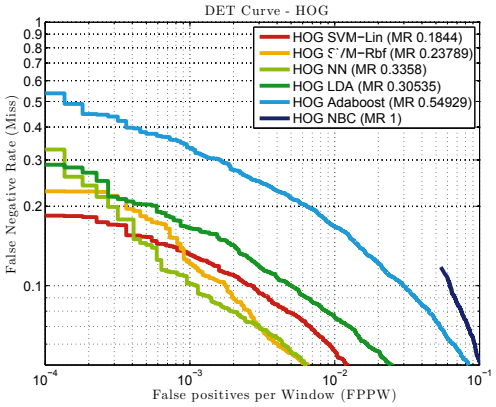
3.6.2. Features

The classification is treated as a supervised pattern recognition problem. Given a set of samples manually labelled by an expert $D = \{\{x_1, y_1\}, \cdots, \{x_n, y_n\}\}$, where $x_i \in R^d$ is a feature vector and $y_i \in \{\pm 1\}$ is a binary label, which establishes its belonging to one of the two classes $\mathcal{C}_1$ and $\mathcal{C}_2$, a decision function is optimized.

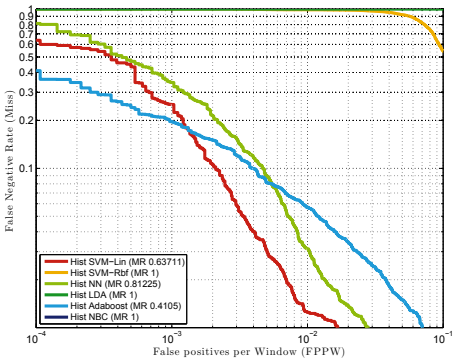
In this section the feature selection is discussed, along with implementation details.

Histograms of Oriented Gradients The Histogram of oriented gradients has been tested, using 5×5 pixel cells. The magnitude of the gradient is linearly interpolated to the two closest orientations. Additionally, each point is bi-linearly interpolated to the neighboring cells. Each cell is then normalized four times with the surrounding cells. Within each cell a 9 bin histogram of orientation between 0 and π radians is calculated. Using unsigned histograms do not improve performance, as shown in section 4.3.2. The resulting descriptor, which closely resembles the Dalal and Triggs version, will be denoted as HOG. Fig. 3.41 represents the DET curves of the described HOG descriptor trained with the LSI far infrared pedestrian database. The DET curves for unnormalized histograms of orientations are plotted in Fig. 3.41b.

HOPE The HOPE descriptor encodes a grid of local oriented histograms extracted from the phase congruency of the images, which is computed from a bank of Gabor filters. This descriptor do not use spatial interpolation or cell normalization, as the HOG descriptor does. Results of the HOPE descriptor trained with different classification methods are plotted in Fig. 3.42.



(a)



(b)

Figure 3.41: DET curves of the HOG descriptor. a) Normalized histograms; b) Unnormalized histograms. Legend states Miss Rate (MR) at 10^{-4} FPPW.

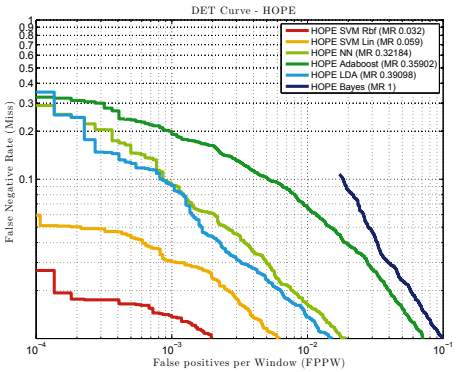


Figure 3.42: DET curves of the HOPE descriptor. Legend states Miss Rate (MR) at 10^{-4} FPPW.

Local Binary Patterns LBP, as introduced in [159], represent the image as a similarity vector of each pixel with their surroundings. This descriptor encodes information as a binary number, where for each pixel, the neighbors with a gray value higher or equal contribute with a one in their position in the binary number, otherwise a zero. Each sample is divided in $3x3$ pixel non-overlapping cells.

The value of the LBP descriptor of a pixel (x_c, y_c), as represented in Fig. 3.43 is given by equation 3.22. Results of the LBP descriptor trained with different classification methods are plotted in Fig. 3.44.

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) \cdot 2^p \quad , s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.22}$$

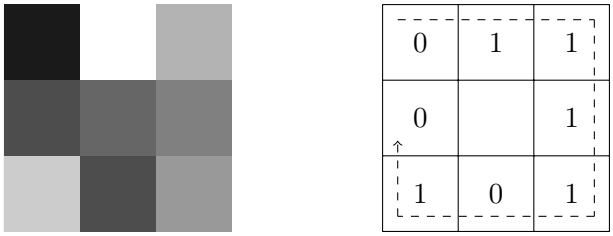


Figure 3.43: Example of an LBP descriptor

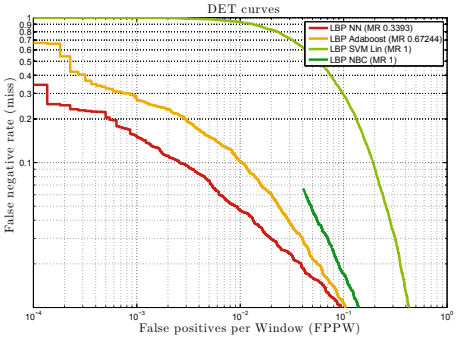
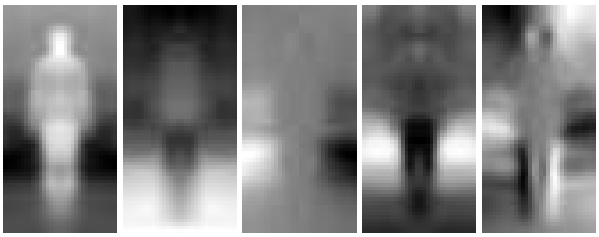


Figure 3.44: DET curves of the LBP descriptor. Legend states Miss Rate (MR) at 10^{-4} FPPW.

PCA We treat PCA [140] eigenvectors as a grey-level feature vector. The initial motivation for applying this approach is that PCA tends to disregard small details at high frequency, as seen in Fig. 3.45, while FIR images usually have poor levels of detail, as they present softness due to motion blur, especially at low resolutions. We retain the 30 most significant eigenvectors, that is, those with the largest eigenvalues. Figure 3.46 shows the DET curves of the PCA descriptor for several classification methods.

Feature Concatenation Descriptor fusion is explored as feature vector concatenation [127]. The resulting feature vector concatenating two descriptors $D_1 \in \mathbb{R}^m$ and $D_2 \in \mathbb{R}^n$ is



(a) Eig(1) (b) Eig(2) (c) Eig(3) (d) Eig(4) (e) Eig(5)

Figure 3.45: First 5 eigenpedestrians

a new higher dimension feature vector $D = D_1 \,||\, D_2 \in \mathbb{R}^{m+n}$, which holds different kinds of complementary information, which can improve the overall performance. Dimensionality is kept low by removing linearly dependable features by means of a discriminant analysis.

3.6.3. Discussion of the Comparative Results

From the presented results, it can be observed that approaches based on local orientated histograms, such as HOG and HOPE, get better results than PCA or LBP. The best performing feature seems to be HOPE, with a miss rate of 0.06 at 10^{-4} false positives (FP) for the SVM-Lin classifier, followed by HOG with a miss rate of 0.18 at 10^{-4} FP. With an RBF kernel performance improves up to 0.03% miss rate at 10^{-4} FP in the case of HOPE .

Regarding feature combination, we have used an SVM-Lin to assess the impact of the features in the classification performance. Both LBP and PCA, does not improve significantly classification when merged with other features. Combining the HOPE descriptor with HOG significantly increases detection rate. This means that both extract complementary information. This is a particularly interesting result, as there are many descriptors that have HOG as part of their feature vectors. If that is the case, adding this descriptor would increase the performance of the classifier, though this would depend on the particular problem at hand.

Concerning the classification methods, SVM-Rbf generally has the best performance followed by SVM-Lin. LDA classifier performs almost as good, or better than Linear SVM for the HOG and HOPE descriptors. The NBC showed the worst performance, except for LBP features.

3.6.4. Statistical Significance of the Results

Statistical significance is assessed with McNemar's approximate test [55]. It is used to compare two classifiers at a particular value of bias. To determine wether classifier ($\mathcal{C}_1$) is significantly better than ($\mathcal{C}_2$), the χ^2 statistic is is used (equation 3.23).

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \tag{3.23}$$

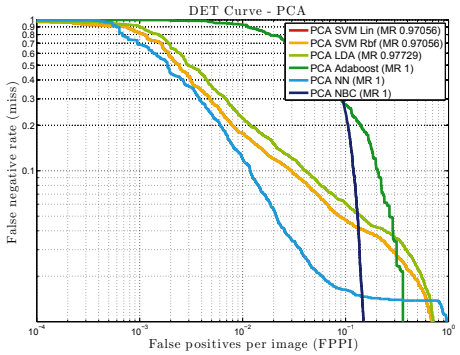


Figure 3.46: DET curves of the PCA descriptor. Legend states Miss Rate (MR) at 10^{-4} FPPW.

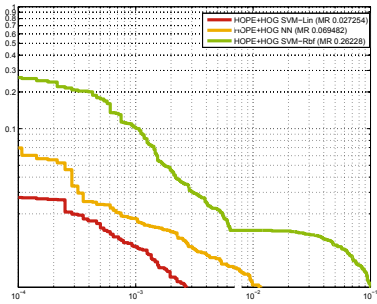


Figure 3.47: DET curves of the combination HOG and HOPE descriptors. Legend states Miss Rate (MR) at 10^{-4} FPPW.

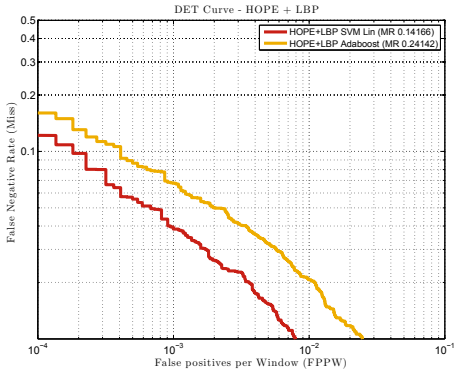


Figure 3.48: DET curves of the combination of LBP with HOG and HOPE descriptors. Legend states Miss Rate (MR) at 10^{-4} FPPW.

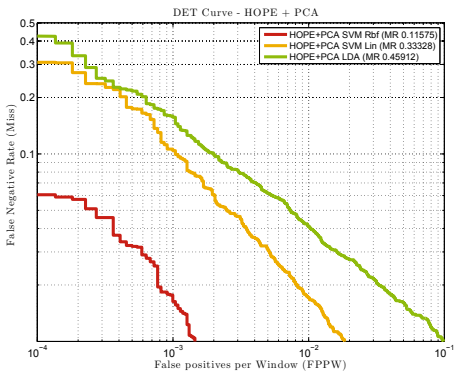


Figure 3.49: DET curves of the combination of PCA with HOPE descriptors. Legend states Miss Rate (MR) at 10^{-4} FPPW.

where n_{01} is a number of cases misclassified by $\mathcal{C}_1$ and classified correctly by $\mathcal{C}_2$, and n_{10} is a number of cases misclassified by $\mathcal{C}_2$ and classified correctly by $\mathcal{C}_1$. The null hypothesis H_0 states that the performance of both classifiers is the same. H_0 may be rejected if χ^2 falls below a probability of 5%, i.e. $\chi^2_{1,0.95} \geq 3.841459$. If that is the case, it can be assumed that one classifier performs significantly better than the other.

Table 3.2 contains χ^2 values for every pair of classifiers used. The bias of all classifiers has been $b = 0$, after rescaling, as it is the value that maximally separates both classes. From these results it can be concluded that the null hypothesis can be rejected for all classifier pairs.

Table 3.2: Results of the McNemar's approximate significance test for every pair of classifiers. The value expressed in the table's fields is χ^2, as stated in equation 3.23

	HOGLin	HOGRbf	HOPELin	HOPERbf	LBP	PCA
HOGLin	0	247.1	81.5	342.1	1852.9	1508.6
HOGRbf	247.1	0	51.6	11.5	2825.5	2443.7
HOPELin	81.5	51.6	0	108.0	2427.6	2056.1
HOPERbf	342.1	11.5	108.0	0	2980.8	2596.7
LBP	1852.9	2825.5	2427.6	2980.8	0	23.8
PCA	1508.6	2443.7	2056.1	2596.7	23.8	0

3.7. Conclusions and Discussion

In this chapter, a variation of the probabilistic template scheme for pedestrian classification in FIR images is presented. From a set of cropped samples containing images, and with information about the sensors temperature, a probabilistic template is created by averaging the thresholded samples. The main purpose of this method is to add invariance to ambient temperature to this scheme. Similar methods usually rely on image statistics to select a fixed threshold. In a non-refrigerated microbolometer the results degrade as the temperature rises, because the histogram of the image shifts with sensor temperature.

In this chapter a new descriptor for classification of pedestrians in far infrared images has been presented. This approach exploits information from low resolution, uncalibrated, non-refrigerated microbolometer sensors. The main application of the system is to be used on night-time images, though the performed tests prove that good performance can be achieved in a wide range of temperature and illumination conditions.

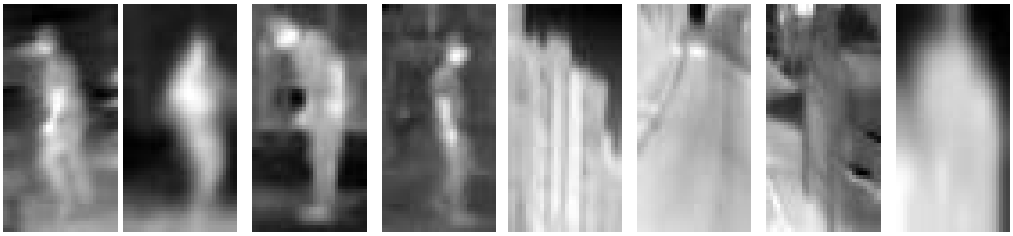
Several combinations of descriptor and classification methods have been tested in a new FIR dataset. By our best knowledge this is the first complete FIR based pedestrian classification and detection dataset publicly available for benchmarking, in the area of ITS. The classification scheme here presented has been tested as part of a detector, using an sliding window approach.

Though Phase Congruency is known to be vulnerable to noise, results suggest that the detector here presented can cope with fairly high levels of noise.

From the experimental results reported in the previous sections it can be concluded that histogram based features perform best than LBP or PCA features. In terms of classification methods, SVM achieved the best performance. The RBF kernel can significantly reduce misclassifications compared with a linear kernel, but is more computationally demanding. This is a critical factor in computer aided transportation applications.

The combination of the HOPE and HOG features into a single descriptor considerably increases performance. This may be useful in detectors that have histograms of gradients as part as a multi-feature classifier.

Finally, a qualitative inspection of misclassified samples suggests that ambient temperature has a determinant impact on performance. Sequences collected at a high environment temperature or under direct sun light present the most false positives, and also the highest miss rate. A qualitative evidence of this issue is shown in Fig.3.50 where we present examples of misclassified positives and negatives. Other sources of misclassification are motion blur, which in FIR images appear frequently, and pose variation. False positives appear mostly in negative examples with a high vertical symmetry.



(a) False negatives (b) False positives

Figure 3.50: Misclassified samples. (a) False negatives due to low resolution, motion blur and pose variation. (b) ExaFalse positives of areas with a high vertical symmetry.

This chapter also cover a new descriptor that exploits a majority voting scheme of randomly selected weak classifiers based on features computed with integral images. The resulting classifier, Int-HOPE has slightly better performance than HOPE, and the Random Forest Classifier used is quicker in training and testing than the SVM used for the holistic descriptor.

The results presented in this chapter suggests that FIR images are a very useful source of information for pedestrian classification and detection, having similar performance to that found in state of the art in VL images, with advantage in low visibility applications.

4

Detection

4.1. Introduction

Pedestrian detection is defined as finding the position of an *a priori* unknown number of pedestrian on a set of full images. Detectors usually are composed of two steps. The first one is the selection of regions of interest in the image, which can be silhouette-based or rectangular. Silhouette-based approaches attempt to segment pixels belonging to the object of interest. In practice, most detectors follow the bounding box approach, where the position of the pedestrian is defined as a rectangular region, that also contain part of the background. Methods that select rectangular regions often follow the sliding window approach, a dense search at many scales of the image.

The most straightforward approach when densely scanning an image is to create one model for the N scales at which the detector should search. A more common approach is to train just one model and resize the image N times, in an image pyramid. This may lead to different detection rates at each scale. When up-sampling an image to match the model size, small objects appear blurry and without detail. If the image is down-sampled, some its information is lost. The detector would then achieve its best performance if the model size encompasses both rich detail and blurry shapes. Variations of this two approaches have been proposed. In [58] suggest that features at nearby scales can be approximated. This leads to only a fraction of images (N/K) in the pyramid to be computed. Features computed from the pyramid images are then used to approximate the features at nearby scales. With less images to process, the detectors can be greatly accelerated. In [9] the inverse concept is suggested. Instead of resampling a subset of images and approximate features at nearby scales, they propose using a single image scale and training a set of N/K models. Each model is tuned to accept responses from K sizes.

The second step of a detection algorithm is the classification of those regions. Using an sliding window leads to one score for every position of the sliding window. A good detector would render higher scores for windows that are spatially close to a pedestrian, and lower scores as the window is slid away. In an exhaustive search, each pedestrian is contained in multiple windows, so by thresholding the score, one pedestrian may cause multiple detections. A third optional step in the algorithm is to group detections that are spatially close to one another. The two most common methods used in pedestrian detection are Pair-Wise max suppression and mean-shift.

A typical workflow of a pedestrian detector is shown in Fig. 4.1. Two searching paradigms are depicted in it: dense search, as explained before, and selection of regions

of interest. Image areas that likely contain pedestrians are extracted, and then fed to the classifier.

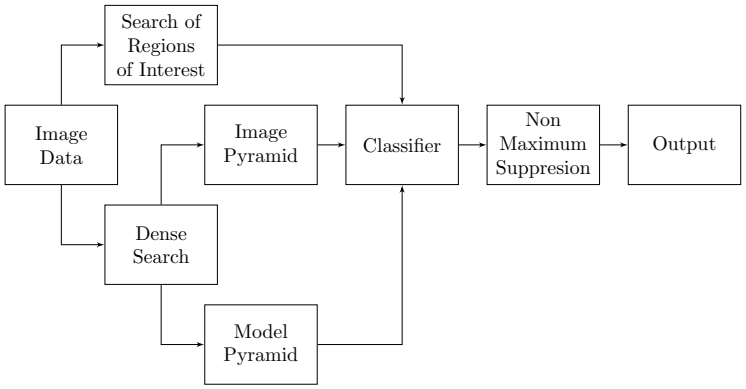


Figure 4.1: Flow chart of a pedestrian detector in images.

4.1.1. Chapter Structure

This chapter is structured as follows.

- Section 4.2 introduces the LSI Detection Database.
- Section 4.3 contains an extensive experimental study on detection performance in FIR images, following an sliding window approach. The election of parameters such as non-maximum suppression overlap, and ROI density is discussed. The experiments have been conducted using the LSI Database as well as the OSU database. From the evaluation, it is has been highlighted that the HOPE descriptor fails specially on small pedestrians.
- Section 4.4 tackles with of detection of small pedestrians. In it, a method for approximating the features of up-sampled image to the original image is presented: for each upscaled image, an specific bank of gabor filters is used to computed the phase congruency feature. The minimum frequency of those filters is shifted, proportionally to the scaling of the image.
- In section 4.5 the opposite concept is proposed. Features in the central scale are approximated to upscaled images. The purpose of this is to reduce the computational time of the algorithm.
- Finally, conclusions and future work are presented in section 4.7.

4.2. Detection Dataset

The detection dataset contains the original images from which the classification dataset are extracted, along with manual annotations of the pedestrian's positions.

The detection dataset was acquired in 13 different sessions, each containing a varying number of images. It comprises 15224 14-bit one-channel images, with dimension 164×129 pixels. The train set contains 6159 images, and the test set contains 9065 images. A representative subset of the Detection Dataset is depicted in figure 4.2

Each session occurred at a different location and with different illumination and temperature conditions. Out of those sessions 6 were used to extract the *Train* set, leaving the remaining 7 for *Test* set. This ensures that *Train* and *Test* are independent from one another. The temperature at which they were shot, which in turn affects the grey level and the histogram spread, causes the most important difference in appearance between sequences. Fig. 4.3 contains the histogram of the mean grey level value of the Train and Test Detection Databases.

Only images having pedestrians with more than 20% of the area of the original bounding box occluded behind other obstacles are considered in this evaluation.

4.3. Sliding Window Approach

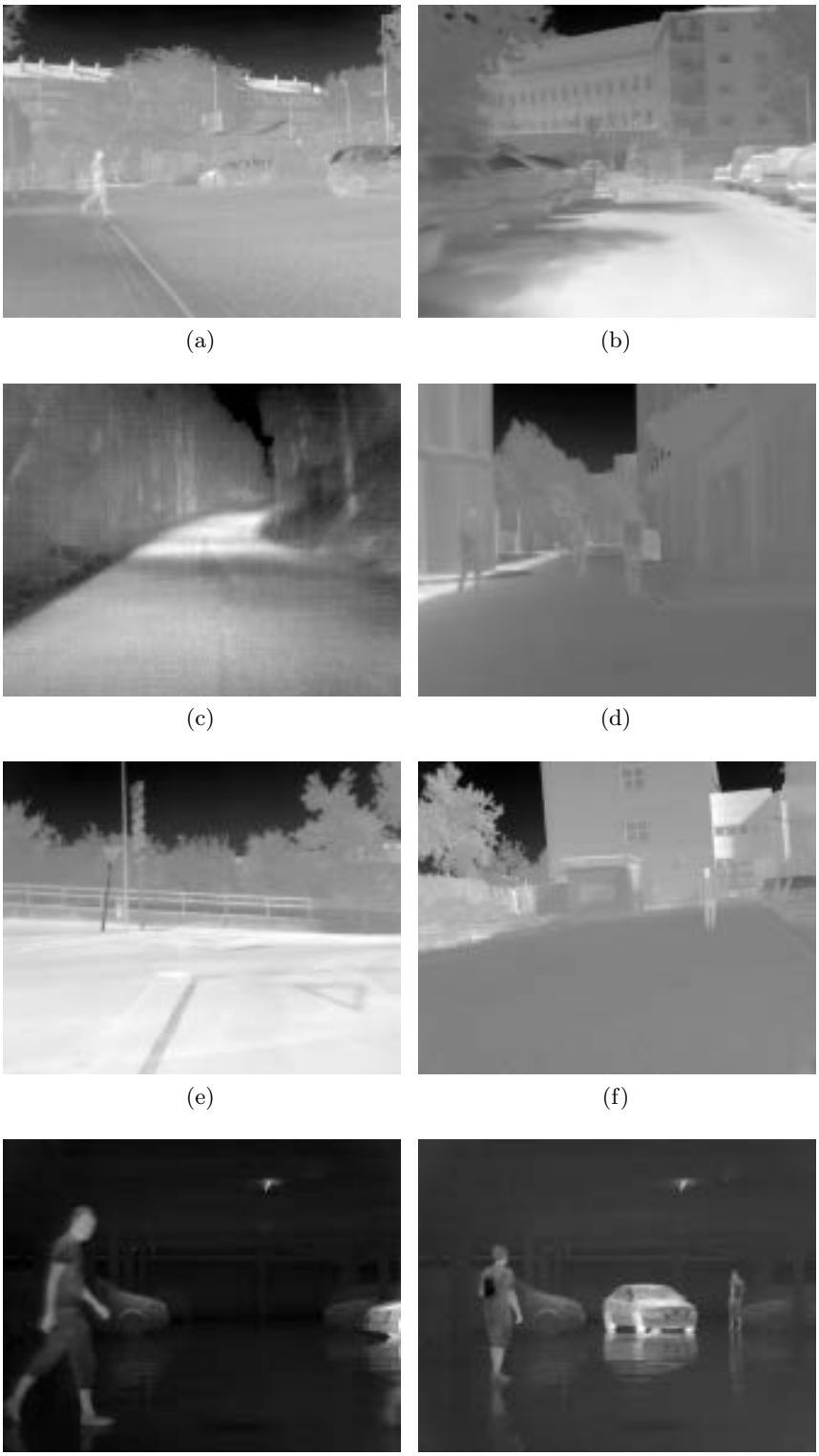
In this section, the HOPE-Lin descriptor is evaluated using a sliding window approach. A constant-sized window is slid over the image. At each location the classifier computes an score of similarity with a pedestrian. This process is repeated at several scales in an image pyramid to detect pedestrians at a range of distances from the camera. The descriptor is evaluated based on its detection accuracy, that is, In the remainder, the evaluation methodology is described. The presented descriptor is the compared with the best performing classifiers tested in the classification database, Int-HOPE and HOG-Lin. Results with an Rbf kernel are omitted, as the computation of the descriptor scores is too demanding for a detection problem.

4.3.1. Evaluation methodology

Selection of Regions of Interest For multi-resolution detection purposes the input image is resized to N scales per octave, from one octave up to one octave down, in an image pyramid. A rectangular, single-aspect ratio, window spatially scans the image at all scales (Fig. 4.4).

Pascal Criteria The detection task will be evaluated by the Pascal Criteria [69], plotting results in DET curves. Detections are considered true or false positives based on the area of overlap with ground truth bounding boxes. To be considered a correct detection, the area of overlap a_o between the predicted bounding box B_p and ground truth bounding box B_{gt} must exceed 50% by the equation 4.1, as depicted in Fig. 4.5.

$$a_o = \frac{\text{area}(B_p \cap B_{gt})}{\text{area}(B_p \cup B_{gt})} \tag{4.1}$$



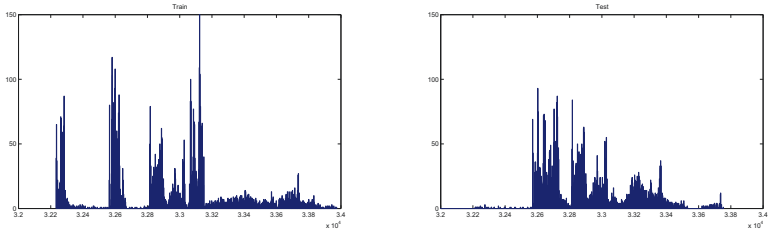
(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 4.2: Subset of the Pedestrian Detection Dataset.



(a) Train Database (b) Test Database

Figure 4.3: Histogram of mean gray level of the images in the Train and Test Databases.

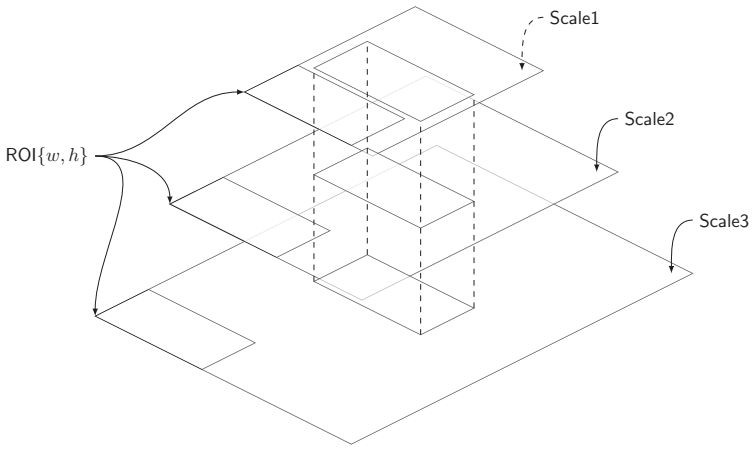


Figure 4.4: Sliding Window approach.

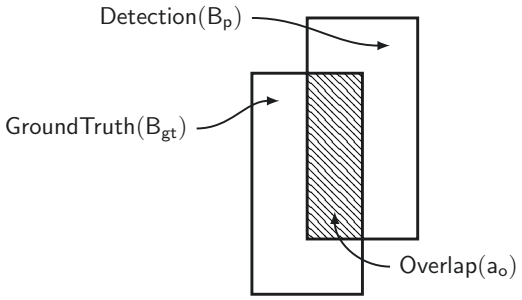
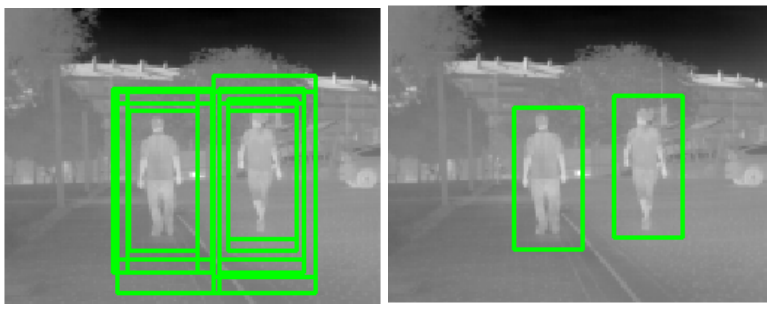


Figure 4.5: Overlapping area of Ground Truth and Detection.

Arguably, this criteria is more significant in an object detection framework for general purposes. In the specific area of obstacle avoidance in ADAS, the exact location of the pedestrian is rarely needed. Likewise, having multiple detections per pedestrian do not interfere with the warning system. For a detector to be effective it should render at least one detection per pedestrian, even with coarse detection accuracy. However, the Pascal Criteria has been adopted by the ITS community as the standard for measuring the performance of detection algorithms. A restrictive criteria encourages the development of better detectors.

Non-Maximum Suppression For each pedestrian, it is usual that more than one detection appear in the neighbourhood around the ground truth bounding box. If two or more detections match the same ground truth bounding box, only the one with the higher score would be considered a true positive. Other overlapping detections are considered false positives. To minimize the number of repeated detections, a greedy non-maximum suppression (NMS) algorithm, pairwise max (PM) suppression [74], is applied to all bounding boxes. It selects iteratively detections with higher scores than their neighbourhood, discarding detections with lower scores over an overlapping percentage. This overlap is again calculated with equation 4.1. Fig. 4.6 shows an example of multiple detections for the same pedestrians. Figure 4.6a shows the rectangular bounding boxes that have achieved a classification over the threshold. In Fig. 4.6b the result after applying the NMS algorithm is shown. Only the best bounding boxes remain. This example is shown to illustrate the concept. Effectively, the NMS algorithm is applied to all bounding boxes, whatever their classification scores are.



(a) NMS off (b) NMS on

Figure 4.6: Example of non maximum suppression of multiple detections for each pedestrian.

4.3.2. Results using the LSI dataset

Scales per Octave The number of scales per octave affects the detection rate, as well as the computation time of the detector. Figure 4.7 shows that the detection rate gets considerably lower by using two scales per octave, instead of one. By increasing the number of scales only a small improvement is found. In order to balance the detection performance and the computation time, two scales per octave will be used in the remainder of the evaluation.

Impact of Pedestrian size Pedestrian size has a big impact on detection results. Pedestrians standing far away from the vehicle appear at a lower resolution on the image and, as because of that, have lower detection rates. The impact of pedestrian resolution is experimentally assessed by varying splitting the test images, based on the presence of pedestrians within a height range. It the image contains a pedestrian taller, or shorter that the considered height range, it is not included for evaluation. Figure 4.8 shows the the results of splitting the images into the following ranges: very small pedestrians ($h \leq 25$),

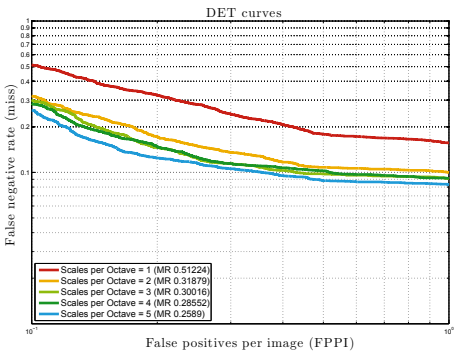


Figure 4.7: DetectionScale.

small pedestrians ($25 < h \leq 35$), medium pedestrians ($25 < h \leq 55$), large pedestrians ($55 < h \leq 85$) and very large pedestrians ($h > 85$). Figure 4.8 shows the detection rates for these subsets. The main conclusion that can be drawn is that small pedestrians get significantly worse performance. By only considering pedestrian in the medium subset of larger, it is clear that detection rates get significantly better. These results led to an evaluation to find the cause of this behavior and provide a solution that improves the results in small pedestrians. The methodology developed and the conclusions extracted are presented in section 4.4.

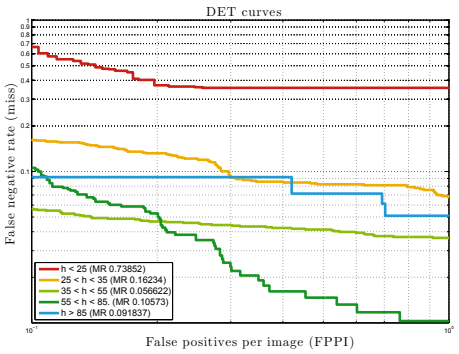


Figure 4.8: Miss rate at 10^{-1} FPPI for the detector in the *Small*, *Medium* and *Large* and *Very Large* Test Subsets.

Spatial accuracy An accurate pedestrian detector, do not only have to only have to render a high true positive rate and a low false positive rate, but also be accurate on the location of the pedestrian. The standard methodology to establish if a detection is correct is the one described in the Pascal Challenge, which requires that the detection overlaps by at least $ov = 0.5$. In Fig. 4.9 the DET curves for different overlap threshold are showed. It can be concluded that loosing the overlap threshold do not have a significant impact in miss rate over 0.3 at 10^{-1} FPPI. Increasing the threshold deteriorate the overall detector, a fact that is specially evident for overlapping thresholds $ov \geq 0.7$. These two facts indicate that the spatial accuracy of the detector is limited by the scanning window density.

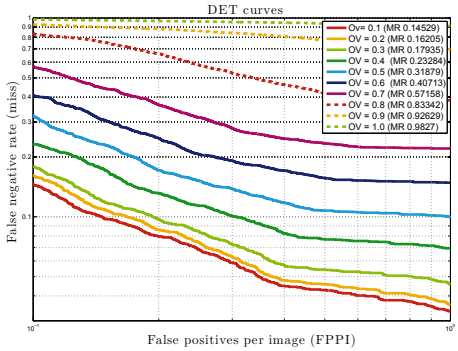


Figure 4.9: Pascal Overlap Criteria.

Comparative Results For reference, in Fig. 4.10 the DET curves of the best performing detectors, as described in chapter 3, are plotted, using an NMS overlap of $a_o = 0.5$. The Int-HOPE descriptors used have histograms of 6 bins and 8×8, and 4×4 cells receptively, and it has been trained with 200 trees. The HOG detector has been trained with the Linear SVM classifier, and results for signed and unsigned orientations are included, which are virtually identical. Rbf kernels have been omitted from evaluation, as its computational complexity is unfit for a detection problem, least high end processors or GPUs are available. Based on these results it might seem that there is a correlation between classification and detection results.

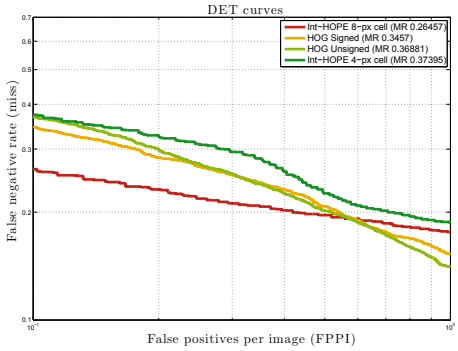


Figure 4.10: Detection DET Curves after applying the PM NMS algorithm with an overlap threshold of $a_o = 0.5$. Legend states Miss Rate (MR) at 0.1 FPPI.

4.3.3. Results in the OSU Database

Additionally, the HOPE descriptor has been tested against the OSU Thermal pedestrian Database (fig. 4.11). This dataset contains a small number of images, grouped into 10 different sequences, acquired from a static location and in similar temperature and illumination conditions. An exception is sequence number 3, where images were captured at a higher ambient temperature is higher than in the rest. The imaging device used to acquire this dataset is of ferroelectric nature. As such, it suffers from the *halo effect*. Bright

objects have around them an area that is falsely identified as having a low temperature.

In this experiment, HOG-Lin and HOPE-Lin are trained using the *Train* subset of the LSI classification database. For each sequence in the OSU database a DET curve is plotted.

The evaluation methodology used is identical as the one explained earlier, except in the selection of ROIs. The images in the OSU database were captured from a static camera and have a constant field of view. The camera position limits the range of sizes that pedestrians can have in the images. In this case, it is preferable to use an evaluation methodology more suited to a video surveillance system. Instead of searching for pedestrians with all possible heights, the ROI size is restricted to only include pedestrians between a minimum and a maximum size. Two different configurations have been used, called *Dense* and *Sparse*. In the Dense process, a pyramid of four scales with minimum scale $m_s = 1.6$ and a stepping scale of $ss = 1.1$ is created. The *Sparse* method, extracts only one scale at $s = 1.6$.

For each ROI selection method, two sets of experiments are conducted. In the first one, the detector are trained using the LSI database. In the second, the detectors are trained using images in the OSU database. For reference, other methods using the same database as a benchmark include [52] and [124].

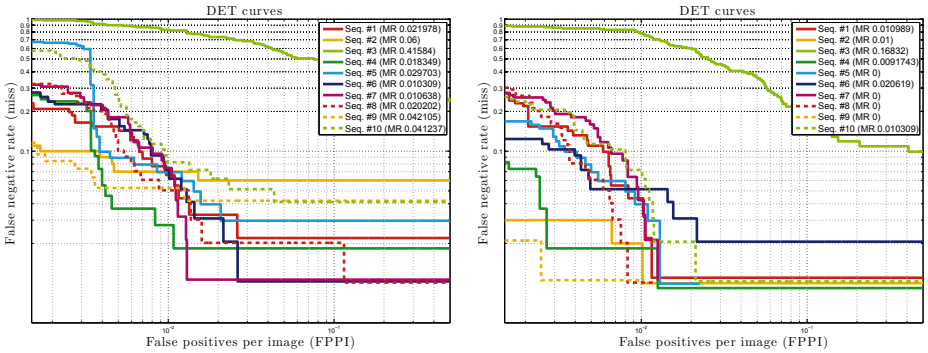
(a) (b)

Figure 4.11: Processed images from the OSU Thermal Pedestrian Database.

Trained on the LSI database The detectors used in this section have been trained on the LSI database using the same methodology as explains in chapter 3. Results of the HOPE descriptor are plotted in Fig. 4.12, and results of the HOG descriptor in Fig. 4.13. It is clear that results for both are much better than in the LSI database. This indicates that a detector trained on microbolometer images can be successfully used in ferroelectric devices, when used at low temperatures. For higher temperatures, the halo effect of the ferroelectric devices, not present in microbolometer images, degrades considerably the results.

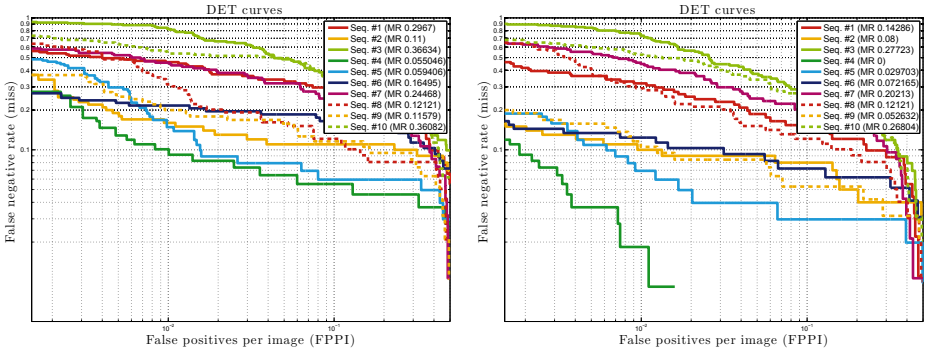
The ROI density do not seem to have any significant impact on detection performance. Dense methods improves detection by just an average miss rate of 1% at 0.1 FPPI.

Trained on the OSU database In this experiment, the two same detectors are trained using images from the OSU database. For each sequence to be tested, positive and negative samples are extracted from the other nine sequences. This ensures that the same image is never user for training and testing the same detector. Nevertheless, as background is unchanged between sequences, the classifiers tend to be over-fitted for this particular



(a) *Sparce* method (b) *Dense* method

Figure 4.12: DET curves for the 10 sequences of the OSU database, using the HOPE detector trained on the LSI database.

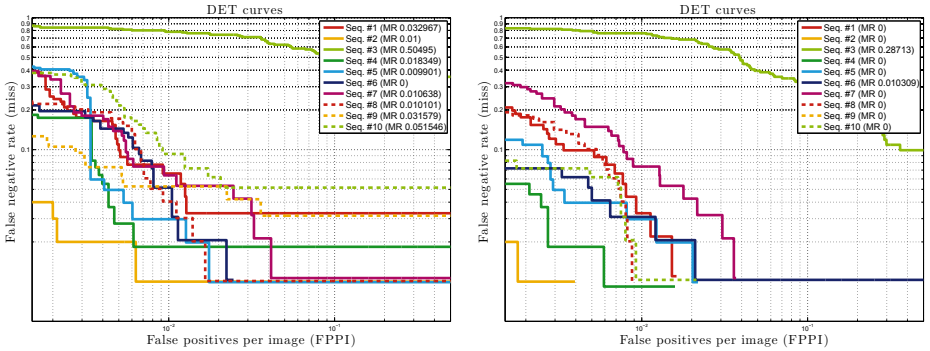


(a) *Sparce* method (b) *Dense* method

Figure 4.13: DET curves for the 10 sequences of the OSU database, using the HOG detector trained on the LSI database.

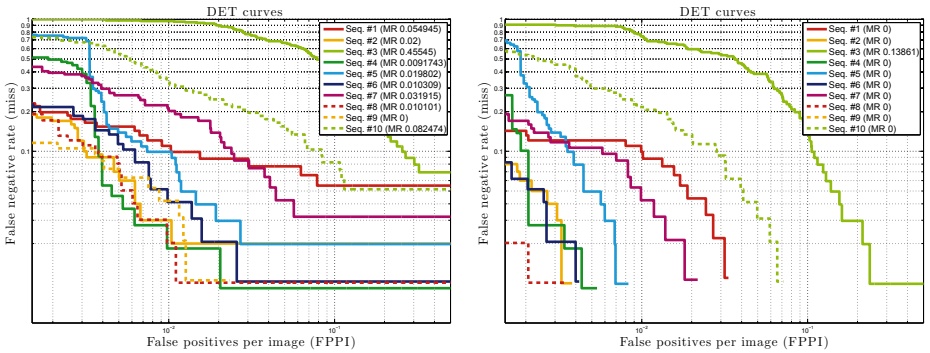
database. Background samples are randomly selected from images containing pedestrians. The maximum overlap admissible to have a background samples with a pedestrian is $ov = 0.3$.

Results of the HOPE descriptor are plotted in Fig. 4.12, and results of the HOG descriptor in Fig. 4.13. As expected, both detectors achieve better results when trained with images from the OSU database. It should be noted that results for sequence 3 do not get considerably better. Every other sequence was acquired at low temperatures, making their appearances quite different from the ones in sequence 3. The solution to achieve better results on that particular sequence would be to add to the training set with similar images. However, future improved classifiers in FIR images, should be able to cope with this order of generalization.



(a) *Sparce* method (b) *Dense* method

Figure 4.14: DET curves for the 10 sequences of the OSU database, using the HOPE detector trained on the OSU database.



(a) *Sparce* method (b) *Dense* method

Figure 4.15: DET curves for the 10 sequences of the OSU database, using the HOG detector trained on the OSU database.

4.3.4. Latent-SVM

In chapter 3 it was stated that the addition of multi-resolution information to a descriptor based on histograms of orientation can help to improve classification results by adding into one model both coarse shape and fine detail. Intuitively, finer grained descriptors retain more details, so classification should benefit from it. However, due to non-rigid deformations of pedestrians, and noise in the images, the overall classification scores for high resolution descriptors are low. To overcome this limitations, and others, Felzenszwalb et al. introduced in [74] and [75] their latent SVM detector. In this section, it is shown that the Latent-SVM approach can also be applied to FIR images.

This classification method relies on a set of filters: a low resolution root filter and a set of high resolution part filters that define a hidden or latent structure. The location of the parts of the pedestrian that best define its presence on the image, their size and quadratic cost function coefficient are the latent variables z. While training, the exact location of the ground truth bounding box of positive examples is also a latent variable. This allows for auto-correcting mistakes made while labeling the dataset. As such, this kind of approach cannot be trained using a traditional classification database, but it is instead trained using full-sized images with annotations.

The score of a filter m over a sample x is calculated as the sum of the score of a root filter r plus the sum over all parts y of the maximum score of each part minus the cost of the parts c (eq. 4.2) The cost function for each part is minimum at a specific location of the root filter. If the part is found at any other location it is penalized.

$$s(m,x) = s(r,x) + \sum_{y \in parts} \max_{y}(s(p,y) - c(p,x,y)) \tag{4.2}$$

The detection is treated as a binary classification problem in a sliding window approach. Given a training set $D = \{\{x_1, y_1\}, \cdots, \{x_n, y_n\}\}$, where $x_i \in R^d$ is a feature vector and $y_i \in \{\pm 1\}$ is a binary label, each region of interest of the image is assigned a score,

$$f_{\beta,Z}(x) = \max_{z \in Z(x)} \beta \cdot \Phi(x,z) \tag{4.3}$$

where x is the region of interest, z are latent values, and β is a vector of model parameters. The function $\Phi(x,z)$ is the feature vector assembled from the root and parts filters. β should then minimize

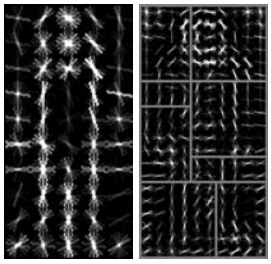
$$\frac{1}{2}\| \max_{i=i,\ldots,k} \beta\|^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i \cdot f_{\beta,Z}(x_i)) \tag{4.4}$$

where k is the number of components in the model, and the second term determines the softness of the SVM margin.

The training process can be divided into two parts. The first one trains a *root* filter of low resolution descriptors by warping positive samples to simulate a large number of

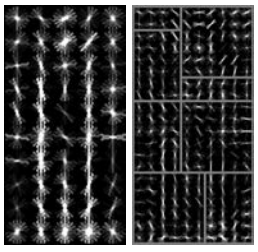
human poses. The original method considers pose variation by a mixture of models, i.e. pedestrian heading to the left or to the right. The score of a sample would then be the maximum over the set of models. In our implementation, only one model is used.

To assess the impact on performance of part based detection, two descriptors have been trained using the Latent-SVM approach: HOG and HOPE. Detection performance is evaluated using the same methodology explained in section 4.3.1, except for the following: for each scale computed in the image pyramid, an additional scale is added with double the resolution, in order to extract the part models. The latent-SVM models for both the HOG and HOPE descriptors have been trained using the same parameters. They have been initialized for $k = 8$ latent parts and root filters with (6×6) cells. Fig. 4.16 represents the filters trained with the HOG descriptor and Fig. 4.17 the filters trained with the HOPE descriptor.



(a) Root Filter (b) Parts

Figure 4.16: Latent SVM filters using HOG features.



(a) Root Filter (b) Parts

Figure 4.17: Latent SVM filters using HOPE features.

The DET curves in Fig. 4.19 compares the performance of root, latent and parts detectors for the HOG and HOPE descriptors.

From these results some initial conclusions may be extracted. In the first place, the performance of the root filter improves the results of a naively trained classifier. This leads to the conclusion that warping positive samples is a relevant step in the algorithm, and one that could be easily implemented in the training algorithm of the descriptors described in chapter 3. However adding the part models seem to degrade them slightly. In low resolution

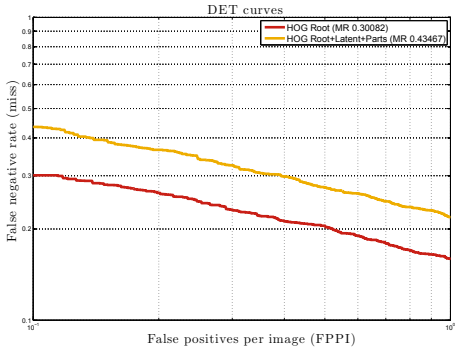


Figure 4.18: Detection DET Curves of Latent-SVM HOG after applying the PM NMS algorithm. Legend states Miss Rate (MR) at 0.1 FPPI.

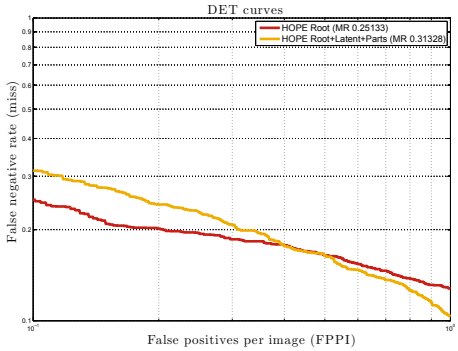


Figure 4.19: Detection DET Curves of Latent-SVM HOPE after applying the PM NMS algorithm. Legend states Miss Rate (MR) at 0.1 FPPI.

images, such as the ones presented in this database, latent parts are hard to find in the smallest objects because they are computed at twice the resolutions as the root filter and as such, much of the information is lost. The evaluation of the performance of the part based approach would need a new database, of higher resolution images.

4.4. Small pedestrians

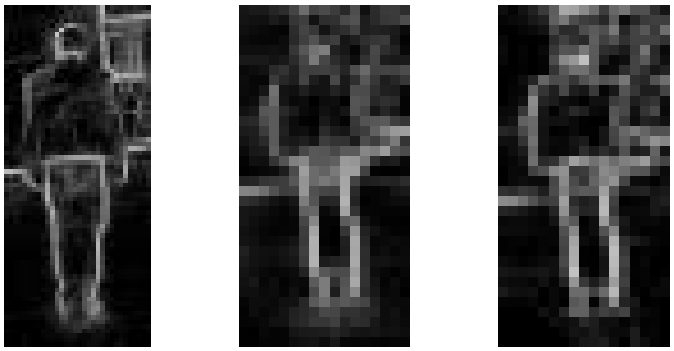
In section 4.3.2 it was stablished that the detector achieves worse detection rates for small pedestrians, than it does for large pedestrians. This may be explained by considering that small pedestrians appear blurry when upscaling the images to search for small candidates. Object borders in up-sampled images extend over a wider spatial range that in images resized to a scales smaller than $s = 1$. This result in a high intra-class variance. The classification module, upon which the detector relies, has been trained with pedestrians that, in most cases, have a height greater that 40 pixels, as shown in chapter 3. Therefore, a linear classifier tends to separate the most common appearance. This match approximately to the appearance of pedestrian extracted in the original scale, in the detection problem. There are possible approaches that address this issue: gather an over-complete database that equally represent all possible appearance and training several models, possibly with different descriptors, for the different subclasses, or modify the descriptor to better generalize the intra-class variability. In stead of that, in this section the following contribution is proposed: to approximate features at different scales to the appearance that would have the object if it was extracted from the central scale.

Phase congruency features extracted from different scales of the image pyramid can be approximated to the central scale by modifying the log-Gabor filters used to calculate them, as will be demonstrated in the remainder. Figure 4.20 illustrates this concept. Figures 4.20a and 4.20b represent the phase congruency magnitude of a cropped window containing a pedestrian, separated by one octave. Both features have been procesed with the original bank of log-Gabor filters. Figure 4.20c represent the phase congruency magnitude of the same image as Fig. 4.20b, calculated with the new bank of log-Gabor filters, for a scale of $s = 2$. The result intuitively resembles more the higher resolution image.

The bank of filters are calculated scale-wise by shifting the minimum wavelength so that, for instance, a low resolution pedestrian appears with shaper borders in the phase congruency image. The amount of this phase shift is different in each scale is proportional to its distance to the original scale $s = 1$. This statement is experimentally validated in the remainder.

A definition of a log-Gabor filter is expressed in equation 4.5, where $G(\omega)$ is the value of the filter for frequency ω, ω_0 is the center wavelength of the sinusoid, and k/ω_0 remains a constant for all the filters.

$$G(\omega) = exp\left(-\frac{1}{2}\frac{\left(\log(\omega/\omega_0)\right)^2}{\left(\log(k/\omega_0)\right)^2}\right) \tag{4.5}$$



(a) 64×32 (b) 32×16, λ (c) 32 × 16, λ/s

Figure 4.20: Example of small pedestrian

The frequency of the filter at scale s is:

$$\omega_{0,s} = \frac{1}{\lambda_{0,s}} = \frac{1}{\lambda_0 \cdot \Delta\lambda^{(s-1)}} \tag{4.6}$$

where λ_0 is the minimum wavelength, $\Delta\lambda$ is the distance between filters in the phase spectrum and s is the scale of the filter.

To compare the similarity of the phase congruency features, a set of cropped images containing a pedestrian at different scales, and with different λ_s have been compared using the Peak Signal-to-Noise Ratio (PNRS) magnitude (eq. 4.7).

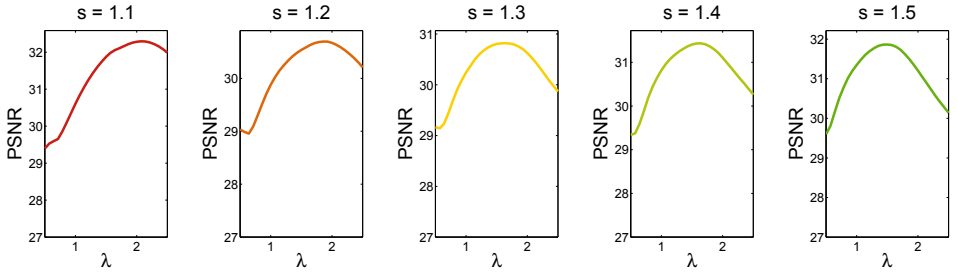
$$PSNR = 10 \cdot \log\left(\frac{1}{\sqrt{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left(C_0(i,j) - C_s(i,j)\right)^2}}\right) \tag{4.7}$$

Where C_0 is the cropped image at scale $s = 1$ and C_s is the cropped phase congruency cropped image calculated with a shifted minimum frequency. The figure 4.21 represents the value of PNSR, varying the value of ω_s. Each of the subfigures represents the similarities between the original scale and the new scale, where $s > 1$ represent up-sampled pedestrians. For down-sampled pedestrians, this approximation is not needed, as once rescaled the appearance is similar to a pedestrian in the central scale. Notice that, as the pedestrian gets smaller (incresing s) the wavelength have to be shifted to smaller values. For values of $s > 2$ it was not possible to confirm if the approximation holds, as the maximum height of a pedestrian in the database is within the first octave.

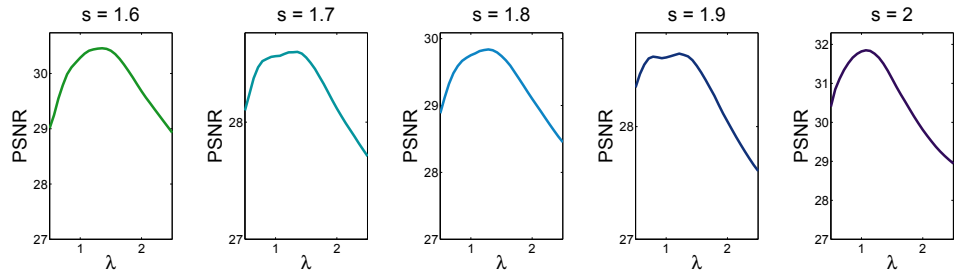
The minimum wavelength that has been found to maximize the appearance of phase congruency features at different scales, when compared with the original scale has been equation 4.8. Arguably, raising the wavelength frequency can set the central filter close to the Nyquist frequency or beyond it. If that is the case, aliasing artifacts arise in the images.

Nevertheless, shifting the bank of filters overall improves detection rates.

$$\omega'_{0,s} = \frac{s}{\lambda_0 \cdot \Delta\lambda^{(s-1)}} \tag{4.8}$$



(a) $s = 1.1$ (b) $s = 1.2$ (c) $s = 1.3$ (d) $s = 1.4$ (e) $s = 1.5$



(f) $s = 1.6$ (g) $s = 1.7$ (h) $s = 1.8$ (i) $s = 1.9$ (j) $s = 2.0$

Figure 4.21: PSNR of a range of values of the minimum wavelength for different scales.

The results of applying this feature approximation in the HOPE detector to upscaled images in the pyramid are shown in Fig. 4.22. The miss rate is consistently reduced for every value of FPPI. This approach do not need additional computation time, as the banks of filters only need to be calculated once. Notice that descriptors are computed at all scales in the pyramid.

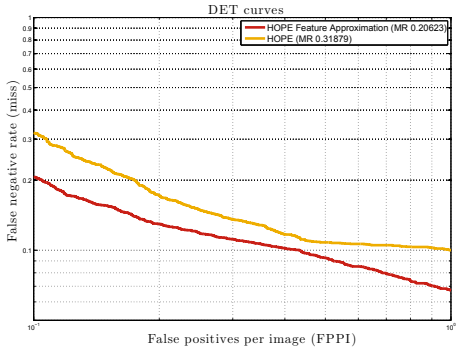


Figure 4.22: Feature approximation.

4.5. Scale Approximation

The feature approximation presented in the previous section can be *reversed*, in the sense that, instead of approximating the appearance of scale s to resemble the central scale, the latter can be approximated to the appearance of scale s. The main purpose of this method is to reduce the computation time. In stead of up-sampling an image and then computing its phase congruency score, the latter can be computed directly on the original images and then be resized, after some modifications to the bank of log-Gabor filters have been applied. This reduces the computation time of the detection algorithm, as processing the phase congruency of large images is time consuming. By applying this method, the largest image to process would have the same size as the central scale. Effectively, this method is only used in images that would have required to be up-sampled. For large pedestrians, the original image is still down-sampled and phase congruency is computed on the smaller-sized image. Computing the phase congruency on the original image and then downsampling the image to computed the descriptor is still possible, but doing so would defeat the purpose of reducing the computation time.

As before, the minimum wavelength of the bank of log-Gabor filters is shifted for each scale needed. This shift maximizes the appearance of the phase congruency image computed on the original images to the phase congruency that would have if computed on a resized version of the same image.

It has been found that the minimum wavelength that minimizes the error can be approximated to (eq: 4.9). Figure 4.23 represent the PSNR for a set of υ values.

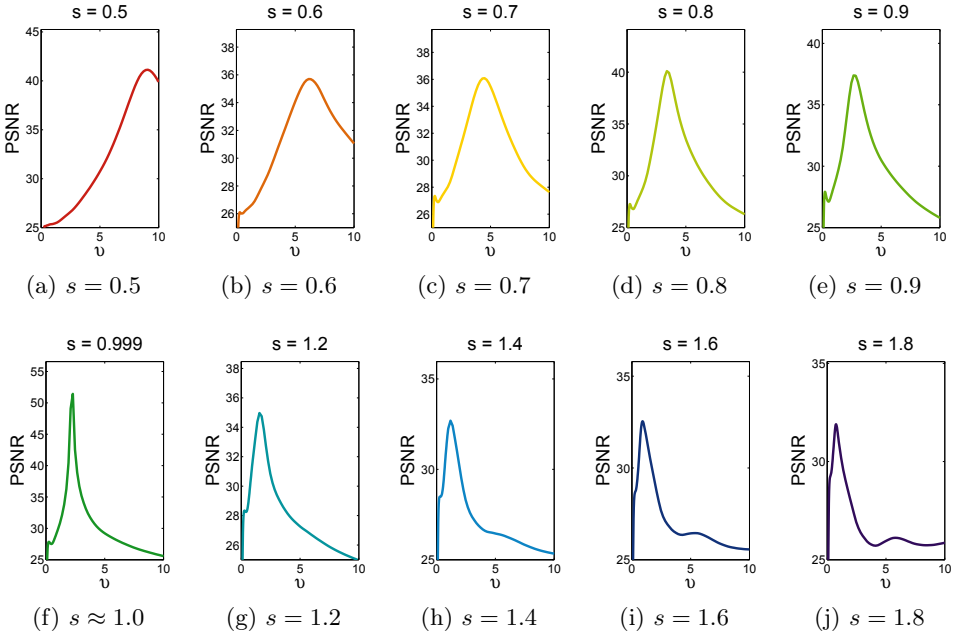
$$\omega''_{0,s} = \frac{1}{\upsilon \cdot \Delta\lambda^{(s-1)}} = \frac{s^2}{\lambda_0 \cdot \Delta\lambda^{(s-1)}} \tag{4.9}$$

The error for this approximation is found to be:

$$e = \sqrt{\frac{\sum_{i=1}^{S} \left(\left(s_i^2 - \arg\max_{\upsilon} \mathrm{PSNR}(I_0, I_s, \upsilon, s) \right)^2 \right)}{S}} = 0.0219 \tag{4.10}$$

Where s is the scale in the image pyramid, S is the total number of scales, PSNR is the peak signal-to-noise function, I_0 is the phase congruency image computed at scale $s = 1$, with a minimum wavelength of υ, which is the parameter to be optimized, and I_s is the phase congruency image at scale s. This error is calculated within the upper octave. For larger values, the approximation does not hold, and new banks of filters may be needed, for instance, to approximate $s = 2$ to scales $s > 2$.

This method is tested using the HOPE-Lin descriptor as a benchmark. The detection evaluation process follows the same methodology as explained before. The detection DET curves for the original definition of the HOPE descriptor and the scale approximation version are plotted in Fig. 4.24. While slightly degrading performance, this approximation closely resembles the original results, using a fraction of the time. On average, using this method requires 58% of the processing times, for the scales per octave tested in section



(a) $s = 0.5$ (b) $s = 0.6$ (c) $s = 0.7$ (d) $s = 0.8$ (e) $s = 0.9$

(f) $s \approx 1.0$ (g) $s = 1.2$ (h) $s = 1.4$ (i) $s = 1.6$ (j) $s = 1.8$

Figure 4.23: Scale approximation

4.3.1. This time only takes into account the time to process the descriptor, not the time that takes the classfication.

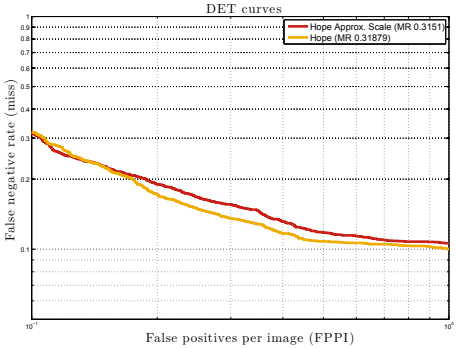


Figure 4.24: Scale approximation.

Finally, the scale approximation algorithm is sumarized in algorithm 1.

Algorithm 1 Scale approximation

Require: s in Scales in Pyramid
  Precomputed the log-Gabor filters
  for s **do**
    if $s > 1$ **then**
      compute log-Gabor filters with $\omega''_{0,s}$, where $\omega''_{0,s} = \frac{s^2}{\lambda_0 \cdot \Delta\lambda^{(s-1)}}$
    else
      compute log-Gabor filters with $\omega'_{0,s}$, where $\omega'_{0,s} = \frac{s}{\lambda_0 \cdot \Delta\lambda^{(s-1)}}$
    end if
  end for
  Detection
  for s **do**
    if $s > 1$ **then**
      Compute phase congruency on original image
      Resample phase congruency image by s.
    else
      Resample image by s.
      Compute phase congruency on resampled image.
    end if
  end for
  Compute descriptor
  Classifify

4.6. Improving the performance

Pedestrian detection is a very broad area, with many topics involved. Two of the most relevant ones that hasn't been discussed yet in this dissertation are generation of regions of interest and occlusion handling. In this section, some notions about those topics, derived from initial research, are proposed.

4.6.1. Selection of Regions of Interest

Detecting pedestrians in images involves identifying regions of the image that have characteristics consistent with belonging to a person. Specialized classifiers are able to distinguish the target object from the background, or other similar objects. One of the most common approaches, and the one followed in section 4.3, consists of an exhaustive search, applying the classifier at every position in the image. However, intuitively it is easy to see that large area of the images do not contain relevant information. Applying a complex classifier to those irrelevant areas of an image is a waste of time, a very important factor in ADAS. In this section two methods for selecting regions of interest (ROI) are proposed.

4.6.1.1. Selection of regions of interest by temperature segmentation.

Far infrared images have a very valuable advantage over the visible light ones. They do not depend on the illumination of the scene. The output of those cameras is a projection on the sensor plane of the emissions of heat of the visible objects, which is proportional to their temperature. Some authors have develop classification methods based on the temperature distribution of the human body. In chapter 3 a method for classifying pedestrians in FIR images, based on temperature templates was proposed. In this section, the described methodology for thresholding images based on object temperature is used to extract regions of interest. The outline of the algorithm is as follows. The FIR images are thresholded based on the temperature of the human head. The position in the image of the top the extracted blobs is reprojected to world coordinates and, assuming that the ground is flat a ROI is generated in the image. These ROIS are rectangular and with constant aspect ratio. In order to project between image an ground coordinates the projective model has to be defined, which follows.

Projective model The camera is modeled as a pin-hole. The intrinsic parameters are know and so is the position and orientation of the camera. The world system of coordinates O is placed on the ground plane, moving along with the vehicle and so does the camera position O' (Fig. B.3).

The position of the pedestrian is modeled as a gaussian distribution in the xy plane of the ground. To determine accurately its distance to the camera, the homography of the ground plane onto the sensor is calculated for each frame. The projection of a 3D point in the image plane can be done if it is known its relative position to a certain plain. In this case, the camera position relative to the ground plane is known and can be assumed that it

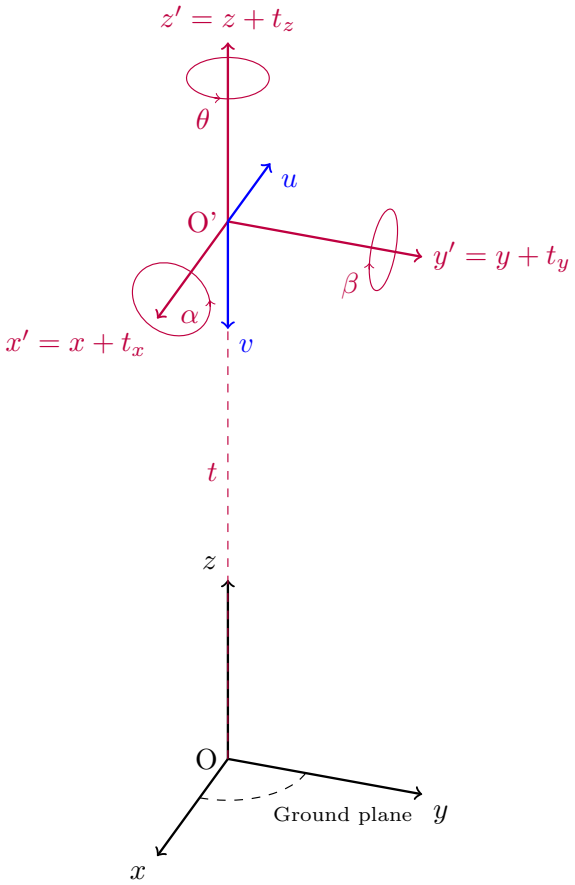


Figure 4.25: Reference system of world and camera coordinates.

is constant. A more detailed explanation of the system setup can be found in section annex B. The rotation of the camera is known via a three degrees gyroscope. The homography function is stated in equation 4.11,

$$\begin{bmatrix} U \\ V \\ S \end{bmatrix} = M \cdot \left(R \cdot \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \right) \tag{4.11}$$

$$M = \begin{bmatrix} f_u & 0 & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{bmatrix} \tag{4.12}$$

$$R = R_x \cdot R_y \cdot R_z \tag{4.13}$$

$$R_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha + \Delta\alpha) & \sin(\alpha + \Delta\alpha) \\ 0 & -\sin(\alpha + \Delta\alpha) & \cos(\alpha + \Delta\alpha) \end{bmatrix} \tag{4.14}$$

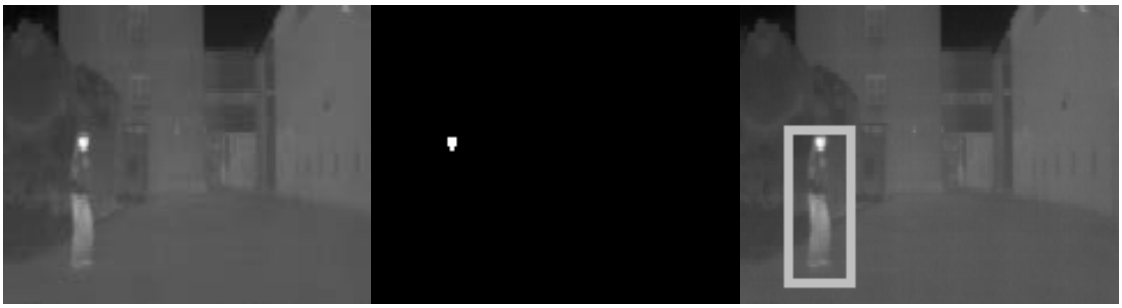
$$R_y = \begin{bmatrix} \cos(\beta + \Delta\beta) & 0 & -\sin(\beta + \Delta\beta) \\ 0 & 1 & 0 \\ \sin(\beta + \Delta\beta) & 0 & \cos(\beta + \Delta\beta) \end{bmatrix} \tag{4.15}$$

$$R_z = \begin{bmatrix} \cos(\theta + \Delta\theta) & \sin(\theta + \Delta\theta) & 0 \\ -\sin(\theta + \Delta\theta) & \cos(\theta + \Delta\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.16}$$

where f_u and f_v are the focal lengths of the lens in the u and v directions; (c_u, c_v) is the optical center of the sensor in point O' of figure B.3 , R is the rotation matrix around the three axis x, y, z and t if the translation vector of the optical center of the camera from the coordinates origin. The camera is located directly above of the ground. Its coordinate system is imaginary and correlated with the axis of the camera. Because of it, the translations in the xy plane are always zero ($t_x = t_y = 0$). The only translation is the distance of the optical center to the ground t_z. Initially, the rotation of the sensor plane is $\alpha = \frac{\pi}{2}$, $\beta = \pi$ and $\theta = 0$. Using an IMU the increments of the angles ($\Delta\alpha$, $\Delta\beta$, $\Delta\theta$) are updated every frame. U and V are the image homogenous coordinates. The true pixel coordinates are $u = \frac{U}{S}$ and $v = \frac{V}{S}$.

The results of the projection of a world point in the image plane are specially sensitive to variations of the skew angle $\Delta\alpha$ (See figure B.3). In the presented system this angle is know for each capture frame with the help of an gyroscope with three degrees of freedom attached to the same base as the camera.

Most of the time, while driving in urban environments, the roll angle is close to zero. If that restriction can be applied, the projection is greatly simplified.



(a) Original Image. (b) Hotspots. (c) Regions of Interest.

Figure 4.26: Selection of regions of interest.

$$u = c_u - \frac{f_u \cdot w_x}{t_z \cdot \sin(\Delta\alpha) - w_y \cdot \cos(\Delta\alpha)} \tag{4.17}$$

$$v = \frac{t_z \left(\mathrm{c_v} \cos\left(\frac{\pi}{2} + \Delta\alpha\right) + \mathrm{f_v} \sin\left(\frac{\pi}{2} + \Delta\alpha\right)\right) - w_y \left(\mathrm{f_v} \cos\left(\frac{\pi}{2} + \Delta\alpha\right) - \mathrm{c_v} \sin\left(\frac{\pi}{2} + \Delta\alpha\right)\right)}{t_z \cos\left(\frac{\pi}{2} + \Delta\alpha\right) + w_y \sin\left(\frac{\pi}{2} + \Delta\alpha\right)} \tag{4.18}$$

where f_u and f_v are the focal lengths on the u and v directions of the image; c_u and c_v are the coordinates of the center of the image; c_u and c_v are the image coordinates from the upper left corner. These four parameters are measured in pixels. t_z is the height of the camera over the ground.

4.6.1.2. Extraction of warm areas.

Extraction of the warm areas is done by thresholding the image in two phases: the first one tries to extract the heads of the pedestrians in the images; the second, the whole pedestrian silhouette. Objects within the normal temperature of the human body are thresholded. The threshold selection process was discussed in chapter 3. The result is a binarized image, containing blobs that can represent parts of the human body, specially heads and hands (figure 4.26). Since this first step searches for the pedestrian head, those blobs that are not in the upper half of the image are ignored. Those blobs that are not within some geometric restictions are also excluded.

Once the head candidates have been selected, a first set of regions of interest are generated. The highest point of the head is also the top of the box, while the lowest point is at the closest point of the ground at that resolution. This way, the whole body of the pedestrian is included in the box, if there is any (figure 4.26c).

At this point only the position of the head in the image is know, thus these bounding boxes have to be big enough to contain any pedestrian, no matter what height. A first approach is to suppose that the head is at a height h from the ground plane on which the pedestrian is standing. The distance of the pedestrian to the camera (w_y) is given by

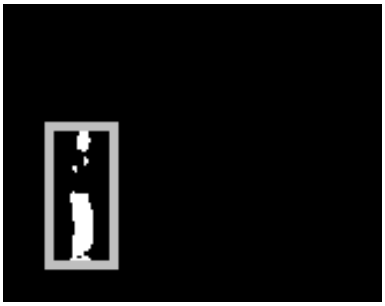


Figure 4.27: Search of the pedestrian inside the previously calculated ROI.

equation (4.19), where w_z is the camera distance to the ground plane, v is the position of the top of the region of interest, in image coordinates, h is the height of the pedestrian and f_v is the vertical focal length. The base of the region of interest is calculated with equation (4.18), for this new distance. The width of the bounding boxes is set to be 1/2 of the height.

$$w_y = \frac{f_v(h - w_z)}{v - c_v} \tag{4.19}$$

The regions of interest generated from the original image are now binarized with a threshold of t_1, that is the lower temperature established for the human body. Since most pedestrians height is less than $190cm$, h is set to $200cm$. The distribution of temperatures inside the ROI for most pedestrians will only seize a fraction of it. The window is then resized, keeping the same proportions, assuming that the lowest part of the pedestrian are the feet, and that those are resting on the flat ground ahead the vehicle (figure 4.27).

4.6.1.3. Selection of ROIS by edge density

In this section a fast way to discard areas with low probability of containing a pedestrian is detailed. The outline of this method is as follows. First, a set of regions of interest is created as rectangular boxes with a aspect ratio of 1/2. To avoid searching for pedestrian in unlikely areas some geometric restrictions are applied. Only pedestrians on the ground plane and inside the trajectory of the vehicle are looked for. Knowing the intrinsic parameters of the pin-hole modeled camera and its position and orientation over the ground plane it is possible to establish an homography projection of the ground plane over the sensor plane. The regions of interest are created at fixed ranges of distances to the camera. The world system of coordinates is placed on the ground plane, moving along with the vehicle and so does the camera position.

The phase congruency score for every pixel in the image is determined by equation 3.9. Regions of the image without significant phase transitions have a low phase congruency score. On the other hand, regions with high scores are spatially distinguishable from its neighbors. Smaller regions with constant phase congruency values are more likely to contain useful information of borders belonging to pedestrians (Fig. 4.28). The relevance of each pixel is set to be the inverse of its distance to the closer border. To evaluate the weight of

each pixel a watershed segmentation [143] is performed over the distance transform of the magnitude image of the phase congruency. As an example, Fig. 4.28d contains a watershed segmentation of the distance transform image (Fig. 4.28c).

As explained before, the pixels of the watershed image have a weight based on the size of the blob they belong to. Bigger blobs have lower weights, as they are smooth areas with less important information. The score of each region of interest is the sum of the weight of every pixel in it, normalized by the size of the ROI (equation 4.20).

$$S = \frac{\sum_{x=0}^{w} \sum_{y=0}^{h} \phi_{x,y}}{w \cdot h} \tag{4.20}$$

Where S is the score of the ROI, w is the width, h is the height and $\phi_{x,y}$ is the weight of each pixel. If the score of the box is below a certain threshold, that region can be ignored and won't be fed to the classifier. In Fig. 4.28e only the surviving ROIs with a high score are represented. Only this reduced subset will be further processed, thus reducing the processing time.

4.6.2. Part-based detection

This approach presents an initial research on part based detection. The outline of which is as follows: on heavily occluded pedestrians full-body detector usually have low detection rates. However, the parts that are still visible can give a hint of the presence of the pedestrian. This approach suggests combining the responses of parts in a Markov Logic Network (MLN) that then decides if there are enough detected parts for the sample to be considered a pedestrian. In practice a matrix of non-overlapping histograms of orientation are used as features for the parts. Two approaches have been tested: building a predefined set rules, and letting the Logic Network seek iteratively relations among all possible rules. The resulting weights allow inferring the presence of pedestrians from incomplete samples by looking for hidden part structures, outperforming rigid models looking for complete objects, in the case of heavily occluded pedestrian.

Regarding the application of MLNs to pedestrian detection, Oliveira et al. presented in [160] a Lidar-based system, immune to partial segmentation of data. The system infers the relationship of the sub-segments and their context. In [161] their work is extended in a multisensory scheme that fuse visual information with Lidar data, based on spatial relationship of parts-based classifiers, via a MLN.

Markov logic networks (MLN) are a first order knowledge base, based on a set of logic rules that define the occurrences of events or the relations or conditionality between them. Each logical formula f_i has an associated weight w_i, which is trained using discriminative learning [187] from a labeled database, and assumed to have equal prior probability. Thus each logical statement is no longer binding, but the events they that imply will be more likely to be true based on its weight.

The network models the joint distribution of the events. In this case, events are binary variables that answer whether sample is positive or not. If any of these formulas are true,

the network implies that the detection is positive. The probabilistic inference that responds to the query *isPedestrian* is calculated on the minimum subset of events using the Lifted Belief algorithm [188]. In equation 4.21, the joint probability of the events given their responses to the logic formulation is defined as the normalized exponential product of the formula-weight.

$$P = \frac{e^{(\sum_i w_i f_i)}}{Z} \tag{4.21}$$

Where w_i are the weights, f_i are the set of logic formulae, and Z is the normalization factor.

The detection window is divided into subparts and an SVM is trained for each one (Fig. 4.29). In the MLN, a positive detection of a part would imply that the query *isPedestrian* is true.

For each cell in the pedestrian descriptor an SVM classifier calculates the boundary between the pedestrian and background classes by searching the hyperplane that maximally separates the training set. The SVM is trained with a subsample of the train dataset, on which pedestrians have a lateral occlusion between 0 and 50%. Pedestrian and non-pedestrian images are resized to have the same dimensions. The decision function in equation 4.22 is optimized so that $y_k(x)$ maximizes the distance between the nearest point (x_i) and the hyperplane. The linear discriminant function is:

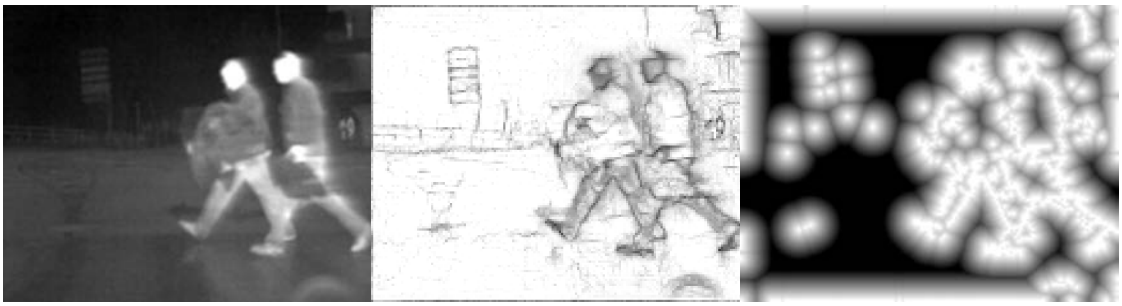
$$y_k(x) = w^T \cdot \Phi(x_k) + b_k = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \sum_{l=1}^{m} w_{ijl} \Phi_l(c_{ij}) + b_k \tag{4.22}$$

Where w is normal to the hyperplane, b_k is the classification bias of part k, and $\frac{b_k}{||w||}$ is the perpendicular distance from the hyperplane to the origin. $\Phi(c_{ij}) \in \mathbb{R}^m$ is the corresponding histogram of gradients with m cell bins, at pixel c_{ij}. Φ is the kernel function that is used to project the samples. In this evaluation a linear kernel has been used. The sample is assigned to one of the two classes by thresholding the decision function, where a sample with a score of $y(x) > b_k$ is classified as a pedestrian and as background otherwise.

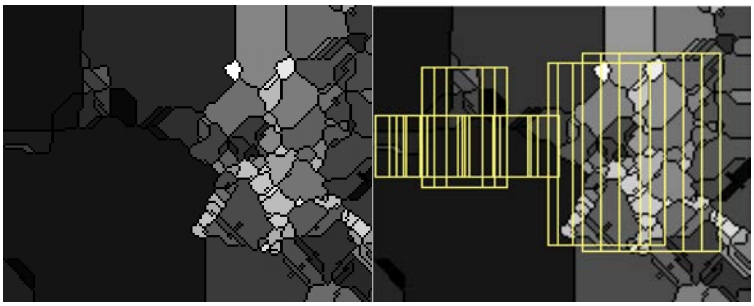
Each pedestrian part classifier $(y(p_{g_i}))$ is evaluated by the area under the ROC curve (a_{uc}). Only those with an a_{uc} over a threshold (thr) are used to calculated the MLN weights. For each pedestrian classifier with $a_{uc} >$ thr, the bias is selected by calculating the optimal operating point of the ROC curve, that is, the one on which the curve intersect with the line with slope

$$S = \frac{\mathrm{c}(P|N) - \mathrm{c}(N|N)}{\mathrm{c}(N|P) - \mathrm{c}(P|P)} \cdot \frac{N}{P} \tag{4.23}$$

Where c is the cost of misclassifying one sample as a member of the opposite class, and P=TP+FN and N=TN+FP are the total number the positive and negative samples, respectively.



(a) Original Image. (b) Phase congruency. (c) Distance transform.



(d) Watershed Clustering. (e) Regions of Interest.

Figure 4.28: Selection of ROIS by edge density.

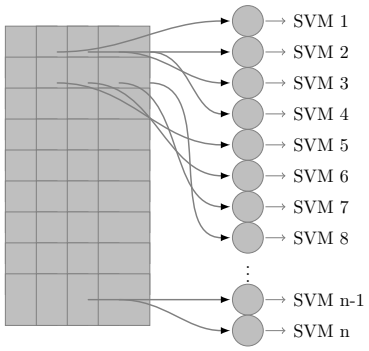


Figure 4.29: Descriptor blocks of pedestrian parts. For each block an SVM classifier is trained.

Fixed set of First Order Logic rules The knowledge base is made from one rule for each part classifier of a sample, plus a rule for the full-body detector. If a part is classified as pedestrian-part the network imply that the sample is a pedestrian.

$$\begin{aligned} \forall w, \texttt{Body}(w) &\Rightarrow \texttt{isPedestrian}(w), \forall i \in n_p \\ \forall w, \texttt{isPart}_i(w) &\Rightarrow \texttt{isPedestrian}(w), \forall i \in n_p \end{aligned}$$

Where n_p are the set of parts that have classifiers with an area under the roc curve $a_{uc} > \text{thr}_i$, $\texttt{isPart}_i(w, t)$ is true if part i of window w if classified as positive. Figure 4.30 shows the relative classification weight of pedestrian parts on the FIR database.

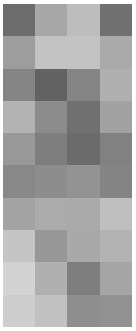


Figure 4.30: Relative classification weight of pedestrian parts scaled in $w_p = \{0, 1\}$.

Structure Learning The set of logic rules can be automatically learned from the ground truth database, along with their respective weights. The relation between each possible combination of clauses is tested and added to the set of rules if there exist statistical significance. This approach is used to detect incomplete pedestrians, be it because they are occluded of because only part of it falls inside the image. Figure 4.31 shows a representation of a latent structure made up from five parts for left-side occluded pedestrian. Occlusion percentage is calculated as depicted in Fig. 4.32. The result is a MLN trained to find latent structures of heavily occluded pedestrians. In this implementation, clauses are found using a beam search algorithm [115].

The following results are based on set of bounding boxes containing occluded pedestrian, cropped from the full-sized images, and resized to a common size. The test dataset has been divided into five subsets based on the percentage of lateral occlusion. Figure 4.33 represents the true positive rate plotted against the false positives per image on the FIR database. Those are the results of applying the part-based model, where Fig. 4.33a are results of using the fixed set of logic rules, and Fig. 4.33b represent the results of the latent structure. For reference, Fig. 4.34 represents the roc curve of full body classification on the FIR database for different percentages of occlusion. It can be appreciated that, for large values of occlusion, classification degrades. In both part-based detection approaches, the classification results get significantly better for large values of occlusion, while it is slightly

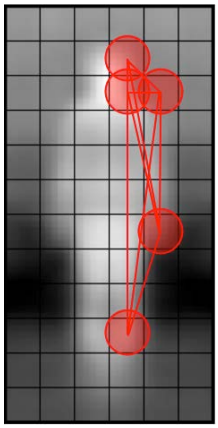


Figure 4.31: Occlussion

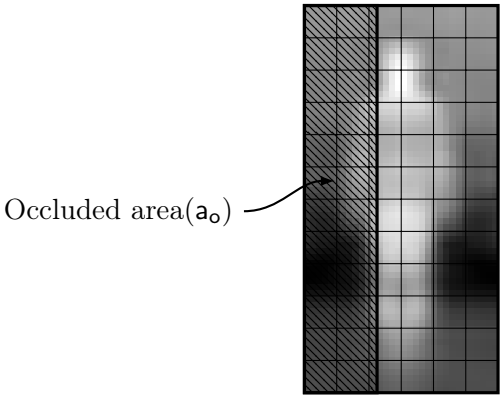


Figure 4.32: Occluded area of the region of interest.

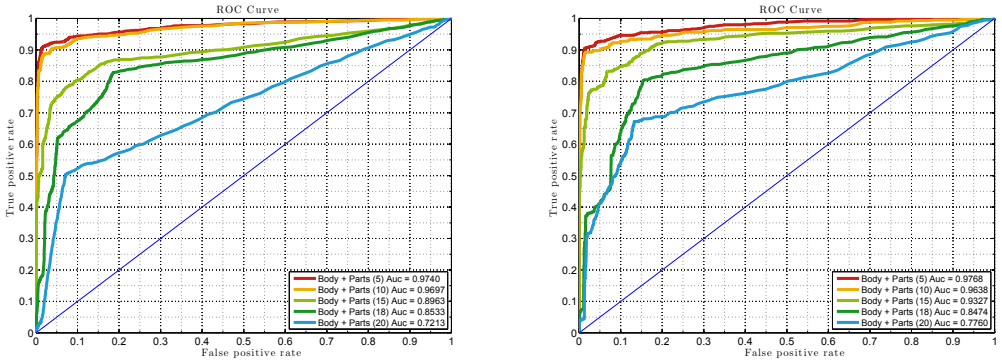
worse than the full-body approach for samples on which most or all of the pedestrian is visible. That being said, detection rates of largely occluded pedestrian are still not comparable with full body samples.

Figure 4.35 represents the roc curves of each individual SVM trained on a part of the pedestrian. Only roc curves with an area under the curve $a_{uc} > thr$ are plotted.

4.7. Conclusions and Discussion

In this chapter a detection framework for pedestrian detection in FIR images is presented.

In section 4.2 the LSI FIR pedestrian detection dataset is presented. It is made from sequence of FIR images captured from a static or moving vehicle is urban environments. The images comes with annotations, where pedestrians are labeled as rectangular bounding boxes of constant aspect ratio. The database is intended to be used as benchmarking for new detectors of pedestrians in FIR images. It is also released with the idea to be useful for



(a) Roc curve of logic inference based on the fixed set of rules for different percentages of occlusion.

(b) Roc curve of logic inference of latent structure for different percentages of occlusion.

Figure 4.33: Roc curves for occluded pedestrians using a full model and part based detection based on logic inference. Legend states occlusion and area under the curve (auc).

training new descriptors that need context information and cannot be trained on a classic classification database, thus it is divided in a train and a test subsets.

The detection framework and the evaluation methodology used to evaluate the pedestrian detectors are also discussed in this chapter. From the results it can be concluded that FIR images contain useful information for the task of detecting pedestrian. Even in challenging images of the detection dataset, such as the ones captured on hot summer days, the presented descriptors achieve high detection rates. These results may lead to reconsidering the role assigned to FIR cameras, as night vision devices. A detection system that is independent of external illumination conditions, and that is able to properly detect pedestrians both in day or night, is a very useful addition to an ADAS system. The experimental study of detection performance in full-size images suggest that there is correlation between the per-window results of the classifiers and their per-image performance.

The results presented in the evaluation section has led to a methodology that addresses the issue found on detection of small pedestrians. After resizing the image, in order to find small pedestrians, those appear with poor detail and spread borders, which makes their appearance quite different from that of a larger pedestrian. The propose solution is to computed the HOPE descriptor using a shifted version of log-Gabor filters that approximate the appearance of small pedestrian to that found on pedestrians in the central scale of the image pyramid. The application of the described method improves detection rates at all values of FPPI.

Computing the phase congruency of the images in the scale pyramid takes a large fraction of the time needed to compute the descriptor. In this chapter, a method for reducing that time is presented. In stead of computing it on the up-sampled images of the pyramid, it is processed on the central scale by using a new set of shifted log-Gabor filters, and then resampled. The results of this approximation closely match the ones of the original HOPE descriptor, while considerably reducing the computation time.

This chapter also addresses two important topics on pedestrian detection in images:

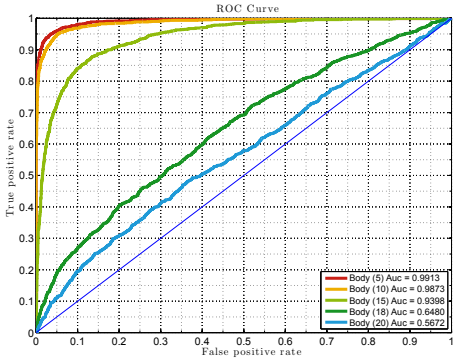


Figure 4.34: Roc curve of full body classification for different percentages of occlusion. Legend states occlusion and area under the curve (auc).

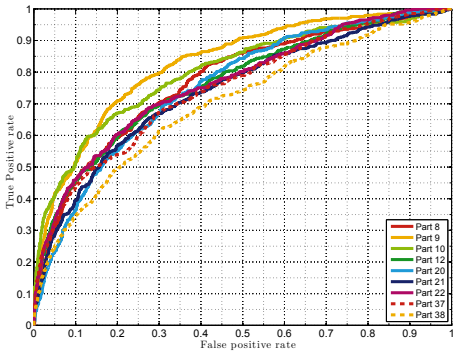


Figure 4.35: Roc curves of the best performing part classifiers. Only parts classifiers with an area under the curve $a_{uc} > thr$ are plotted.

ROI generation and occlusion handling. The presented ideas in these areas constitute initial research, and will be further developed. In the section of ROI generation, two methods for selecting interesting parts of the images have been described. In the first one, pedestrians are segmented by their apparent temperature. The second is based on edge density. From the phase congruency computed at the central scale, areas that do not hold enough detail are discarded. The main purpose of a ROI algorithm is to reduce the computation time of the overall detector. An evaluation of the computation time has been stablished as future research. That same section also propose a method for improving the classification of occluded pedestrians. A Markov Logic Network is used to infer the presence of a pedestrian based on the responses of a full-body descriptors and its parts. The initial results presented suggest that this method improves detection in largely occluded pedestrians. However, as future research, a through evaluation of its merits should be done.

Other topics that will be addressed in future research are the following. In [58] it is demonstrated that features, such as histograms of orientation can be calculated in one image and then be approximated to nearby scales. This idea can be applied to the phase congruency scale approximation presented in this chapter. The resulting algorithm would then not need to resample the approximated phase congruency to calculate the descriptor. In stead of that, the descriptor would be computed in the approximated phase congruency image and then be approximated to the equivalent descriptor at a different scale. This would make unnecessary to resample the image, and thus the time to process the image pyramid could be greatly reduced. This procedure would be specially useful in the case of the Int-HOPE descriptor, which can benefit from both techniques. Finally, a new line of research will be to combine different detector, that achieves better detection rates for different pedestrian size.

5

Tracking

Tracking is defined as the identification of a particular pedestrian in a sequence of images. There are many advantages to the use of a tracking step in any pedestrian detection system. First, the pedestrian's future trajectory can be anticipated, a concept of great interest to the topic of ADAS. By integrating the movement information from a set of consecutive frames it is possible to infer the speed and direction of the pedestrian. The position of the pedestrian in the immediate future may be used as a preemptive warning to the driver. Furthermore, the position of the pedestrian can be refined by filtering abrupt variations in its trajectory.

Tracking can also be used to improve the detection performance. False positives happen when an object, that may resemble the shape of a pedestrian, is incorrectly identified as being one. Those mis-detections are usually isolated and are not repeated in successive frames. A tracking algorithm can disregard not recurring detections, reducing the number of false positives. Another benefit of using a tracking algorithm is the reduction of false negatives. A detector may sporadically fail to correctly identify a particular pedestrian in a number of frames. A tracking algorithm may be able to infer that it is still there, thus reducing the number of false negatives. Similarly, a pedestrian being momentarily occluded behind an object is still of interest and should be detected. Without tracking there is no way to tell those occluded pedestrians are really there.

Figure 5.1 shows a pedestrian detected in a subset of images captured in a sequence. Every detection is displayed as a bounding box that enclose the body of the pedestrian. In Fig. 5.2 the predictions of the pedestrians after updating are shown. Notice that each individual pedestrian is enclosed in a bounding box with a persistent color. In the fourth frame a occluded pedestrian is still labeled, though there has not been a positive detection for that image. In the same image, an incomplete pedestrian, falling partially outside the image is still detected.

Tracking involves estimating the state of the position of a pedestrian from measurements in successive images. The kalman filter uses a series of measurements observed over time. It can also cope with noisy measurements and mis-detections. The resulting predictions of the state tend to be more accurate than an isolated measurement. A thorough explanation of the kalman filter can be found in [212]. In this section the kalman filter is used to track the coordinates of the bounding box of the detected pedestrian in the previous step. Other tracking methods, such as Particle Filters have been considered but discarded due to their computational complexity.

Figure 5.1: Detections

Figure 5.2: Predictions

5.1. Kalman Filter Variables

Initial conditions The state $\hat{x}$ is a six-dimensional gaussian vector with covariance P, made up from the coordinates of the upper left corner of the bounding box and its height (u, v, h) and their velocities $(\delta u, \delta v, \delta h)$ as stated in equation 5.1. The initial state is the coordinates of the first detected bounding box with zero velocity (equation 5.2). Initially the covariance matrix of the state is heuristically set to be equation 5.3.

$$\hat{x} = \begin{bmatrix} u & \delta u & v & \delta v & h & \delta h \end{bmatrix} \tag{5.1}$$

$$x_0 = \begin{bmatrix} u_0 & 0 & v_0 & 0 & h_0 & 0 \end{bmatrix} \tag{5.2}$$

$$P_0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.3}$$

5.1.1. Time Update

5.1.1.1. Static Model

Tracking pedestrians from a moving vehicle imply tracking their position in the image, which changes due to two factors: the movement of the pedestrian and the ego-motion of the vehicle. The latter is usually unavailable in pedestrian datasets. If that is the case, the following assumption has to be made: if the acquisition time is small enough, the motion of the vehicle between two consecutive measurements is accounted for in the measurement noise. As such, the time update model is set to be:

$$\hat{x}_k^- = A\hat{x}_{k-1} + w_k \tag{5.4}$$

Where A is the state transition model (equation 5.5). Assuming constant velocity, the position in time k is the one in $k-1$ plus its increment between updates.

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{5.5}$$

5.1.1.2. Motion Model

In this section, the time update process using IMU information and the Unscented Transform (UT) is explained.

For every time increment $k-1|k$ the state of the filter is updated. There are two reason the state changes between two measurement: the motion of the pedestrian and the motion of the vehicle. This movement may be simplified as a roto-translation transformation (ϕ_{rt}) of the coordinate axis of the vehicle on the xy plane: a translation Δt_y in the y and a rotation $\Delta\theta$ around the yaw axis. The motion model is represented in Fig. 5.3. The translation and rotation are measured with an IMU device.

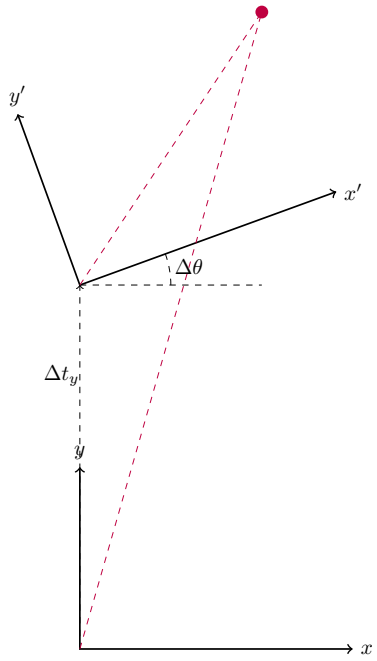


Figure 5.3: Representation of the movement of the vehicle between two consecutive frames. The perspective of the object, represented as a circle, changes due to a roto-translation of the camera.

After the roto-translation the coordinates of the bounding box have to be updated. Equation 5.6 is the new state after applying the transformation ϕ_{rt}, where u_c (equation 5.7) and u_v (equation 5.8) are the updated coordinates of the upper left corner of the bounding box. Since the height of the pedestrian is unknown, the height of the bounding box is set to be the same as before plus δh. This approximation holds for the tested sequences, while driving in urban environments.

$$\hat{x}'_{k-1} = \phi_{rt}(\hat{x}_{k-1}) \tag{5.6}$$

$$u_c = \frac{({c_u}^2 f_v + {f_u}^2 f_v - c_u f_v(u + (h/4)))\sin(\Delta\theta) + c_u \Delta t_y f_u (v + h - c_v) + f_u f_v u \cos(\Delta\theta)}{(c_u f_v - f_v(u + (h/4)))\sin(\Delta\theta) + \Delta t_y f_u (v + h - c_v) + f_u f_v \cos(\Delta\theta)} - \frac{w}{2} + \delta u \tag{5.7}$$

$$v_c = c_v - h - \frac{f_v}{\frac{f_u\, f_v\, \cos(\Delta\theta)}{c_v\, f_u - f_u\, (v+h)} - \Delta t_y + \frac{f_v\, \sin(\Delta\theta)\, (c_u - (u + (h/4)))}{c_v\, f_u - f_u\, (v+h)}} + \delta v \tag{5.8}$$

The overall time update is defined in equation 5.9, where k is the update step, B is the control model and u is the control input. The process noise w is a gaussian vector with mean zero and covariance Q (equation 5.10), where q is heuristically set to $q = 10^{-4}$. There is no control over the movement of the pedestrian, thus the control input is set to $c_{k-1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T$.

$$\hat{x}_k = \hat{x}'_{k-1} + Bc_{k-1} + w_{k-1} \tag{5.9}$$

$$Q = \begin{bmatrix} q & 0 & 0 & 0 & 0 & 0 \\ 0 & q & 0 & 0 & 0 & 0 \\ 0 & 0 & q & 0 & 0 & 0 \\ 0 & 0 & 0 & q & 0 & 0 \\ 0 & 0 & 0 & 0 & q & 0 \\ 0 & 0 & 0 & 0 & 0 & q \end{bmatrix} \tag{5.10}$$

The described time update transformation is highly non-linear, which breaks one of the conditions to use the Kalman Filter equations. The Unscented Kalman Filter (UKF) [108] extends the general Kalman filter to non-linear transformations of a random variable without the need of linearization, as the Extended Kalman Filter (EKF) does [206]. The UKF achieves better results than the EKF for highly non-linear transformations with approximately the same computational demands.

The Unscented Transformation propagates the random variable across the non-linear system using a minimal set of deterministically chosen weighted sigma points. The mean and variance of the transformed variable are accurate up to the second order of Taylor series expansion.

For an augmented random variable of dimension n with mean $\bar{x}$ and covariance P the sigma points χ are in equation 5.11.

$$\begin{aligned} \chi_0 &= \bar{x} \\ \chi_i &= \bar{x} + \sqrt{(n+\lambda)P} && i = 1, \ldots, n \\ \chi_i &= \bar{x} - \sqrt{(n+\lambda)P} && i = n+1, \ldots, 2n \end{aligned} \tag{5.11}$$

where $\lambda = \alpha^2(n + \kappa)$ is an scaling factor that determines how much spread are sigma points around the mean $\bar{x}$. In this case the values of α and κ are heuristically set to $\alpha = 0.01$ and $\kappa = 200$.

The selected weighted sigma points are propagated though the non-linear function f and the mean and covariance of the state are approximated.

For each sigma point, two weights are calculated, w^c and w^m in equation 5.13.

$$w_0^c = \frac{\lambda}{n+\lambda} + 1 - \alpha^2 + \beta \tag{5.12}$$

$$w_0^m = \frac{\lambda}{n+\lambda} \tag{5.13}$$

$$w_i^m = w_i^c = \frac{1}{2(n+\lambda)}$$

Where $\beta = 2$, as noise is initially considered to follow a Gaussian distribution.

Once the selected sigma points are propagated though the non-linear function f (equation A.31), weights w^m are used to approximate the mean (equation A.32) and w^c to approximate the covariance (equation A.33) of the state.

$$\gamma_i = f(\chi_i) \qquad i = 0, \ldots .2n \tag{5.14}$$

$$\hat{x}_{k|k-1} = \sum_{i=0}^{2n} w_i^m \gamma_i \tag{5.15}$$

$$P_{k|k-1} = \sum_{i=0}^{2n} w_i^c [\gamma_i - \hat{x}_{k|k-}][\gamma_i - \hat{x}_{k|k-1}]^T \tag{5.16}$$

5.1.2. Measurement Update

The measurement update equation is 5.17. The measurement $\hat{z}$ is the three-dimensional vector of the bounding box location (u, v, h), where (u, v) are the coordinates of the upper left corner of the bounding box and h is the height of the box, measured in pixels.

$$\hat{z}_k = H\hat{x}_k + v_k \tag{5.17}$$

The measurement model, denoted as H in equation 5.18, remaps the six-dimensional state vector to the three-dimensional measurement vector. The measurement v is a gaussian vector with zero mean and covariance R (equation 5.19), where r is heuristically set to $r = 4$.

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{5.18}$$

$$R = \begin{bmatrix} r & 0 & 0 \\ 0 & r & 0 \\ 0 & 0 & r \end{bmatrix} \tag{5.19}$$

The remaining steps of this approach follows the same equations as the original Kalman Filter. For brevity, those equations are stated in section A.1.3.

5.2. Detection Matching

A detection may be considered to be the tracked pedestrian if it minimizes the Munkres' assignment algorithm [152]. The cost of assigning a detection to a kalman filter is defined by the square Mahalanobis distance between the state and the measurement in equation 5.20.

$$d(z) = (z - Hx)^T \cdot \frac{1}{HPH^T - R}(z - Hx) \tag{5.20}$$

The cost of assignment is thresholded, so if any measurement has a distance $d(z_i) > \text{thr}$ it is set that $d(z_i) = \infty$. In this implementation the threshold is set to $thr = 20$. After assignment, three sets of predictions are generated:

- Pr: is the set of detections that have been matched with any of the tracked pedestrians.
- U: is the set of kalman filters that have not found a match among the last set of measurement. This set does not undergo a measurement update. The fact that there has not been a positive match may be due to an isolated mis-detection or to the disappearance of the pedestrian from the field of view of the camera. An internal counter of every member of this set is incremented by one. If it reaches a maximum value of k_u it is removed from tracking. If it is again found in future assignment steps, that counter is set to zero. In the experimental section of this chapter this maximum number of consecutive mis-detections is set to $k_u = 5$.
- N is the set of detections that have not found a match among the kalman filters. This may be due to a new pedestrian or to a false positive. A new kalman filter is created for every member of the N set. However, those kalman filters are not yet assigned the designation of *pedestrian* but instead are labeled as *uncertain*. The kalman filter is moved to the Pr set once it is detected in a minimum of k_n consecutive frames. In the experimental section of this chapter this minimum number of consecutive mis-detections is set to $k_n = 5$.

5.3. Experimental Results

The presented tracking algorithm has been tested in five sequences of the LSI database. The experimental results section is divided into several special cases: multiple non-occluded

pedestrians, multiple occluded pedestrians and tracking of pedestrians from a moving vehicle. The evaluation methodology used in this section is the same as the one described in chapter 4. A detection is considered positive if it overlaps in more than $ov = 0.5$ with an annotated ground truth bounding box. The performance of the overall pedestrian detectors is assessed by comparing the Precision-Recall (PR) curves of the detectors with an without the tracking algorithm. For visualization purposes, the detections are projected and plotted on the xy plane of the ground. Raw detections from the detector are plotted as blue dots. Predictions of the filter are plotted with different colors, one for each individual pedestrian.

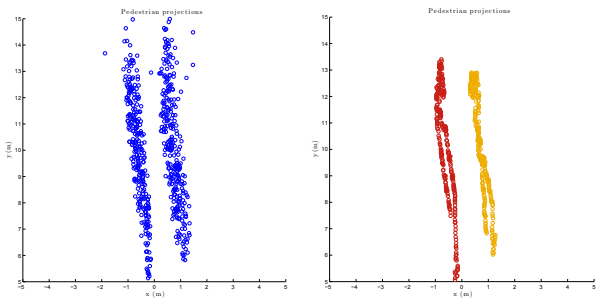
5.3.1. Non-occluded Pedestrians

Sequence #1 Figure 5.4 shows a subset of sequence #1. In it, two pedestrians walk heading getting further from the camera and eventually turning around and coming back. The detections, projected on the ground plane are plotted in Fig. 5.5.

(a) Frame 150 (b) Frame 200 (c) Frame 250

(d) Frame 300 (e) Frame 350 (f) Frame 400

Figure 5.4: Samples of tracking test sequence # 1



(a) Unfiltered tracks (b) Filtered tracks

Figure 5.5: Pedestrian detections projected on the ground plane (sequence # 1).

In Fig. 5.6 the PR curves for tracked and untracked detections are plotted. It should be noted that there are not precision values for all possible recall values. This is because the threshold for a detection to be considered a pedestrian is set deliberatively high. Though the miss rate is rather high, the number of false positives is set to a low value. The tracking algorithm removes the remaining false positives, which appear in a non-recurring fashion. The tracking algorithm is also able to infer the presence of the pedestrian, disregarding the sporadic mis-detections. Interestingly, for this particular sequence the average precision after the tracking algorithm is $AvPR = 1$.

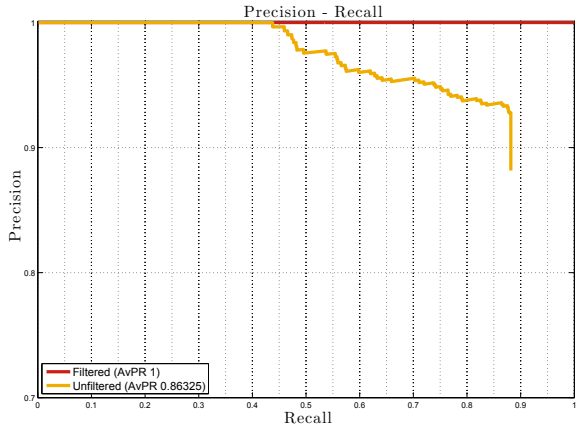


Figure 5.6: Pedestrian detections projected on the ground plane (sequence # 1).

Sequence #2 Figure 5.7 shows a subset of sequence #2. This particular sequence was shot in daylight on a hot day. It should be noted that the number of false positives is greater than in sequence #1, as can be seen on the left side of figure 5.8. The area where there is a greater density of false positives match an area of the image under direct sunlight. The high decision threshold means that difficult pedestrians are missed, as can be seen in Fig. 5.7e.

In Fig. 5.6 the PR curves for tracked and untracked detections are plotted. The application of the tracking algorithm actually make results worse for low recall values, due to frequent mis-detections appearing in the same area.

5.3.2. Occluded Pedestrians

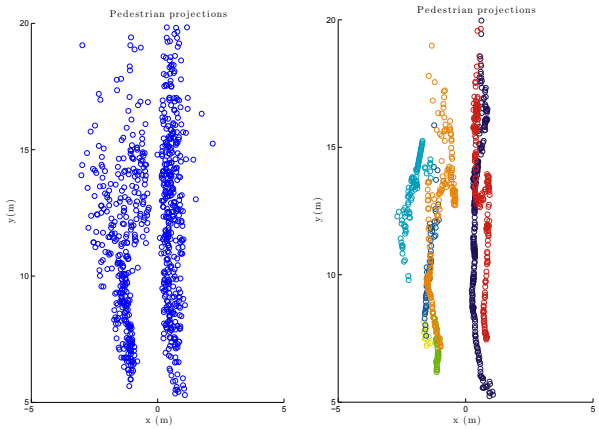
In this subsection a more challenging scenario, pedestrian tracking under occlusion, is considered. The two following examples contains pedestrian crossing in front of each other.

Sequence #3 Figure 5.10 shows a subset of sequence #3, shot in front of a busy zebra cross, with pedestrians crossing the street in both directions. From the filtered tracks in Fig. 5.11 it is noticeable that, though the algorithm is able to track each individual pedestrian, their positions in the 3d world are not accurate. Though by analyzing the sequence it is

(a) Frame 300 (b) Frame 400 (c) Frame 500

(d) Frame 600 (e) Frame 700 (f) Frame 800

Figure 5.7: Samples of tracking test sequence # 2



(a) Unfiltered tracks (b) Filtered tracks

Figure 5.8: Pedestrian detections projected on the ground plane (sequence # 2).

evident that the pedestrians follows a rectilinear path, the projected detections have errors of several meters.

Figure 5.12 shows the PR curves for the filtered and unfiltered detections. It shows that the precision remain higher for most of the recall range in the filtered sequence.

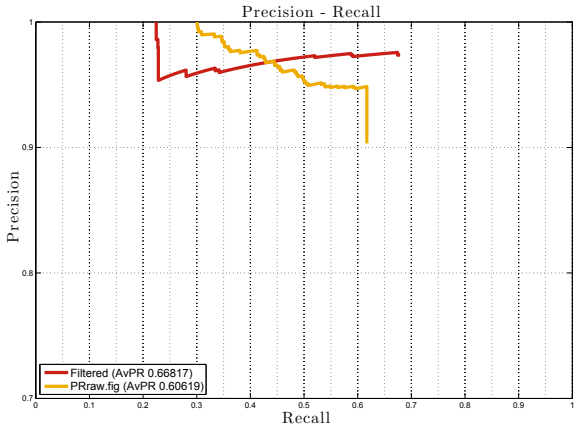
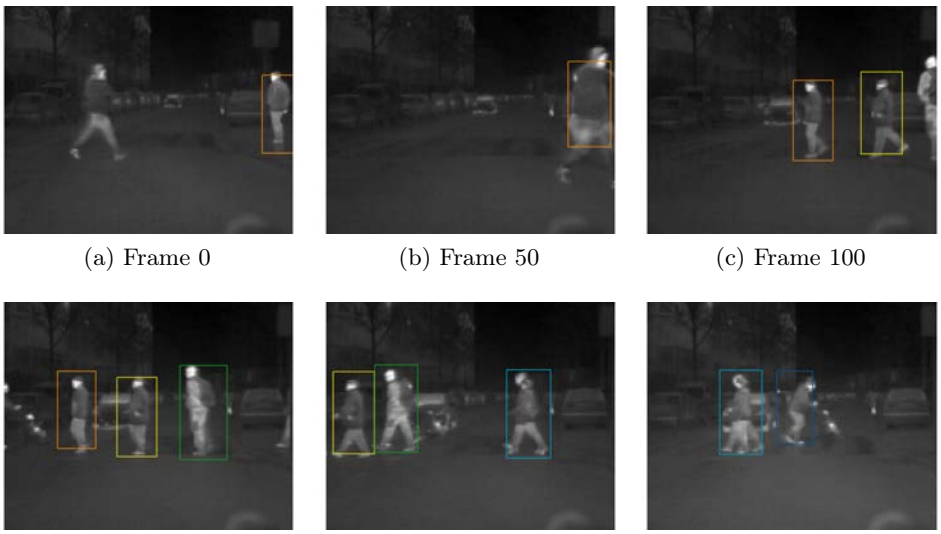


Figure 5.9: Pedestrian detections projected on the ground plane (sequence #2).



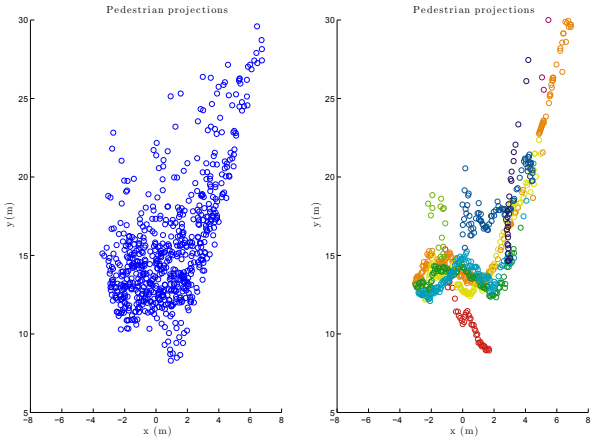
(a) Frame 0 (b) Frame 50 (c) Frame 100

(d) Frame 150 (e) Frame 200 (f) Frame 250

Figure 5.10: Samples of tracking test sequence # 3

Sequence #4 Figure 5.13 shows a subset of sequence #4. In it two pedestrians walk in complex trajectories falling several times out of the field of view of the camera.

The PR curve in Fig. 5.15 shows that filtering the detection allows for every pedestrian to be detected, as there is a value of precision for a value of recall of one. However, the overall precision rate degrades. This is due to the filter following the pedestrians once they leave the field of view of the camera. Each tracker allows for the pedestrian to be missing for k_u frames. The ground truth bounding boxes are only annotated for visible pedestrians, so predictions falling outside the image are considered as false positives.



(a) Unfiltered tracks (b) Filtered tracks

Figure 5.11: Pedestrian detections projected on the ground plane (sequence # 3).

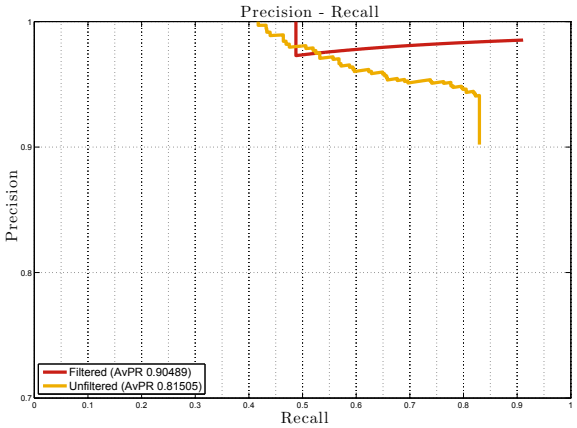


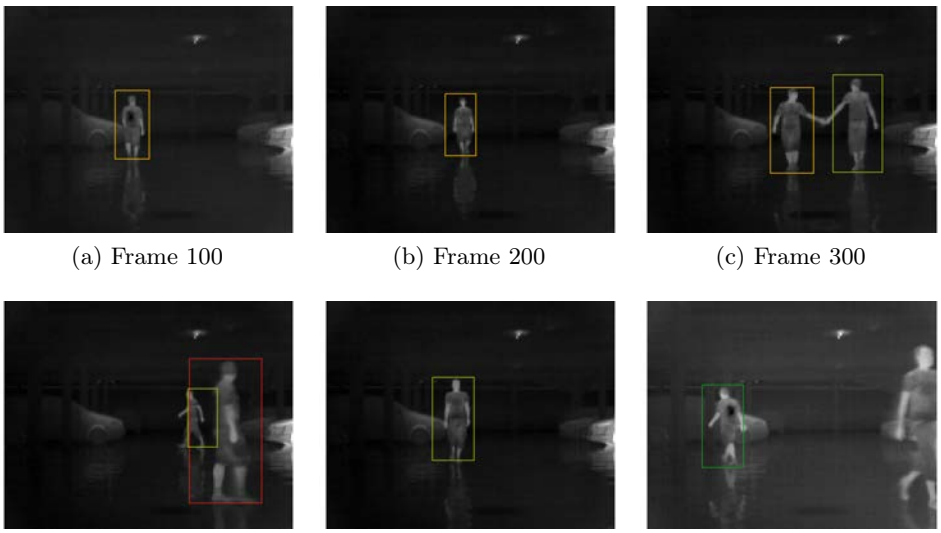
Figure 5.12: Precision-Recall curves of filtered and unfiltered tracks (sequence # 3).

5.3.3. Motion Model

Sequence #5 Figure 5.16 shows a subset of sequence #5. In it, a pedestrian walking towards the camera is tracked from a moving vehicle.

Figure 5.17 shows the raw detections as well as the predictions of the filter using the static and the dynamic model. In Fig. 5.17a there can be seen large gaps in the pedestrian position. Those are not so evident after applying the static filter (Fig. 5.17b). When the pedestrian is close to the camera, the displacement of the detection window is more pronounced between two consecutive frames.

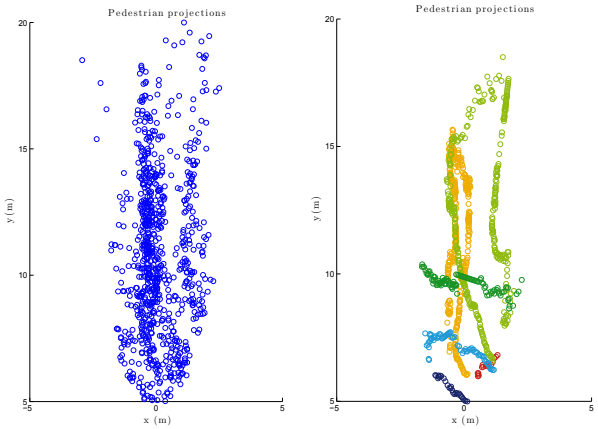
In Fig. 5.18 the PR curve of sequence #5 is shown. The overall precision of the filtered detections is higher than the unfiltered ones. It is to be noted that the average precision for the static and motion models is the same. While driving at low speeds in urban environments the movement of the vehicle between two consecutive frames is small enough for the static model to apply. As seen in Fig. 5.17 the static motion filter *loses*



(a) Frame 100 (b) Frame 200 (c) Frame 300

(d) Frame 400 (e) Frame 500 (f) Frame 600

Figure 5.13: Samples of tracking test sequence # 4



(a) Unfiltered tracks (b) Filtered tracks

Figure 5.14: Pedestrian detections projected on the ground plane (sequence # 4).

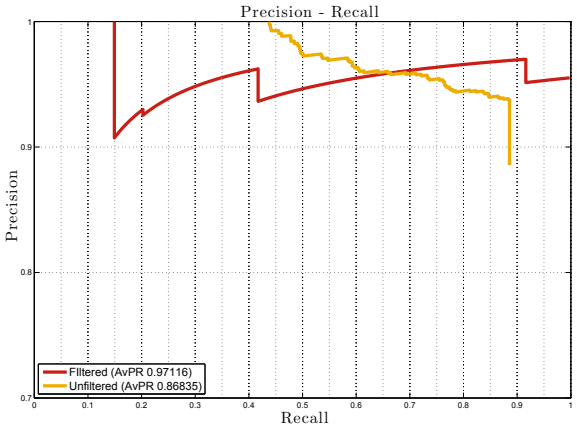
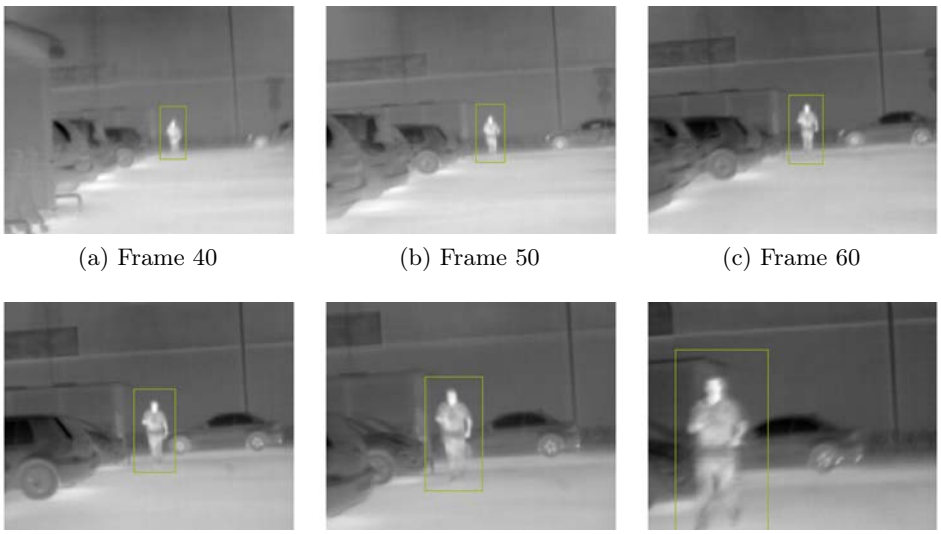


Figure 5.15: Precision-Recall curves of filtered and unfiltered tracks (sequence # 4).



(a) Frame 40 (b) Frame 50 (c) Frame 60

(d) Frame 70 (e) Frame 80 (f) Frame 90

Figure 5.16: Samples of tracking test sequence # 5

the pedestrian and a new filter is created for the same pedestrian. However, there is a correct detection for all intermediate frames because, though the location of the detection bounding box is slightly misplaced, it still overlaps with the ground truth bounding box.

Sequence #6 Figure 5.19 shows a subset of sequence #6. The sequence was recorded from a moving vehicle and two special situations may be found in it. The first one is the approximation to a zebra cross, where multiple pedestrians are crossing the street or waiting to cross it on the sidewalk. While crossing, the pedestrians are occluded by other vehicles stopped at the zebra cross. The second part of the sequence involves driving around a roundabout. The lateral movement of the camera causes a *ghosting* effect on the images, lowering the detection accuracy of pedestrians walking on the sidewalk. In Fig. 5.20 the pedestrians detections are projected on the ground plane. The motion model is specially helpful in tracking pedestrians while the vehicle is driving around the roundabout. While the vehicle is driving straight, both models behave the same way.

In Fig. 5.21 the PR curves of sequence #6 are shown. The average precision is highest for the motion model filter. However it is evident that, in any case, the average precision is low. This is due to two reasons. First, pedestrians remain occluded for a long time, while the ground truth annotations assert that they are still there. Secondly, there are a large number of small pedestrians that are mis-detected due to the *ghosting* effect caused by the lateral movement of the camera.

5.4. Conclusions

In this section the tracking step of the pedestrian detection algorithm has been presented. An Kalman FIlter has been used to track moving pedestrians from a static or moving vehicle, based on the coordinates of the bounding boxes generated by the detector. Only detections with a high SVM score are used to track pedestrians, which increases the number of false negatives. However, experiments shows that the tracking step of the algorithm is able to successfully track a pedestrian, even though there may be isolates mis-detections. The tracking algorithm is also able to filter out false positives, by disregarding erratic detections. In some occasions, this approach may also degrade the detection performance. A pedestrian is removed from the tracking stack if there is no matching detection in a number of consecutive images. Before it is removed, the tracking algorithm is inferring its position from previous information, which may no longer apply. Also, for a detection to be included in the pedestrian stack, there needs to be a minimum amount of consecutive detections. This keeps the number of false positives low, but also increases the number of false negatives. These issues may be solved by applying a different methodology to the evaluation of the results. In ADAS systems, the most important factor is to fed the driver with pertinent and timely information. An study on reaction times of the driver is proposed as a future work. This study will allow for a refined evaluation methodology, where a detection will be considered correct if the information provided to the driver is useful in preventing an accident, and incorrect if there is no need for the driver to act.

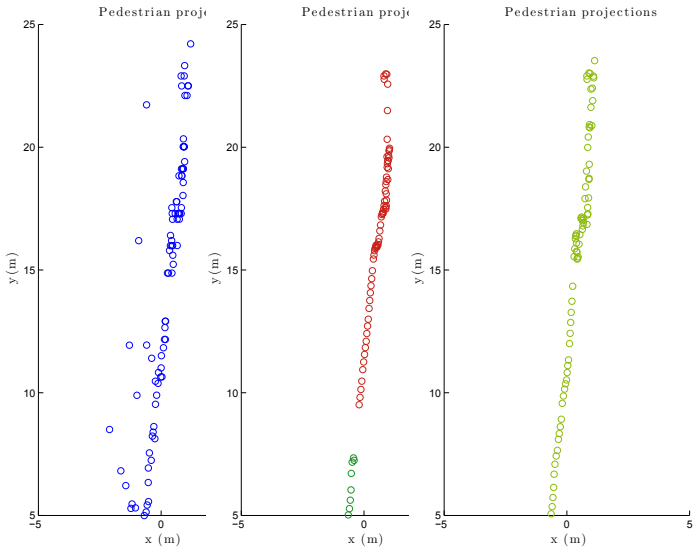


Figure 5.17: Pedestrian detections projected on the ground plane (sequence # 5). a) Unfiltered tracks; b) Filtered tracks (Static Model); c) Filtered tracks (Dynamic Model).

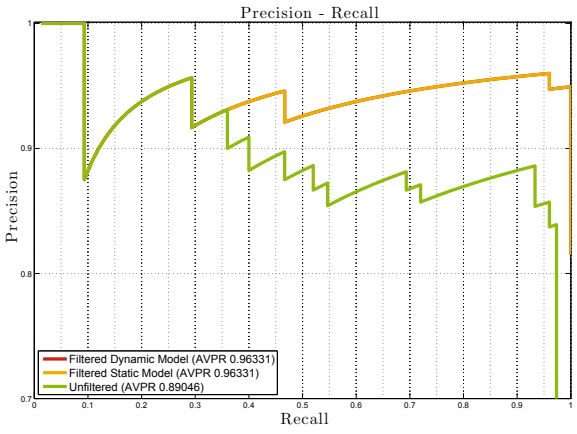
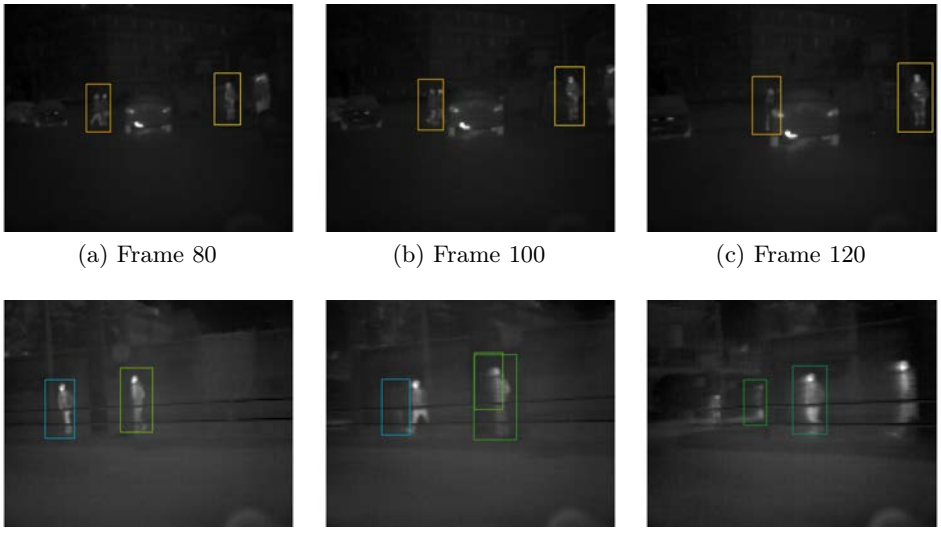


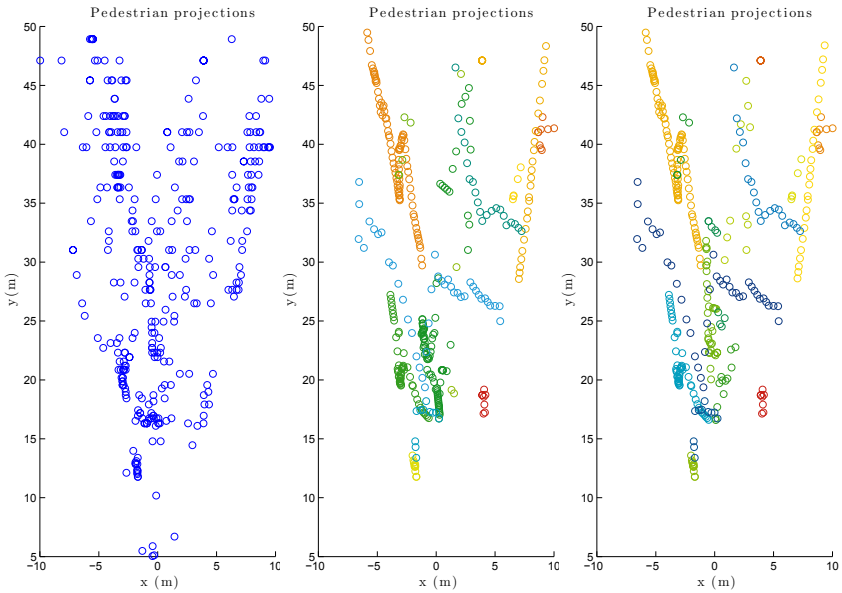
Figure 5.18: Pedestrian detections projected on the ground plane (sequence # 5).



(a) Frame 80 (b) Frame 100 (c) Frame 120

(d) Frame 620 (e) Frame 640 (f) Frame 660

Figure 5.19: Samples of tracking test sequence # 6



(a) (b) (c)

Figure 5.20: Pedestrian detections projected on the ground plane (sequence # 6). a) Unfiltered tracks; b) Filtered tracks (Static Model); c) Filtered tracks (Dynamic Model).

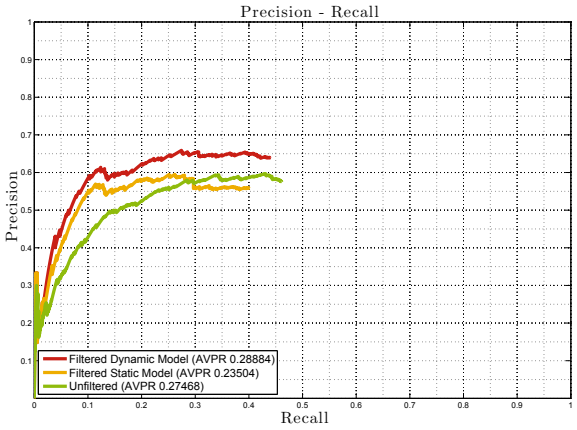


Figure 5.21: Pedestrian detections projected on the ground plane (sequence # 6).

6

Conclusions and Future Work

Pedestrian detection in computer vision is a challenging task. Though it is a research field that can be dated a few decades back, the interest on it is still growing. There are a large number of topics involved, some which are yet not quite solved. Pedestrians present an extraordinary variation in shape, size, clothing, and temperature. The development of classification schemes that can efficiently separate pedestrian from background is one of the more active areas. Others include occlusion handling, tracking and machine learning algorithms.

This thesis is framed within the topic of pedestrian detection in FIR images, which is a relatively new area in this field; one is not as developed as computer vision using VL imagery. Both present similarities: the same variability of shapes, sizes and distances to the camera are present. However, there are also few key differences the most important of which is the illumination need. The image from a FIR camera present magnitude that is proportional to the temperature of the scene. As such, there is no color information and texture is less noticeable but have the great advantage of being able to work without external illumination. There are also drawbacks to the application of FIR cameras to pedestrian detection in ITS. The most immediate of which is budget-related. Though the prices of these kinds of cameras are getting lower they are still expensive, when compared with a VL camera. For the purpose of object recognition there are a number of issues derived from the use of these cameras, namely, the relation between contrast of the image and ambient temperature, or the blurring effect seen on hot objects due to camera motion and the clothing, which also adds variability to the appearance.

The objective of a pedestrian detection in an ADAS scheme is to alert the driver of dangerous situations. The driver would then have more reaction time, and an accident may be avoided.This thesis aims to provide a system for analysis of the driving environment capable of detecting the presence of pedestrians by means of a micro-bolometer camera, with sensitivity in the far infrared range of the spectrum. An artificial perception system identifies pedestrians in front of the vehicle and determines whether there is any risk that endangers the integrity of pedestrians as well as passengers. In practice, the mere presence of a pedestrian in the field of view of the camera is considered as a dangerous sign.

The structure of this document follows one of the most used detection paradigms. It is divided in classification techniques, the application of those techniques to full-sized images and temporal tracking of the detected pedestrians. It also contains an evaluation of the techniques in the state of the art, which focus on techniques developed in the topic of pedestrian detection, be it using FIR or LV images.

The term classification, in this work, has been loosely used to express the combination of a descriptor and a machine learning algorithm to give a similarity score to a cropped image that may, or may not contain a pedestrian. The first work developed in this topic has been a variation of the probabilistic template scheme for pedestrian classification in FIR images. As discussed before, the output of a FIR camera based on a microbolometer changes with variations of its internal temperature. In order to assess if a cropped sample contains a pedestrian or any other object, a probabilistic template based on temperature is used. The main purpose of this method is to achieve invariance to ambient temperature. Though achieving surprisingly good results, for such a straightforward method, the next part of the classification chapter focus on the development of a descriptor with invariance to contrast but that also defines the shape of a pedestrian in a more descriptive manner.

The new descriptor, which has been dubbed as HOPE, has its foundation on the histograms of gradients descriptors, which was originally though to work on VL images, although it has been probed that its application to FIR images is also possible. This descriptor addresses contrast invariance by applying a normalization of neighboring histograms. The proposed descriptor addresses two requirements of a pedestrian detection in FIR images. In the first place, the descriptor follows the scheme of encoding the shape of the underlying object as a grid of histograms of orientation, thus encoding the shape in a descriptive manner. The second objective is to make it invariant to contrast and changes in illumination. Instead of using the gradient of the image, a different feature should be used, one that is intrinsically invariant to contrast. The phase congruency feature has been chosen because, unlike gradient or local energy, this feature will render the same magnitude for the shape of an object in two images with different illumination. Particularly, the values are alway in the $[0-1]$ range, where a point where all its Fourier component are in phase will produce a 1. A point where none are in phase will produce a 0. This response is repeatable across different images of the same object captured under varying conditions of contrast or illumination. The various parameters on which the descriptor relies have been tested and discussed, focusing on the most relevant ones.

One of the contributions to this work is the LSI Far Infrared Pedestrian Dataset. This collection of images is released with the hope that it may be useful to develop new algorithms for pedestrian detection in FIR images. The database contains a large number of labeled images, where each pedestrian is represented as rectangular box of fixed ratio. It has been recorded at different location, and with different illumination and temperature conditions, both statically and from a moving vehicle. The sequences captured with the car moving have been taken in real urban traffic environments. These include different actions, for example, pedestrian crossing traffic lights, other vehicles on the road and occluded pedestrians walking on the sidewalk. Regarding pedestrian databases, different sources of information may increase the performance of a detector, by fusing data from multiple sensors. As future work, the LSI Pedestrian Dataset will be expanded to include data from other sensors, such a visible cameras, stereoscopic systems (both VL and FIR), rangefinder measurements, and contextual information, such location of the vehicle (GPS) and accelerations (IMU).

The presented database was used to compare different descriptors, which are commonly used in VL images. The experiments suggest that the approaches based on histograms of

orientation work best, for the samples used.

This chapter also includes an evaluation of simple features that can be computed using the integral image paradigm. These evaluation relies on the paradigm of the random forests classifiers. The integral features used were combined with each other to train a set of classifiers. From the results, it has been concluded that phase congruency is a relevant feature on its own, and outperforms gradient or gray-level. In some of the experiments it even shows better performance than histograms of un-normalized gradients. The results are based on two sets of experiments: non-overlapping square features and random rectangular features. By comparing both of them it is clear that the election of a grid of square features captures the shape of the pedestrian precisely. Adding a large number of random features do not significantly increase the performance.

The results presented in this chapter suggests that FIR images are a very useful source of information for pedestrian classification with advantage in low visibility applications. However, a qualitative inspection of misclassified samples suggests that there are some issues to be taken into account in future research, namely, motion blur, pose variation and occlusion handling.

This work proceeds to a detection framework, where the ideas developed in the classification chapter are put into practice in full-size images. The chapter presents the evaluation methodology followed to asses the performance of the presented descriptors. From the results it can be concluded that FIR images contain useful information for the task of detecting pedestrian. Even in challenging images of the detection dataset, such as the ones captured on hot summer days, the presented descriptors achieve high detection rates. These results may lead to reconsidering the role assigned to FIR cameras, as night vision devices. A detection system that is independent of external illumination condition, and that is able to properly detect pedestrians both in day or night, serves is a very useful addition to an ADAS system. The experimental study of detection performance in full-size images suggest that there is correlation between the per-window results of the classifiers and their per-image performance. As proof of concept, the Latent-SVM detector has been tested in the detection dataset. This detector trains the models in two parts: first positive samples are warping, creating a much larger training dataset. This approach help the classifier to adapt to pose variation. In the second part, hidden structures of parts are searched in a double-resolution version on the descriptor. This methodology is generic in the sense that, a wide range of descriptor can benefit from it, specially if they encode shape as local histograms. The experiments performed showed that this approach can be successfully be adapted to work on, and also to use another descriptor, FIR images. The results, however suggest, that the performance of this classifier should increase, be the images larger. As future work, the evaluation of this approach on a larger-image database is proposed.

The results presented in the evaluation section has led to a methodology that addresses the issue found on detection of small pedestrians. After resizing the image, in order to find small pedestrians, those appear with poor detail and spread borders, which makes their appearance quite different from that of a larger pedestrian. The proposed solution is to compute the HOPE descriptor using a shifted version of log-Gabor filters that approximate the appearance of small pedestrian to that found on pedestrians in the central scale of the

image pyramid. The application of the described method improves detection rates at all values of FPPI. Computing the phase congruency of the images in the scale pyramid takes a large fraction of the time needed to compute the descriptor. In this chapter, a method for reducing that time is presented. In stead of computing it on the up-sampled images of the pyramid, it is processed on the central scale by using a new set of shifted log-Gabor filters, and then resampled. The results of this approximation closely match the ones of the original HOPE descriptor, while considerably reducing the computation time. The integral features presented in the classification chapter be calculated in one image and then be approximated to nearby scales. This idea can be applied to phase congruency scale approximation presented in this chapter. The resulting algorithm would then not need to resample the approximated phase congruency to calculate the descriptor. In stead of that, the descriptor would be computed in the approximated phase congruency image and then be approximated to the equivalent descriptor at a different scale. This would make unnecessary to resample the image, and thus the time to process the image pyramid could be greatly reduced. This procedure would be specially useful in the case of the Int-HOPE descriptor, which can benefit from both techniques.

This chapter also addresses two topics on pedestrian detection in images: ROI generation and occlusion handling. The presented ideas in these areas constitute initial research, and will be further developed in the future. In the section of ROI generation, two methods for selecting interesting parts of the images have been described. In the first one, pedestrians are segmented by their apparent temperature. The second is based on edge density. From the phase congruency computed at the central scale, areas that do not hold enough detail are discarded. The main purpose of a ROI algorithm is to reduce the computation time of the overall detector. An evaluation of the computation time has been stablished as future research. That same section also propose a method for improving the classification of occluded pedestrians. A Markov Logic Network is used to infer the presence of a pedestrian based on the responses of a full-body descriptor and its parts. The initial results presented suggest that this method improves detection in largely occluded pedestrians. However, as future research, a through evaluation of its merits should be done.

Finally, the work proceeds to the last stage of the proposed algorithm: temporal tracking of targets. A Kalman Filter has been used to track moving pedestrians from a static or moving vehicle, based on the coordinates of the bounding boxes generated by the detector. Only detections with a high SVM score are used to track pedestrians, which increases the number of false negatives. However, experiments shows that the tracking step of the algorithm is able to successfully track a pedestrian, even though there may be isolates mis-detections. The tracking algorithm is also able to filter out false positives, by disregarding erratic detections. In some occasions, this approach may also degrade the detection performance. A pedestrian is removed from the tracking stack if there is no matching detection in a number of consecutive images. Before it is removed, the tracking algorithm is inferring its position from previous information, which may no longer apply. Also, for a detection to be included in the pedestrian stack, there needs to be a minimum amount of consecutive detections. This keeps the number of false positives low, but also increases the number of false negatives. These issues may be solved by applying a different methodology to the evaluation of the results.

In ADAS systems, the most important factor is to fed the driver with pertinent and timely information. An study on reaction times of the driver is proposed as a future work. This study will allow for a refined evaluation methodology, where a detection will be considered correct if the information provided to the driver is useful in preventing an accident, and incorrect if there is no need for the driver to act.

A

Introduction to the Kalman Filter and its derivates

The Kalman filter (KF) is widely used in tracking and estimation tasks, given its simplicity and robustness. However, its relaibility depends heavily on the linearity of the model. Traditionally, if the the system present a large nonlinearity the *Extended Kalman Filter* (EKF) used instead, which is a linearization of the system around the working point, so the Kalman filter equations can be applied. However, experience shows that its implementation is complicated and is reliable only in a few cases. To compensate for the shortcomings of the EKF, *Julier and Uhlman* proposed in [108] the *Unscented Kalman Filter* (UKF). Besides being much more robust for high nonlinearities, system noise can be non-Gaussian, a major constraint imposed by the EKF. For an overview of the KF and the EKF refer to [212]

A.1. The Kalman Filter

A.1.1. Constraints

The Kalman filter creates a model of the system state that maximizes the posterior probability, given a series of measurements. In this case the *a posteriori* probability is the final probability, once all measurements from start to the present moments have been acquired. To apply it, the system must meet certain restrictions:

- Lineal: The system at time k can be expressed as a matrix multiplied by the state at time $k-1$. Nonlinear systems can not be expressed using matrix algebra. This condition is never fulfilled in practice, however the Kalman filter is effective for low nonlinearities of the system and measurement instruments.
- Measurement Noise: The measurements provided by the sensors always have an uncertainty involved. The Kalman filter imposes that the sensor noise has to be white, i.e. not correlated in time. That is, the noise level will not increase or decrease from one measurement to the next.
- Process Noise: The Kalman filter imposes that it has to be Gaussian. It is the most important restriction and usually the first that is not met. The Kalman filter is

based on the accumulation of information over time by multiplying the state, which is defined as a Gaussian distribution. If the noise is not Gaussian (or pretty close to it) the results would be unpredictable.

It should be noted here that some of these restrictions do not apply to UKF. One of its main advantages is that it allows systems with non-Gaussian noise. It is also adaptable to systems with high nonlinearity.

A.1.2. Principles

Given two measurements of the same variable, captured with two different sensors, these can be parameterized as a mean and variance associated with it, i.e. a Normal or Gaussian distribution. The mean is the measurement of each sensor, and the variance would be an estimate of the quality of the sensor. The better the sensor the smaller would be this variance. Figure A.1 shows two Gaussian distributions for two measurements with different mean and standard deviation.

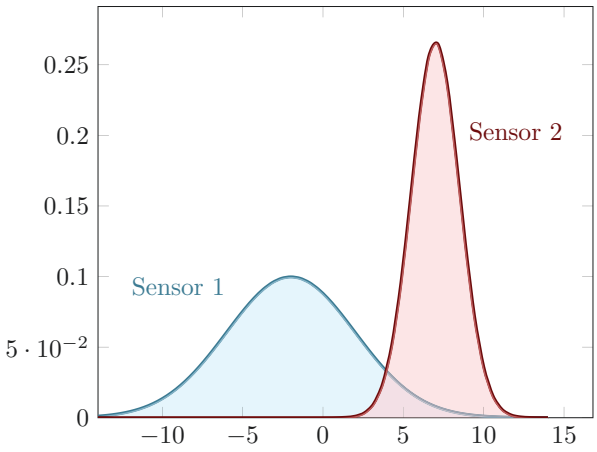


Figure A.1: Gaussian distributions, representing the measurements of two different sensors of the same variable.

A.1.2.1. One Dimensional Example

The probability distribution function (PDF) of a Gaussian in one dimensional space is defined in equation A.1

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma}\right)^2\right]} \tag{A.1}$$

Where the probability of x depends just on the mean ($\bar{x}$) and the variance (σ) of the PDF.

The fundamental property on which the Kalman Filter is based is that the combined probability of two Gaussian distributions is another Gaussian distribution. The equation A.2 is the combined probability distributions of the measurements of both sensors.

$$p_{12}(x) = \frac{1}{\sigma_1\sqrt{2\pi}} \cdot \frac{1}{\sigma_2\sqrt{2\pi}} \cdot e^{\left[-\frac{1}{2}\left(\frac{x-\bar{x}_1}{\sigma_1}\right)^2\right]} \cdot e^{\left[-\frac{1}{2}\left(\frac{x-\bar{x}_2}{\sigma_2}\right)^2\right]} \tag{A.2}$$

Again the combined probability is only dependent on the mean and variance of the two measurements. The results is, indeed, another Gaussian distribution with mean and variance other than above.

The mean of this new Gaussian is at its maximum. It is calculated as the point of the distribution where the first derivative becomes zero (equation A.3). A Gaussian is never to have a value of zero probability for any value of x, thus there is only one point with zero derivative: its mean.

$$\frac{dp_{12}}{dx}\Big|_{\bar{x}_{12}} = -\left[\frac{\bar{x}_{12}-\bar{x}_1}{\sigma_1^2} + \frac{\bar{x}_{12}-\bar{x}_2}{\sigma_2^2}\right] \cdot p_{12}(\bar{x}_{12}) = 0 \tag{A.3}$$

As the probability of x can never be zero, the term of equation A.4 must be equal to zero in order to annul the derivative.

$$\frac{\bar{x}_{12}-\bar{x}_1}{\sigma_1^2} + \frac{\bar{x}_{12}-\bar{x}_2}{\sigma_2^2} = 0 \tag{A.4}$$

In this way we can solve for the average of the combined probability, which will be a function of the mean and variance of the two measurements. The new mean and variance are defined in equations A.5 and A.6, respectively.

$$\bar{x}_{12} = \left(\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right) x_1 + \left(\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right) x_2 \tag{A.5}$$

$$\sigma_{12}^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} \tag{A.6}$$

Given a new measurement, the state estimate depends on that measurement and on the prior state, as that defined in equations A.7 y A.8.

$$\hat{x}_2 = \hat{x}_1 + \frac{\hat{\sigma_1}^2}{\hat{\sigma_1}^2+{\sigma_2}^2}(x_2 - \hat{x_1}) \tag{A.7}$$

$$\hat{\sigma_2}^2 = \left(1 - \frac{\hat{\sigma_1}^2}{\hat{\sigma_1}^2+{\sigma_2}^2}\right)\hat{\sigma_1}^2 \tag{A.8}$$

Where both terms include the update gain K (equation A.9).

$$K = \frac{\hat{\sigma_1}^2}{\hat{\sigma_1}^2 + {\sigma_2}^2} \tag{A.9}$$

It can then be defined a new state, which is a function of de update gain K (equations A.10 y A.11)

$$\hat{x}_2 = \hat{x}_1 + K(x_2 - \hat{x_1}) \tag{A.10}$$

$$\hat{\sigma_2}^2 = (1 - K)\,\hat{\sigma_1}^2 \tag{A.11}$$

Fig. A.2 show the optimum state of the systems, given those two measurements.

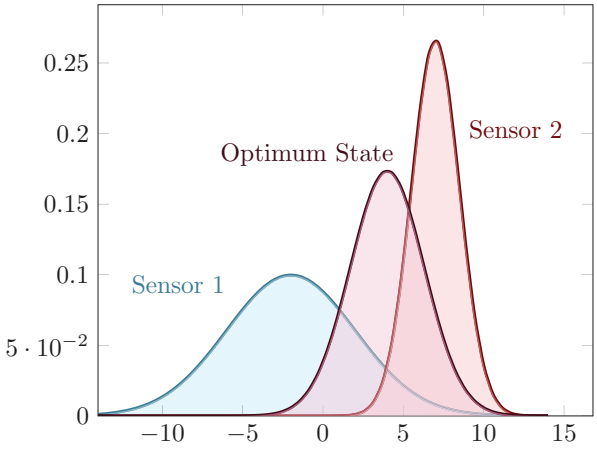


Figure A.2: The combination of two Gaussian distribution is also Gaussian.

A.1.2.2. Dynamic Systems

In case the system is dynamic all information available about how it changes in time must be included in the filter. This information can be of three types.

Dynamic System information . It is derived from what we expect the measurement to be, given the last measurement. Any system to use the Kalman filter must have a defined model which, as mentioned, must be non-linear.

Control Information . In controlled systems, such as robots, the output of the system can be changed by applying a control action. In that case, the system evolves due not only to its intrinsic dynamics but also depending on the control input. For example, if a robot moves at a constant speed and sends the order to accelerate, it is expected that its position in the next moment is different than it would be if it had not received that order.

Random Noise . Finally, the system is allow to have a random temporal evolution, knowing that there will be a noise associated with the process and with the acquisition stage.

A.1.3. Equations.

The recursive Kalman filter algorithm is summarized in Fig. A.3. It comprises a state prediction stage and a measurement update stage. In summary, Fig.A.3 includes the equations of each of the stages.

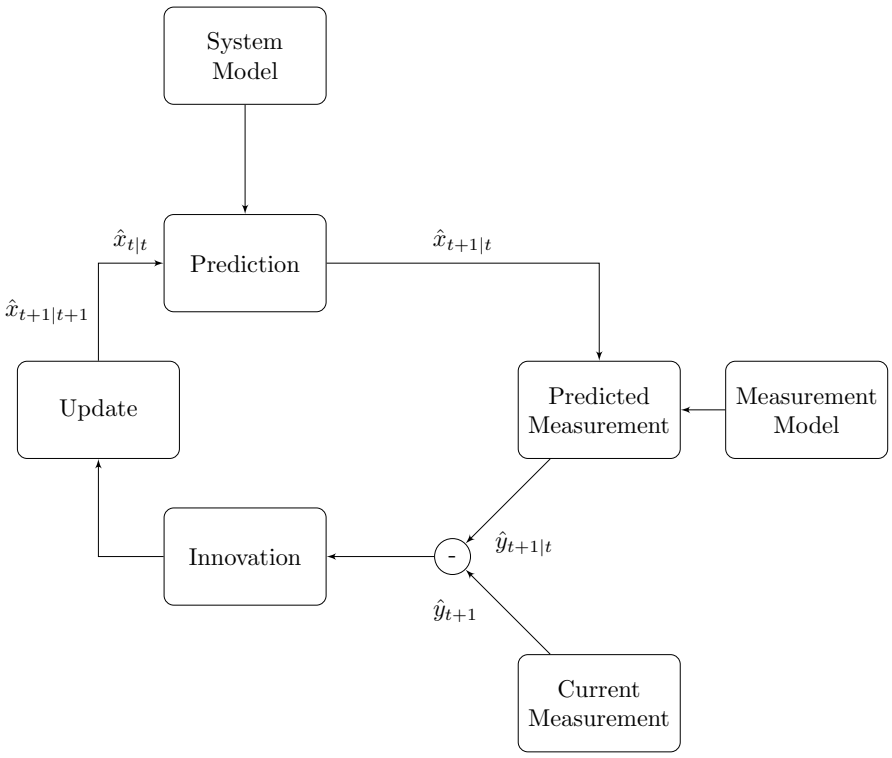


Figure A.3: Recursive Kalman filter algorithm.

A.1.3.1. Prediction

Given a system state $\hat{x}_{t|t}$ a prediction step ot the mean (equation A.12) and variance (equation A.13) is performed. In equation A.12, A is the update matrix that summarizes temporal evolution model of the system, from the instant $k-1$ until time k. The matrix B tranforms the control input u, if any, to the appropriate output.

The variance is updated with the same model (equation A.13). Any non-linearity is included in this update within the error Q of P covariance.

$$\hat{x}_k^- = A\hat{x}_{k-1} + Bu_{k-1} \tag{A.12}$$

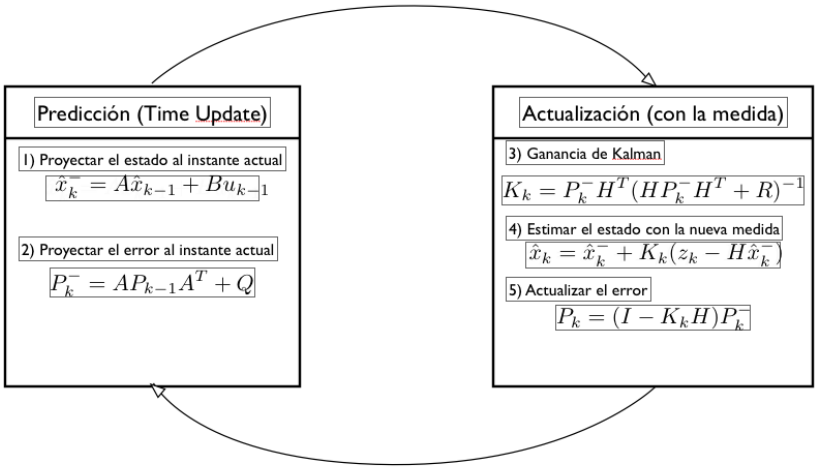


Figure A.4: Kalman filter equations.

$$P_k^- = AP_{k-1}A^T + Q \tag{A.13}$$

A.1.3.2. Measurement Update

Upon collecting a new measurents, the state is updated, taking into account the difference between the expected and actual measurements. The Kalman gain K_k relates the difference between the measured and the estimated states. Note that it is assumed that the state is proportional to the measurement.

$$K_k = P_k^- H^T (HP_k^- H^T + R)^{-1} \tag{A.14}$$

$$\hat{x}_k = \hat{x}_k^- + K_k(z_k - H\hat{x}_k^-) \tag{A.15}$$

$$P_k = (I - K_k H)P_k^- \tag{A.16}$$

The matrix R has an influence on the behavior of the filter. If its value is high, the filter will tend to give more importance to the system model so it would have an smaller reaction to new measurements. In the case where R is small, the filter would have a quicker reaction to the state measurement update.

Figures A.6 y A.6 are two examples of the behavior of a Kalman filter (green line), for the same measurements (red crosses) but with a different R matrix.

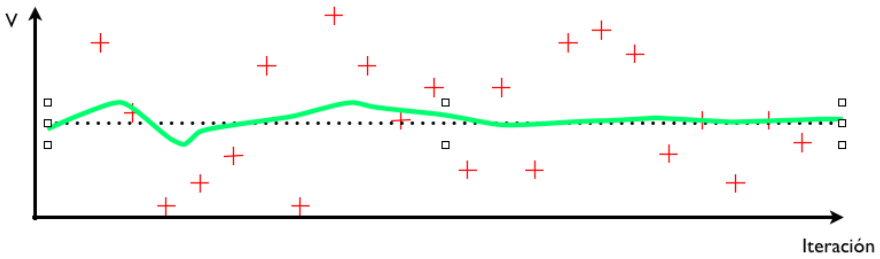


Figure A.5: Behaviour of the Kalman filter for large values of R.

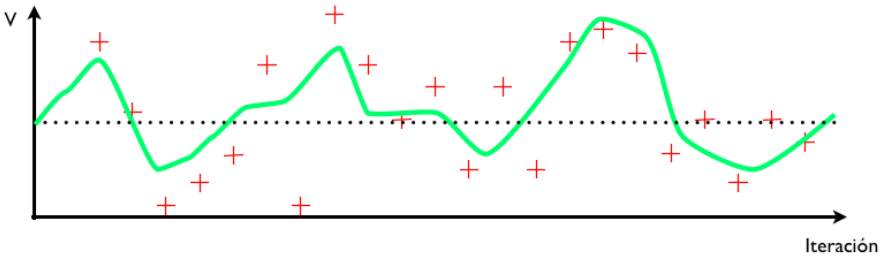


Figure A.6: Behaviour of the Kalman filter for small values of R.

A.1.4. Kalman Filter Variants

The Kalman filter relies on restrictions that are not always met. There are hardly any linear systems on which to use the filter. Therefore other techniques have been developed which preserve the basic operation of the Kalman filter, extending its use to non-linear systems, or Gaussian noise.

Fig. A.7 is a linear representation of the propagation of a random variable with Gaussian noise through a linear system. It can be seen that the output distribution is also Gaussian. This property does not occur if the system is not linear, as shown in Fig. A.8.

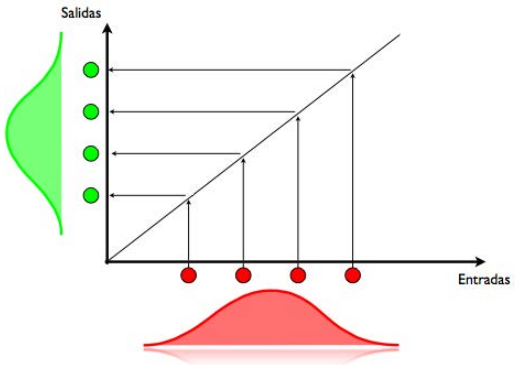


Figure A.7: In a linear system, if the input is Gaussian noise, the output will as well be.

A.1.4.1. Extended Kalman Filter

The Extended Kalman Filter (EKF) has been traditionally used in nonlinear systems. The algorithm is identical to the original, except that a linearization step is performed for each iteration of the matrices A (eq. (A.18)) and H (eq. (A.20)).

The modified EKF equations are:

- Update the state to current time

$$\hat{x}_k^- = f(\hat{x}_{k-1}, u_{k-1}) \tag{A.17}$$

- Update the error to current time

$$A_k = \frac{\delta f}{\delta x}|_{\hat{x}_{k^-|k^-}, u_k} \tag{A.18}$$

$$P_k^- = AP_{k-1}A^T + Q \tag{A.19}$$

- Kalman Gain

$$H_k = \frac{\delta h}{\delta x}|_{\hat{x}_{k|k^-}} \tag{A.20}$$

$$K_k = P_k^- H^T (HP_k^- H^T + R)^{-1} \tag{A.21}$$

- Estimation of the state with new measurement

$$\tilde{y}_k = z_k - h(\hat{x}^-) \tag{A.22}$$

$$\hat{x}_k = \hat{x}_k^- + K_k \tilde{y}_k \tag{A.23}$$

- Error update

$$P_k = (I - K_k H)P_k^- \tag{A.24}$$

A.2. Unscented Kalman Filter

The Unscented Kalman Filter (UKF) [108] extends the Kalman filter in highly nonlinear transformations of a random variable without linearization, as does the Extended Kalman Filter (EKF). This is particularly useful in the information acquisition process by a visual system. Therefore, the use of UKF over EKF is justified. Moreover, because it is not necessary to calculate the Jacobian, its computation requires significantly less time, while at the same time achieving better results.

The Unscented Kalman Filter propagates a random variable through a nonlinear system using a minimal set of *sigma* weighted points. The mean and variance of the variable after

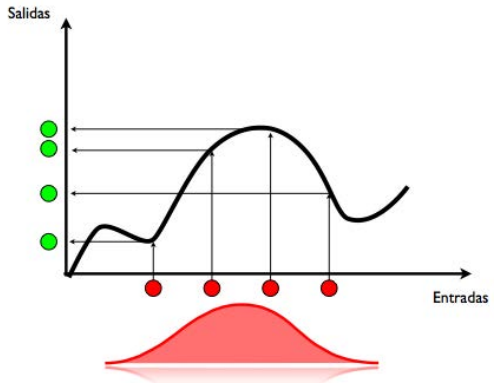


Figure A.8: In a non linear system, for a Gaussian input, the output distribution is not a Gaussian.

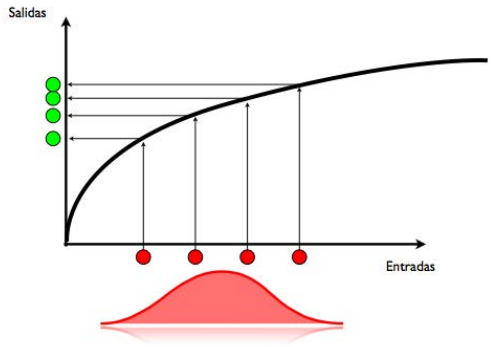


Figure A.9: If the systems is fairly linear the propagated veriable can be aproximated to a Gaussian.

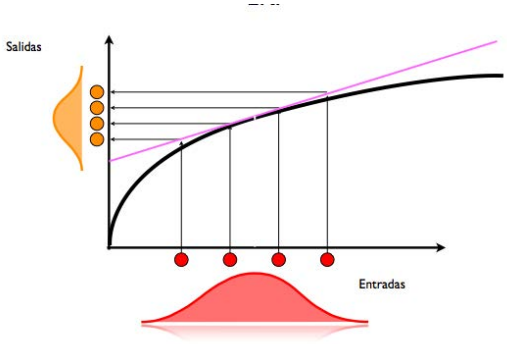


Figure A.10: Linearization around the working point.

processing is accurate, at least to the second order of the Taylor expansion series. Fig. A.11 the propagation of a random variable through a nonlinear system using EKF and UKF is represented. It also represents the propagation of a large number of samples (*sampling*). As it can be seen, the distribution of the variable is much more accurate in the case of UKF comparing the results with those of the EKF.

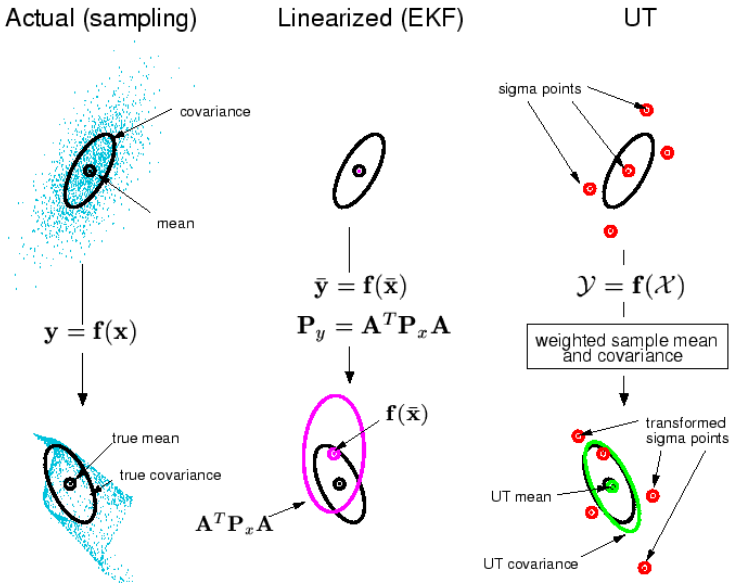


Figure A.11: Visual representation of UKF, EKF and sampling approaches (Eric A. Wan and Rudolph van der Merwe).

For a random variable x of dimension n with mean $\bar{x}$ and covariance P, the points *sigma* are:

$$\chi_0 = \bar{x} \tag{A.25}$$

$$\chi_i = \bar{x} + \sqrt{(n+\lambda)P} \qquad i = 1, \ldots, n \tag{A.26}$$

$$\chi_i = \bar{x} - \sqrt{(n+\lambda)P} \qquad i = n+1, \ldots, 2n \tag{A.27}$$

where $n + \lambda = \alpha^2(n + \kappa)$ is a scaling factor that determines the extent to which the *sigma* points are scattered around the mean $\bar{x}$.

Each *sigma* point is asigned a weight:

$$w_0^c = \frac{\lambda}{n+\lambda} + 1 - \alpha^2 + \beta \tag{A.28}$$

$$w_0^m = \frac{\lambda}{n+\lambda} \tag{A.29}$$

$$w_i^m = w_i^c = \frac{1}{2(n+\lambda)} \tag{A.30}$$

Once the selected points are propagated through the nonlinear function g (equation A.31), these weights are used to approximate the new mean and covariance (equations A.32 and A.33).

$$\gamma_i = f(\chi_i) \qquad i = 0, \ldots .2n \tag{A.31}$$

$$\bar{y} = \sum_{i=0}^{2n} w_i^m \gamma_i \tag{A.32}$$

$$P_y = \sum_{i=0}^{2n} w_i^c [\gamma_i - \bar{y}][\gamma_i - \bar{y}]^T \tag{A.33}$$

The following models represent the equations used for tracking object in world coordinates from a moving platform.

A.2.1. Prediction.

In this step the movement of the object may be implemented as an update of a simple Kalman filter, given the period of observation is relatively small. The movement of the object is modeled as uniform and rectilinear in the interval between two measurements. The real acceleration (which will always be non-zero, but small) and any non-linearity is included in this update within the error Q of covariance P. As mentioned above, the object position is simplified as the position of its centroid. Thus, the filter tracks a single point moving in a three dimensional space, and which always lies in the same plane.

$$\hat{x}_{t+1} = M \cdot R_h \cdot x_t + t_h \tag{A.34}$$

$$P_{t+1} = M \cdot P_t \cdot (M)^t + Q \tag{A.35}$$

It is expected that the object is moving with constant velocity rectilinear. This model is expressed in equation A.36, where for each measure, the prediction of the state for the next instant is the current state plus the distance traveled over the sampling time.

$$M = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{A.36}$$

The vehicle motion is modeled as a combination of a translation in the ground plane (t_h) and a rotation around axis z, perpendicular to that plane. The matrix R_h rotates the

relative position between vehicle and pedestrian, and the direction of the velocity vector. Information relative to the rotation angle β of the vehicle is known from the fusion of GPS and gyroscopes data. This way the movement of the vehicle, and hence that of the camera, can be compensated and the actual movement of the pedestrian is isolated.

$$R_h = \begin{bmatrix} R_p & 0 \\ 0 & R_v \end{bmatrix} \tag{A.37}$$

$$R_p = R_v = \begin{bmatrix} cos(-\Delta\beta) & -sin(-\Delta\beta) \\ sin(-\Delta\beta) & cos(-\Delta\beta) \end{bmatrix} \tag{A.38}$$

The process noise matrix is given by the equation A.39, where (a_x, a_y) is the acceleration of the vehicle.

$$Q = \begin{bmatrix} \frac{a_x^2 t^3}{3} & \frac{a_x^2 t^2}{2} & 0 & 0 \\ \frac{a_x^2 t^2}{2} & a_x^2 t & 0 & 0 \\ 0 & 0 & \frac{a_y^2 t^3}{3} & \frac{a_y^2 t^2}{2} \\ 0 & 0 & \frac{a_y^2 t^2}{2} & a_y^2 t \end{bmatrix} \tag{A.39}$$

A.2.2. Measurement Update

The measures of the position of the object are determined via a a pin-hole model are non-linear. The mean and covariance of the state prediction are used to generate the *sigma* points as explained above. These points are spread over function f. This function is non-linear and a suitable candidate to use the *Unscented* transformation, since the results obtained with the EKF for such applications can sometimes deteriorate quickly.

$$\gamma_t^i = f(\chi_{t-1}^i) \tag{A.40}$$

As explained in the section B the coordinates on the image of an object resting on the ground plane can be determined knowing the position of the camera relative to the plane where the object is.

The set of *sigma* points propagate through the system, using equation B.7 to project them on the image plane.

The position measurements are derived from the coordinates in the image, while the velocity remains constant as the object motion model. Since the velocity can not be observed directly, it can be assumed to be independent of the image coordinates.

The new sigma *sigma* points are used to obtained the mean and covariance prediction (equations (A.32) and (A.33)).

Finally, the new state is calculated. The last measure y is included in this last step to update the state. The difference between the measure and this prediction is weighted by the Kalman gain (K).

$$P_{xy} = \sum_{i=0}^{2n} \sum_{j=0}^{2n} w_{i,j}^{c} [\chi_{i,t|t-1} - \hat{x}_{t|t-1}][\gamma_{i,t|t-1} - \hat{y}_{t|t-1}]^T \tag{A.41}$$

$$K = P_{xy} P_y^{-1} \tag{A.42}$$

$$\hat{x}_t = \hat{x}_{t-1} + K(y - \hat{y}_{t-1}) \tag{A.43}$$

$$P_t = P_{t-1} - K P_y K^T \tag{A.44}$$

B

Vision System.

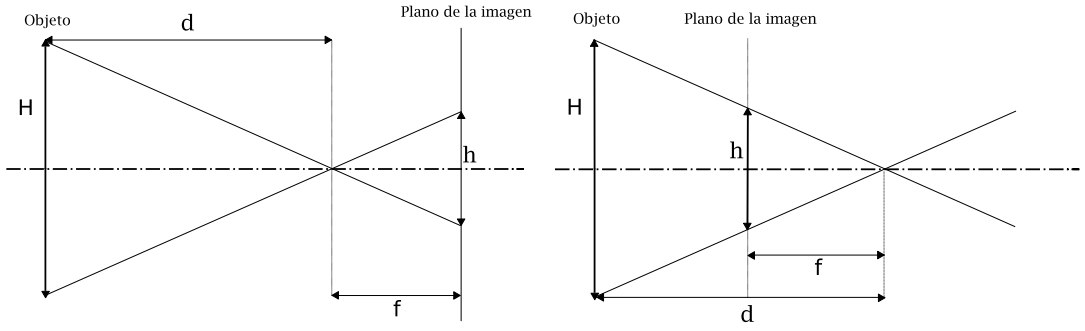
B.1. Calibration of the camera parameters.

The camera is the sensor by which the information is collected from the environment. As such, we need a model by which to obtain a relationship between the three-dimensional world in front of the camera and capture the two-dimensional image captured by the sensor.

In this case, the projective model used is the pinhole camera, wherein the camera optics are reduced to a point at the focal length of the sensor. The actual behavior of any lens is a bit different, however. In a pinhole camera, the small *aperture* allows for very little light to pass on to the sensor, so exposure times tend to be very long. To allow for more light, optics usually uses multiple lenses. Light passes through the lenses in a different way as modeled in the pinhole camera, making them more complicated geometrical models, and also introduce distortions in the images.

By calibration, it is possible to obtain the parameters which model the major distortions of the lens and thereby correct them.

Fig. B.1 represents the pinhole model. Each object point of a tridimensional object is projected on the sensor plane with a straight line that passes trough a point at a distance f from the sensor.



(a) Pinhole model. (b) Inverted pinhole model.

Figure B.1: Pinhole projective model.

B.1.1. Intrinsic parameters.

The projective geometry and the distortion parameters of the optic system are modeled by calibrationg the intrinsic parameters.

B.1.1.1. Projection Matrix

The projection matrix of a pinhole lens (equation B.1) relates the position of a point in the three-dimensional world coordinates with the pixel in the image that represents that point.

$$M = \begin{bmatrix} f_u & 0 & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{bmatrix} \tag{B.1}$$

- f_x: focal length, in pixels, axis x.
- f_y: focal length, in pixels, axis y.
- c_x: distance in pixels of the optical center from the x axis origin.
- c_y: distance in pixels of the optical center from the y axis origin

In a perfect lens the focal lengths in both axes (f_x and f_y) would be equal. However, it is to be expected a slight difference because of the difficulty of manufacturing a lens whose curvature is exactly equal for both axes, and because the optical axis is often not perfectly perpendicular to the sensor. The optical center (c_x, c_y) should coincide with the center of the sensor ($\frac{u}{2}$, $\frac{v}{2}$), but often the lenses are not properly aligned.

B.1.1.2. Distortion Parameters

The distortions in the images are produced mainly because it is easier to fabricate spherical lenses instead of parabolic ones [27]. This results in radial distortions, the most common of which is the *barrel* kind. In this, the points are projected farther from the optical center than it should, and this is displacement is accentuated for points located further away from the center. As a manifestation of the same physical principle, if the distortion projects points closer to the center than it should, the distortion received the name of *pincushion*.

Equation (B.3) gives corrected coordinates of one point of the image, given the radial distortion parameters, k_1, k_2 y k_4.

$$\begin{aligned} x_{corrected} &= x \cdot (1 + k_1 \cdot r^2 + k_2 \cdot r^4 + k_4 \cdot r^6) \\ y_{corrected} &= y \cdot (1 + k_1 \cdot r^2 + k_2 \cdot r^4 + k_4 \cdot r^6) \end{aligned} \tag{B.2}$$

The tangential distortion is due to the optical axis of the lens not being perfectly perpendicular to the sensor. To correct it equation B.4 is applied, where p_1 and p_2 are tangential distortion parameters.

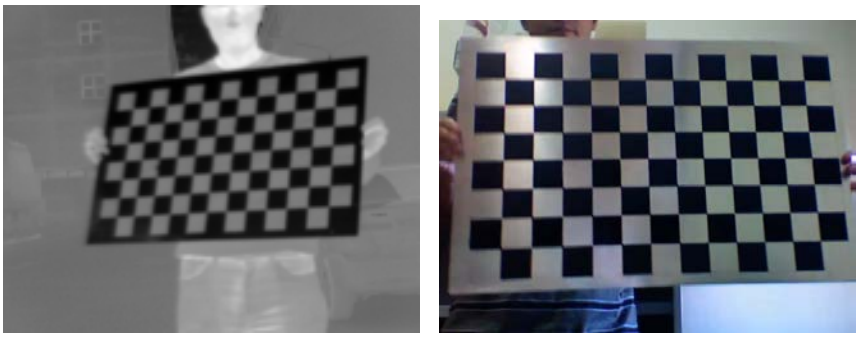
$$\begin{aligned} x_{corrected} &= x + [2p_1 \cdot y + p_2 \cdot (r^2 + 2x^2)] \\ y_{corrected} &= y + [2p_2 \cdot x + p_1 \cdot (r^2 + 2y^2)] \end{aligned} \tag{B.3}$$

B.1.1.3. Chessboard pattern

The calibration has been done with the Caltech *Camera Calibration Toolbox for Matlab* [26]. The calibration algorithm relies on extracting the corners of a chessboard pttern under several views. The geometry of the pattern is perfectly defined by the dimensions of the boxes. A recursive optimization algorithm fits the calibration parameters to the theoretical projection model of the lens.

This method is a very common one for calibrating VL cameras. In far infrared images it is not so simple, because there is no temperature difference between black and white paper and, as such, a normal pattern seems to be of a uniform gray and corners are indistinguishable.

To calibrate the camera parameters from far infrared images using the *Caltech Toolbox* a special pattern was made out of an aluminum foil (see Fig. B.2). It is still a chessboard, where the black squares are covered with acrylic paint, and the white one are aluminum. Polished metal surfaces reflect most of infrared radiations., therefore, pointing skyward, aluminum will appear to be much cooler than painted squares. The resulting images have enough contrast to extract the corners, so that the rest of the calibration is equivalent to that followed with a normal camera.



(a) Aluminum chessboard pattern in FIR.

(b) Aluminum chessboard pattern in VL.

Figure B.2: Aluminum chessboard pattern for FIR cameras calibration.

B.1.2. Extrinsic parameters.

The extrinsic parameters calibration is carried out by taking images of the ground in front of the camera. On it marks have been made at known distances. Given these distances and the fact that all marks are in the same plane it is possible to obtain the extrinsic parameters. Similarly to the intrinsic calibration, the extrinsic parameters are optimized in an iterative process that compares the theoretical projection with the actual one. The process stops when the error between the two falls below a limit.

B.2. Projective Geometry of the World into the Image.

Given the parameter calibration, as explained in section B.1, the intrinsic characteristics are assumed perfectly known, as is the position and orientation of the sensor plane in the world. The world coordinate system origin is in the ground plane, the camera being positioned in the z axis (see Fig. B.3).

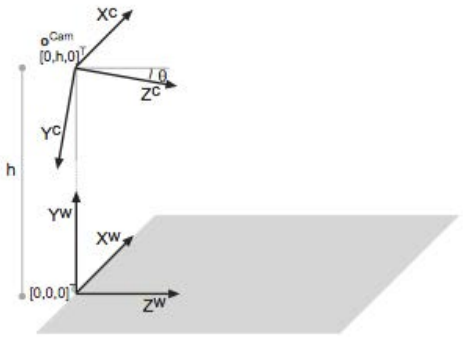


Figure B.3: Coordinate reference system of world and camera.

The virtual projection of a point in a three-dimensional space in the image plane can be calculated knowing its relative position to the ground plane, which is constant and equal to h. Using the pinhole model to project the scene objects in the image plane requires knowing the intrinsic parameters of the camera, such as the horizontal and vertical focal lengths (f_u, f_v) and the center of the image (c_u, c_V). A point with world coordinates (w_x, w_y, w_z) is projected into the image with homogeneous coordinates (U, V, S) with equation (B.4).

$$Image = \begin{bmatrix} U \\ V \\ S \end{bmatrix} = M \cdot R \cdot (World - T) \tag{B.4}$$

where M is the pinhole projection matrix (see equation (B.1)), T is the translation of the camera over the ground plane and $World$ is the matrix of coordinates in the world of the position of the projected object. The first two of these coordinates indicate the position within a two-dimensional plane. The third should always be 1.

R is the rotation of the ground coordinate system to the coordinate system of the camera (equation (B.5)), which is shown in figure B.3. The three rotation angles are $\alpha = \pi/2$, $\beta = 0$ and $\gamma = \pi$. These angles may vary during the course of the car due to vibrations and vehicle inertia when cornering (roll) or braking (pitch).

$$R = R_\alpha \cdot R_\beta \cdot R_\gamma = \tag{B.5}$$
$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & \sin(\alpha) \\ 0 & -\sin(\alpha) & \cos(\alpha) \end{bmatrix} \cdot \begin{bmatrix} \cos(\beta) & 0 & -\sin(\beta) \\ 0 & 1 & 0 \\ \sin(\beta) & 0 & \cos(\beta) \end{bmatrix} \cdot \begin{bmatrix} \cos(\gamma) & \sin(\gamma) & 0 \\ -\sin(\gamma) & \cos(\gamma) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$W - T = \begin{bmatrix} w_x \\ w_y \\ 1 \end{bmatrix} - \begin{bmatrix} 0 \\ h \\ 0 \end{bmatrix} = \begin{bmatrix} w_x \\ w_y - h \\ 1 \end{bmatrix} \tag{B.6}$$

To simplify the calculations, and since the point is always contained in the same plane, the projection matrix can be expressed as an homography, as in equation (B.7).

$$H = M \cdot W = \begin{bmatrix} f_u & 0 & c_u \\ 0 & f_v & c_v \\ 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} R_1 & R_2 & T \end{bmatrix} \tag{B.7}$$

Where R_1 and R_2 are the first and second columns of the rotation matrix (B.5).

Finally, and since the image projection is expressed in homogeneous coordinates, the position in the image is

$$u = \frac{U}{S} \tag{B.8}$$

$$v = \frac{V}{S} \tag{B.9}$$

B.3. Projection of the points of the image into the world

In section B.2 it has been explained how to project of a three-dimensional point in the world towards the image plane, knowing one of its coordinates, the height in this case. This relationship is defined as an homography matrix H. This same projection also works in reverse, so that one can determine the world position of a point from its position in the image. To do this reprojection (equation (B.10)) inverse of matrix H is calculated.

$$World = H^{-1} \cdot Image \tag{B.10}$$

B.4. Calibration of the gain curve of a microbolometer.

For applications based on the temperature of the object in the image, the microbolometer sensor must be calibrated.

In the case of a person, the skin surface can be approximated to a gray body, thas is, a body that emits heat in proportion to its temperature and does not reflect radiation of its environment. The emission would be equivalent to a fraction of the black body for that temperature. From the STEPHAN-BOLTZMANN equation (B.11) it is derived that the flow of energy transmitted per unit of surface depends not only on the temperature of the object but also on the temperature of the sensor.

$$\phi = \epsilon \cdot \delta \cdot (T^4 - T_{sensor}^4) \tag{B.11}$$

Where,

- ϕ : Flow of energy per unit of surface.
- ϵ : Emission factor of the body relative to a black body.
- $\delta = 5,67 \cdot 10^{-8} \frac{W}{m^2 \cdot K^4}$: Stephan-Boltzmann constant.
- T : Temperature of the object.
- T_{sensor} : Temperature of the sensor.

The gray level value of the pixels of the sensor also depends on the distance of the object and the absorption factor of the atmosphere. However, these parameters can be considered very small for short distances such as the range of pedestrian detection. Another factor, which should be considered is the gain of the sensor itself. The camera will be more sensitive to a particular wavelength.

Since the temperature sensor is a known value it is possible to calibrate the sensor sensitivity, relating the temperature of a gray body with gray levels on the image.

The camera sensor is an uncooled microbolometer, and produces images with a depth of 14 bits. Not being cooled, the gray level varies with the temperature sensor. Figure B.4 represents the sensitivity curve obtained in the calibration. Sensitivity curves have been obtained for three representative temperatures of the human body: the maximum and minimum temperature of the head and the minimum temperature of the body. Curves can be approximated, within the operating temperature range of the camera, to a third degree polynomial curve.

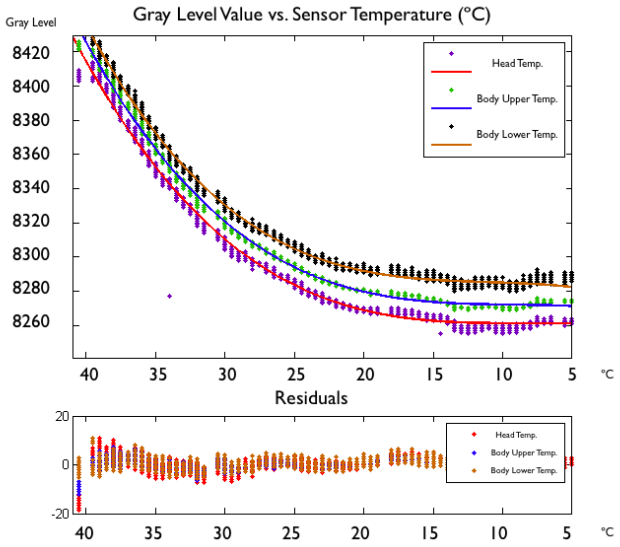


Figure B.4: Gray level of three constant temperatures of the human body, against temperature of the sensor.

References

[1] Ignacio Parra Alonso, David Fernández Llorca, Miguel Ángel Sotelo, Luis Miguel Bergasa, Pedro Revenga de Toro, Jesús Nuevo, Manuel Ocaña, and MA Garcia Garrido. Combination of feature extraction methods for svm pedestrian detection. *Intelligent Transportation Systems, IEEE Transactions on*, 8(2):292–307, 2007. 29, 35

[2] Robert William Gerard Anderson, AJ McLean, MJB Farmer, BH Lee, and CG Brooks. Vehicle travel speeds and the incidence of fatal pedestrian crashes. *Accident Analysis & Prevention*, 29(5):667–674, 1997. 8

[3] M Andriluka, Stefan Roth, and Bernt Schiele. People-tracking-by-detection and people-detection-by-tracking. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008. 39

[4] Richard Arndt, Roland Schweiger, Werner Ritter, Dietrich Paulus, and Otto Lohlein. Detection and Tracking of Multiple Pedestrians in Automotive Applications. In *2007 IEEE Intelligent Vehicles Symposium*, pages 13–18. IEEE, 2007. 32

[5] Max Bajracharya, Baback Moghaddam, Andrew Howard, Shane Brennan, and Larry H Matthies. A fast stereo-based system for detecting and tracking pedestrians from a moving vehicle. *The International Journal of Robotics Research*, 28(11-12):1466–1485, 2009. 35

[6] A Baumberg and D Hogg. An efficient method for contour tracking using active shape models. *Motion of Non-Rigid and Articulated Objects*, 1994. 24

[7] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and L van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008. 26

[8] R Behringer, S Sundareswaran, B Gregory, R Elsley, B Addison, W Guthmiller, R Daily, and D Bevly. The DARPA grand challenge - development of an autonomous vehicle. *Intelligent Vehicles Symposium, 2004 IEEE*, pages 226–231, 2004. 15

[9] Rodrigo Benenson, Markus Mathias, Radu Timofte, and Luc Van Gool. Pedestrian detection at 100 frames per second. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2903–2910. IEEE, 2012. 83

[10] M Bertozzi, A Broggi, L Bombini, and C Caraffi. Vision Technologies for Intelligent Vehicles. *Lecture Notes on Computer Science*, 2007. 35

[11] M Bertozzi, A Broggi, C Caraffi, M Delrose, M Felisa, and G Vezzoni. Pedestrian detection by means of far-infrared stereo vision. *Computer Vision and Image Understanding*, 106(2-3):194–204, May 2007. 19

[12] M Bertozzi, A Broggi, M Carletti, A Fascioli, Thorsten Graf, P Grisleri, and M M Meinecke. IR pedestrian detection for advanced driver assistance systems. *Pattern Recognition*, pages 582–590, 2003. 17, 18, 31

[13] M Bertozzi, A Broggi, M Del Rose, and M Felisa. A symmetry-based validator and refinement system for pedestrian detection in far infrared images. In *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE*, pages 155–160. IEEE, 2007. 19, 31

[14] M Bertozzi, A Broggi, M del Rose, M Felisa, A Rakotomamonjy, and F Suard. A pedestrian detector using histograms of oriented gradients and a support vector machine classifier. *IEEE Intelligent Transportation Systems Conference*, 2007. 27

[15] M Bertozzi, A Broggi, A Fascioli, and Thorsten Graf. Pedestrian detection for driver assistance using multiresolution infrared vision. *Vehicular Technology*, 2004. 31

[16] M Bertozzi, A Broggi, A Fascioli, A Tibaldi, R Chapuis, and F Chausse. Pedestrian localization and tracking system with Kalman filtering. *Intelligent Vehicles Symposium, 2004 IEEE*, pages 584–589, 2004. 32

[17] M Bertozzi, A Broggi, M Felisa, G Vezzoni, and M Del Rose. Low-level pedestrian detection by means of visible and far infra-red tetra-vision. In *Intelligent Vehicles Symposium, 2006 IEEE*, pages 231–236. IEEE, 2006. 19, 34

[18] M Bertozzi, A Broggi, P Grisleri, Thorsten Graf, and M M Meinecke. Pedestrian detection in infrared images. In *Intelligent Vehicles Symposium, 2003. Proceedings. IEEE*, pages 662–667, 2003. 21, 42

[19] M Bertozzi, A Broggi, Cristina Hilario, R Fedriga, G Vezzoni, and M Del Rose. Pedestrian detection in far infrared images based on the use of probabilistic templates. *Intelligent Vehicles Symposium, 2007 IEEE*, pages 327 – 332, May 2007. 21, 24, 42

[20] M Bertozzi, A Broggi, A Lasagni, and M del Rose. Infrared stereo vision-based pedestrian detection. *Intelligent Vehicles Symposium*, 2005. 19

[21] Massimo Bertozzi, Alberto Broggi, and Alessandra Fascioli. Vision-based intelligent vehicles: State of the art and perspectives. *Robotics and Autonomous systems*, 32(1):1–16, 2000. 35

[22] B Besbes, A Rogozan, and A Bensrhair. Pedestrian recognition based on hierarchical codebook of SURF features in visible and infrared images. *Intelligent Vehicles Symposium (IV), 2010 IEEE*, pages 156–161, 2010. 26

[23] E Binelli, A Broggi, A Fascioli, S Ghidoni, P Grisleri, Thorsten Graf, and M M Meinecke. A modular tracking system for far infrared pedestrian recognition. *Intelligent Vehicles Symposium, 2005. Proceedings. IEEE*, pages 759–764, 2005. 10, 21, 32, 42

[24] L Bombini, P Cerri, P Grisleri, S Scaffardi, and P Zani. An evaluation of monocular image stabilization algorithms for automotive applications. *Intelligent Transportation Systems Conference, 2006. ITSC'06. IEEE*, pages 1562–1567, 2006. 17

[25] A Bosch and X Zisserman, A Munoz. Representing shape with a spatial pyramid kernel. *ACM international conference on Image and video retrieval.*, 2007. 61

[26] J. Y. Bouquet. Camera calibration toolbox for matlab. `http://www.vision.caltech.edu/bouguetj/calib_doc/`, 2000. 17, 157

[27] Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library.* O'Reilly, Cambridge, MA, 2008. 156

[28] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 67

[29] A Broggi, P Grisleri, Thorsten Graf, and M M Meinecke. A software video stabilization system for automotive oriented applications. *Vehicular Technology Conference, 2005. VTC 2005-Spring. 2005 IEEE 61st*, 5:2760–2764 Vol. 5, 2005. 17, 18

[30] Alberto Broggi, Massimo Bertozzi, and Alessandra Fascioli. Self-calibration of a stereo vision system for automotive applications. In *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, volume 4, pages 3698–3703. IEEE, 2001. 17

[31] Alberto Broggi, Massimo Bertozzi, Alessandra Fascioli, and Massimiliano Sechi. Shape-based pedestrian detection. In *Intelligent Vehicles Symposium, 2000. IV 2000. Proceedings of the IEEE*, pages 215–220. IEEE, 2000. 19, 23, 31

[32] Alberto Broggi, Alessandra Fascioli, Paolo Grisleri, Thorsten Graf, and M Meinecke. Model-based validation approaches and matching techniques for automotive vision based pedestrian detection. In *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 1–1. IEEE, 2005. 31

[33] Alberto Broggi, RL Fedriga, and A Tagliati. Pedestrian detection on a moving vehicle: an investigation about near infra-red images. In *Intelligent Vehicles Symposium, 2006 IEEE*, pages 431–436. IEEE, 2006. 24

[34] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998. 73

[35] CARE. European Commission: Fatalities at 30 days in EU countries. `http://ec.europa.eu/transport/road_safety/pdf/statistics/2011_user.pdf`, 2011. 2

[36] Juan-Pablo Carrasco, Arturo de la Escalera de la Escalera, José María Armingol, et al. Recognition stage for a speed supervisor based on road sign detection. *Sensors*, 12(9):12153–12168, 2012. 11

[37] Juan Pablo Carrasco Pascual. *Advanced driver assistance system based on computer vision using detection, recognition and tracking of road signs.* PhD thesis, Universidad Carlos III de Madrid, 2009. 11

[38] C-I Chang, Yingzi Du, J Wang, S-M Guo, and PD Thouin. Survey and comparative analysis of entropy and relative entropy thresholding techniques. In *Vision, Image and Signal Processing, IEE Proceedings-*, volume 153, pages 837–850. IET, 2006. 22

[39] J Chang, H Liao, M Hor, and J Hsieh. New automatic multi-level thresholding technique for segmentation of thermal images. *Image and Vision Computing*, 1997. 22

[40] T Chateau, V Gay-Belille, and F Chausse. Real-time tracking with classifiers. *Dynamical Vision*, 2007. 32

[41] Yuxi Chen and Chongzhao Han. Night-time pedestrian detection by visual-infrared video fusion. *Intelligent Control and Automation, 2008. WCICA 2008. 7th World Congress on*, pages 5079–5084, 2008. 35

[42] Bryan Clarke, Stewart Worrall, Graham Brooker, and Eduardo Nebot. Sensor modelling for radar-based occupancy mapping. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 3047–3054. IEEE, 2012. 34

[43] Juan Manuel Collado Hernáiz. *Detección y modelado de carriles de vías interurbanas mediante análisis de imágenes para un sistema de ayuda a la conducción*. PhD thesis, Universidad Carlos III de Madrid, 2009. 12

[44] D Comaniciu. An algorithm for data-driven bandwidth selection. *Pattern Analysis and Machine Intelligence*, 2003. 31

[45] B. Cyganek. Circular road signs recognition with soft classifiers. *Integrated Computer-Aided Engineering*, 14(4):323–343, 2007. 8

[46] Congxia Dai and Yunfei Zheng. Pedestrian detection and tracking in infrared imagery using shape and appearance. *Computer Vision and Image Understanding*, 2007. 33

[47] Congxia Dai, Yunfei Zheng, and Xin Li. Layered Representation for Pedestrian Detection and Tracking in Infrared Imagery. In *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, page 13, 2005. 33

[48] Navneet Dalal. *Finding people in images and videos*. PhD thesis, Institut National Polytechnique de Grenoble-INPG, 2006. 31

[49] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:886–893, 2005. 23, 26, 30, 38, 39

[50] Linh Dang, Buu Bui, P.D Vo, T.N Tran, and B.H Le. Improved HOG Descriptors. In *Knowledge and Systems Engineering (KSE), 2011 Third International Conference on*, pages 186–189, 2011. 26

[51] Thao Dang and Christian Hoffmann. Stereo calibration in vehicles. In *Intelligent Vehicles Symposium, 2004 IEEE*, pages 268–273. IEEE, 2004. 17

[52] J W Davis and M A Keck. A Two-Stage Template Approach to Person Detection in Thermal Imagery. In *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)*, pages 364–369. IEEE, 2005. 91

[53] Larry S. Davis, V Philomin, and R Duraiswami. Tracking humans from a moving platform. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, pages 171–178, 2000. 32

[54] Ernst D Dickmanns. *Dynamic vision for perception and control of motion [electronic resource]*. Springer, 2007. 35

[55] T.G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998. 77

[56] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(4):743–761, 2012. 36, 39

[57] Piotr Dollár, Boris Babenko, Serge Belongie, Pietro Perona, and Zhuowen Tu. Multiple component learning for object detection. *Computer Vision–ECCV 2008*, pages 211–224, 2008. 29

[58] Piotr Dollár and S Belongie. The fastest pedestrian detector in the west. *BMVC 2010*, 2010. 29, 83, 115

[59] Piotr Dollár, Zhuowen Tu, and P Perona. Integral channel features. *BMVC 2009*, 2009. 23, 29, 66

[60] Piotr Dollár, Christian Wojek, Bernt Schiele, and P Perona. Pedestrian detection: A benchmark. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 304–311, 2009. 35, 39

[61] Li Dong, Xinguo Yu, Liyuan Li, and J.K.E Hoe. HOG based multi-stage object detection and pose recognition for service robot. In *Control Automation Robotics & Vision (ICARCV), 2010 11th International Conference on*, pages 2495–2500, 2010. 26

[62] Jan Eichhorn and Olivier Chapelle. Object categorization with SVM: kernels for local features. *In Advances in Neural Information Processing Systems (NIPS)*, 2004. 30

[63] V. Enescu, G. De Cubber, K. Cauwerts, H. Sahli, E. Demeester, D. Vanhooydonck, and M. Nuttin. Active stereo vision-based mobile robot navigation for person tracking. *Integrated Computer-Aided Engineering*, 13(3):203–222, 2006. 8

[64] M. Enzweiler and D.M. Gavrila. Monocular pedestrian detection: Survey and experiments. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(12):2179–2195, 2009. 39

[65] Markus Enzweiler, Angela Eigenstetter, Bernt Schiele, and Dariu M Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 990–997, 2010. 28

[66] Markus Enzweiler and Dariu M Gavrila. Monocular pedestrian detection: Survey and experiments. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(12):2179–2195, 2009. 35

[67] Andreas Ess, B Leibe, Konrad Schindler, and L van Gool. Robust Multiperson Tracking from a Mobile Platform. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(10):1831–1846, October 2009. 35

[68] Andreas Ess, B Leibe, and L van Gool. Depth and Appearance for Mobile Scene Analysis. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, 2007. 39

[69] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 85

[70] Yajun Fang, K Yamada, Y Ninomiya, B. Horn, and I Masaki. Comparison between Infrared-image-based and Visible-image-based Approaches for Pedestrian Detection. In *IEEE IV2003 Intelligent Vehicles Symposium. Proceedings*, pages 505–510. IEEE, 2003. 21

[71] Yajun Fang, K Yamada, Y Ninomiya, B. Horn, and I Masaki. A shape-independent method for pedestrian detection with far-infrared images. *Vehicular Technology, IEEE Transactions on*, 53(6):1679–1697, 2004. 21

[72] B Fardi, U Schuenert, and G Wanielik. Shape and motion-based pedestrian detection in infrared images: a multi sensor approach. *Intelligent Vehicles Symposium*, 2005. 20, 34

[73] William L Fehlman II and Mark K Hinders. Passive infrared thermographic imaging for mobile robot object identification. *Journal of Field Robotics*, pages n/a–n/a, 2009. 37

[74] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multi-scale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 28, 31, 88, 94

[75] P.F Felzenszwalb, R.B Girshick, D McAllester, and Deva Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010. 28, 94

[76] David J Field et al. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 4(12):2379–2394, 1987. 50

[77] Marco Javier Flores, José María Armingol, and Arturo de la Escalera. Driver drowsiness warning system using visual information for both diurnal and nocturnal illumination conditions. *EURASIP Journal on Advances in Signal Processing*, 2010:3, 2010. 12

[78] Marco Javier Flores Calero. *Sistema Avanzado de Asistencia a la Conducción para la Detección de la Somnolencia y la Distracción durante la Noche*. PhD thesis, Universidad Carlos III de Madrid, 2009. 12

[79] U Franke, Dariu M Gavrila, and S Gorzig. Autonomous driving goes downtown. ... *Systems and Their ...*, 1998. 31

[80] Uwe Franke and Armin Joos. Real-time stereo vision for urban traffic scene understanding. In *Intelligent Vehicles Symposium, 2000. IV 2000. Proceedings of the IEEE*, pages 273–278. IEEE, 2000. 32

[81] Yoav Freund, Robert Schapire, and N Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999. 30

[82] Rikke Gade and Thomas B Moeslund. Thermal cameras and applications: a survey. *Machine Vision and Applications*, pages 1–18, 2013. 10, 36

[83] Tarak Gandhi and Mohan M Trivedi. Pedestrian protection systems: Issues, survey, and challenges. *Intelligent Transportation Systems, IEEE Transactions on*, 8(3):413–430, 2007. 9, 15, 35

[84] Dariu M Gavrila. Pedestrian detection from a moving vehicle. *Computer Vision—ECCV 2000*, pages 37–49, 2000. 23

[85] Dariu M Gavrila. A bayesian, exemplar-based approach to hierarchical shape matching. *Pattern Analysis and Machine Intelligence*, 2007. 23

[86] Dariu M Gavrila and S Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *International Journal of Computer Vision*, 73(1):41–59, 2007. 23, 31, 35

[87] Dariu M Gavrila and V Philomin. Real-Time Object Detection for Smart Vehicles. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 87–93 vol.1. IEEE, 1999. 23

[88] DM Gavrila, Jan Giebel, and Stefan Munder. Vision-based pedestrian detection: The protector system. In *Intelligent Vehicles Symposium, 2004 IEEE*, pages 13–18. IEEE, 2004. 23, 31

[89] Junfeng Ge, Yupin Luo, and Gyomei Tei. Real-Time Pedestrian Detection and Tracking at Nighttime for Driver-Assistance Systems. *IEEE Transactions On Intelligent Transportation Systems*, 10(2):283–298, 2009. 21

[90] D. Gerónimo, A. Sappa, A. López, and D. Ponsa. Adaptive image sampling and windows classification for on-board pedestrian detection. In *Proceedings of the International Conference on Computer Vision Systems, Bielefeld, Germany*, 2007. 39

[91] David Gerónimo, Antonio M Lopez, Angel D Sappa, and Thorsten Graf. Survey on Pedestrian Detection for Advanced Driver Assistance Systems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(7):1239–1258, 2010. 8, 10, 35, 39

[92] Neil J Gordon, David J Salmond, and Adrian FM Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113. IET, 1993. 32

[93] Chunhui Gu, J.J Lim, P Arbelaez, and J. Malik. Recognition using regions. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1030–1037, 2009. 23

[94] Lie Guo, Linhui Li, Yibing Zhao, and Mingheng Zhang. Study on pedestrian detection and tracking with monocular vision. In *Computer Technology and Development (ICCTD), 2010 2nd International Conference on*, pages 466–470, 2010. 35

[95] K Hajebi and J Zelek. Dense surface from infrared stereo. *IEEE Workshop on Applications of Computer Vision*, 2007. 19

[96] C Harris and M Stephens. A combined corner and edge detector. *Alvey vision conference*, 1988. 24

[97] B Heisele and C Woehler. Motion-based recognition of pedestrians. In *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, pages 1325–1330, 1998. 20

[98] C Hilario, J Collado, J Armingol, and A De la Escalera. Pedestrian detection for intelligent vehicles based on active contour models and stereo vision. *Computer Aided Systems Theory–EUROCAST 2005*, pages 537–542, 2005. 24

[99] Cristina Hilario Gómez. *Detección de peatones en el espectro visible e infrarrojo para un sistema avanzado de asistencia a la conducción*. PhD thesis, Universidad Carlos III de Madrid, 2008. 31, 42

[100] D Hoiem and AA Efros. Putting objects in perspective. *International Journal of Computer Vision*, 2008. 17

[101] Qian Hong-bo and Han Hao. The Applications and Methods of Pedestrian Automated Detection. In *Measuring Technology and Mechatronics Automation (ICMTMA), 2010 International Conference on*, pages 806–809, 2010. 35

[102] J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982. 73

[103] Sibt Ul Hussain and B Triggs. Feature Sets and Dimensionality Reduction for Visual Object Detection. *British Machine Vision Conference*, pages 112.1–112.10–112.1–112.10, August 2010. 29

[104] A Iketani, A Nagai, Y Kuno, and Y Shirai. Real-Time Surveillance System Detecting Persons in Complex Scenes. *Real-Time Imaging*, 2001. 33

[105] A.K Jain, R.P.W Duin, and Jianchang Mao. Statistical pattern recognition: a review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1):4–37, 2000. 39

[106] G.H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995. 73

[107] Michael J Jones and Daniel Snow. Pedestrian detection using boosted features over many frames. *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4, 2008. 24

[108] S Julier and J Uhlmann. A new extension of the Kalman filter to nonlinear systems. *Int. Symp. Aerospace/Defense Sensing*, 1997. 32, 121, 141, 148

[109] Heewook Jung, Joo Kooi Tan, Seiji Ishikawa, and Takashi Morie. Applying HOG feature to the detection and tracking of a human on a bicycle. In *Control, Automation and Systems (ICCAS), 2011 11th International Conference on*, pages 1740–1743, 2011. 26

[110] M.B Kaaniche and F Bremond. Tracking HoG Descriptors for Gesture Recognition. In *Advanced Video and Signal Based Surveillance, 2009. AVSS '09. Sixth IEEE International Conference on*, pages 140–145, 2009. 26

[111] H. Kakiuchi, T. Kawamura, T. Shimizu, and K. Sugahara. Bypass methods for constructing robust automatic human tracking system. *Integrated Computer-Aided Engineering*, 17(1):41–58, 2010. 8

[112] M Kass, A Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1988. 24

[113] Josef Kittler, John Illingworth, and J Föglein. Threshold selection based on a simple image statistic. *Computer vision, graphics, and image processing*, 30(2):125–147, 1985. 22

[114] M. Kohler et al. *Using the Kalman filter to track human interactive motion: modelling and initialization of the Kalman filter for translational motion.* Citeseer, 1997. 32

[115] Stanley Kok and Pedro Domingos. Learning the structure of markov logic networks. In *Proceedings of the 22nd international conference on Machine learning*, pages 441–448. ACM, 2005. 111

[116] P Kovesi. Image features from phase congruency. *Videre: Journal of Computer Vision Research*, 1(3):1–26, 1999. 47

[117] Peter Kovesi. Phase congruency detects corners and edges. In *The australian pattern recognition society conference: DICTA 2003*, 2003. 49

[118] S.J Krotosky and Mohan M Trivedi. A Comparison of Color and Infrared Stereo Approaches to Pedestrian Detection. *Intelligent Vehicles Symposium, 2007 IEEE*, pages 81–86, 2007. 19

[119] S.J Krotosky and Mohan M Trivedi. On Color-, Infrared-, and Multimodal-Stereo Approaches to Pedestrian Detection. *Intelligent Transportation Systems, IEEE Transactions on*, 8(4):619–629, 2007. 19, 34

[120] Raphael Labayrade, Didier Aubert, and J-P Tarel. Real time obstacle detection in stereovision on non flat road geometry through. In *Intelligent Vehicle Symposium, 2002. IEEE*, volume 2, pages 646–651. IEEE, 2002. 17

[121] Suk Kyu Lee, K McHenry, R Kooper, and P Bajcsy. Characterizing human subjects in real-time and three-dimensional spaces by integrating thermal-infrared and visible spectrum cameras. *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 1708–1711, 2009. 22, 35

[122] Bastian Leibe, Nico Cornelis, Kurt Cornelis, and Luc Van Gool. Dynamic 3d scene analysis from a moving vehicle. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 32

[123] Bastian Leibe, Edgar Seemann, and Bernt Schiele. Pedestrian detection in crowded scenes. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 878–885. IEEE, 2005. 23

[124] Wei Li, Dequan Zheng, Tiejun Zhao, and Mengda Yang. An effective approach to pedestrian detection in thermal imagery. In *Natural Computation (ICNC), 2012 Eighth International Conference on*, pages 325–329. IEEE, 2012. 91

[125] Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–900. IEEE, 2002. 24

[126] Jae S. Lim. *Two-dimensional signal and image processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1990. 62

[127] D.T. Lin and D.C. Pan. Integrating a mixed-feature model and multiclass support vector machine for facial expression recognition. *Integrated Computer-Aided Engineering*, 16(1):61–74, 2009. 76

[128] Zhe Lin, Larry S. Davis, David Doermann, and Daniel DeMenthon. Hierarchical Part-Template Matching for Human Detection and Segmentation. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 29

[129] B. Ling, M. Zeifman, and D. Gibson. Multiple pedestrians detection using ir led stereo camera. In *Proc. SPIE*, volume 6764, pages 9–12, 2007. 9

[130] X. Liu, PH Tu, J. Rittscher, A. Perera, and N. Krahnstoever. Detecting and counting people in surveillance applications. In *Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on*, pages 306–311. IEEE, 2005. 8

[131] Yazhou Liu, Shiguang Shan, Xilin Chen, Janne Heikkila, Wen Gao, and Matti Pietikainen. Spatial-temporal granularity-tunable gradients partition (STGGP) descriptors for human detection. In *ECCV'10: Proceedings of the 11th European conference on Computer vision: Part I*. Springer-Verlag, September 2010. 23

[132] D.F. Llorca, M.a. Sotelo, a.M. Hellín, a. Orellana, M. Gavilán, I.G. Daza, and a.G. Lorente. Stereo regions-of-interest selection for pedestrian protection: A survey. *Transportation Research Part C: Emerging Technologies*, 25:226–237, December 2012. 17, 19

[133] D Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004. 23, 25, 26

[134] D Lowe. SIFT: Scale Invariant Feature Transform. *eecs.umich.edu*, 2007. 25

[135] D G Lowe. Object recognition from local scale-invariant features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 2:1150–1157 vol.2, 1999. 25

[136] O Mateo Lozano. Real-time visual tracker by stream processing. *Journal of Signal Processing Systems*, 2009. 32

[137] M Mahlisch, M Oberlander, O. Lohlein, Dariu M Gavrila, and Werner Ritter. A Multiple Detector Approach to Low-resolution FIR Pedestrian Recognition. In *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.*, pages 325–330. IEEE, 2005. 24

[138] Subhransu Maji, Alexander C Berg, and Jitendra Malik. Classification using intersection kernel support vector machines is efficient. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 31

[139] Lucio Marcenaro, Gianni Vernazza, and Carlo S Regazzoni. Image stabilization algorithms for video-surveillance applications. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 1, pages 349–352. IEEE, 2001. 18

[140] A.M. Martinez and A.C. Kak. Pca versus lda. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):228–233, 2001. 73, 76

[141] U Meis, Werner Ritter, H Neumann, and A DaimlerChrysler. Detection and classification of obstacles in night vision traffic scenes based on infrared imagery. *Intelligent Transportation Systems*, 2003. 21

[142] M Meuter, U Iurgel, S-B Park, and A Kummert. The unscented Kalman filter for pedestrian tracking from a moving host. *Intelligent Vehicles Symposium*, 2008. 32

[143] Fernand Meyer. Topographic distance and watershed lines. *Signal processing*, 38(1):113–125, 1994. 108

[144] R Miezianko and D Pokrajac. People detection in low resolution infrared videos. *Computer Vision and Pattern Recognition Workshops, 2008. CVPR Workshops 2008. IEEE Computer Society Conference on*, pages 1–6, May 2008. 27

[145] K Mikolajczyk and C Schmid. Human detection based on a probabilistic assembly of robust part detectors. *Computer Vision-ECCV 2004*, 2004. 28

[146] S. Milch and M. Behrens. Pedestrian detection with radar and computer vision. *Proceedings of Progress in Automobile*, 9, 2001. 34

[147] A. Miller, P. Babenko, M. Hu, and M. Shah. Person tracking in uav video. *Multimodal Technologies for Perception of Humans*, pages 215–220, 2008. 8

[148] A. Miron, B. Besbes, A. Rogozan, S. Ainouz, and A. Bensrhair. Intensity self similarity features for pedestrian detection in far-infrared images. In *Intelligent Vehicles Symposium (IV), 2012 IEEE*, pages 1120–1125. IEEE, 2012. 10

[149] A Mohan, C Papageorgiou, and T Poggio. Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4):349–361, April 2001. 28

[150] M Concetta Morrone and DC Burr. Feature detection in human vision: A phase-dependent energy model. *Proceedings of the Royal Society of London. Series B, biological sciences*, pages 221–245, 1988. 47

[151] S Munder and Dariu M Gavrila. An Experimental Study on Pedestrian Classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(11):1863–1868, 2006. 39

[152] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial & Applied Mathematics*, 5(1):32–38, 1957. 123

[153] Basam Musleh, Fernando García, Javier Otamendi, José M^{a} Armingol, and Arturo De la Escalera. Identifying and tracking pedestrians based on sensor fusion and motion stability predictions. *Sensors*, 10(9):8028–8053, 2010. 11

[154] Basam Musleh, David Martin, Arturo de la Escalera, and José María Armingol. Visual ego motion estimation in urban environments based on uv disparity. In *Intelligent Vehicles Symposium (IV), 2012 IEEE*, pages 444–449. IEEE, 2012. 11

[155] H Naci, D Chisholm, and TD Baker. Distribution of road traffic deaths by road user group: a global comparison. *Injury Prevention*, 15(1):55–59, 2009. 2

[156] H Nanda. Probabilistic template based pedestrian detection in infrared videos. *Intelligent Vehicle Symposium*, 2002. 21, 24, 42

[157] Hoang Thanh Nguyen and B Bhanu. Tracking pedestrians with bacterial foraging optimization swarms. *Evolutionary Computation (CEC), 2011 IEEE Congress on*, pages 491–495, 2011. 33

[158] NHTSA. National Highway Traffic Safety Administration: Pedestrian Statistics. `http://www-fars.nhtsa.dot.gov/People/PeopleAllVictims.aspx`, 2010. 2

[159] T Ojala, Matti Pietikainen, and T Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002. 29, 38, 76

[160] Luciano Oliveira and Urbano Nunes. Context-aware pedestrian detection using lidar. In *Intelligent Vehicles Symposium (IV), 2010 IEEE*, pages 773–778. IEEE, 2010. 108

[161] Luciano Oliveira, Urbano Nunes, Paulo Peixoto, Marco Silva, and Fernando Moita. Semantic fusion of laser and vision in pedestrian detection. *Pattern Recognition*, 43(10):3648–3659, 2010. 108

[162] D. Olmeda, A. de la Escalera, and J.M. Armingol. Far infrared pedestrian detection and tracking for night driving. *Robotica*, 29(04):495–505, 2011. 17

[163] Ronan O'Malley, Edward Jones, and Martin Glavin. Detection of pedestrians in far-infrared automotive night vision using region-growing and clothing distortion compensation. *Infrared Physics and Technology*, 53(6):439–449, November 2010. 10, 27

[164] Oren, C Papageorgiou, P Shinha, E Osuna, and T Poggio. A trainable system for people detection. *Proc. of Image Understanding Workshop*, 1997. 24

[165] N Otsu. A Threshold Selection Method from Gray-Level Histograms. *Systems, Man and Cybernetics, IEEE Transactions on*, 9(1):62–66, 1979. 22

[166] G Overett and L Petersson. Large scale sign detection using HOG feature variants. *Intelligent Vehicles Symposium (IV), 2011 IEEE*, pages 326–331, 2011. 26

[167] Y Owechko, S Medasani, and N Srinivasa. Classifier Swarms for Human Detection in Infrared Imagery. *Computer Vision and Pattern Recognition Workshop*, 2004. 27, 31

[168] C Papageorgiou and T Poggio. A trainable system for object detection. *International Journal of Computer Vision*, 38(1):15–33, 2000. 22, 24, 39

[169] Dennis Park, Deva Ramanan, and Charless Fowlkes. Multiresolution models for object detection. In *ECCV'10: Proceedings of the 11th European conference on Computer vision: Part IV*. Springer-Verlag, September 2010. 28

[170] Eero Pasanen. Driving speeds and pedestrian safety: a mathematical model. Technical report, Transportation Research Board of the National Academies, 1992. 8

[171] V Philomin, R Duraiswami, and Larry S. Davis. Pedestrian tracking from a moving vehicle. *Intelligent Vehicles Symposium, 2000. IV 2000. Proceedings of the IEEE*, pages 350–355, 2000. 32

[172] Fatih Porikli. Integral histogram: a fast way to extract histograms in Cartesian spaces. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, pages 829–836, 2005. 23, 66

[173] C. Premebida and U.J.C. Nunes. Fusing lidar, camera and semantic information: A context-based approach for pedestrian detection. *The International Journal of Robotics Research*, 2013. 8, 34

[174] Cristiano Premebida, Gonçalo Monteiro, Urbano Nunes, and Paulo Peixoto. A lidar and vision-based approach for pedestrian and vehicle detection and tracking. In *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE*, pages 1044–1049. IEEE, 2007. 34

[175] Richard A Retting, Susan A Ferguson, and Anne T McCartt. A review of evidence-based traffic engineering measures designed to reduce pedestrian-motor vehicle crashes. *American Journal of Public Health*, 93(9):1456–1463, 2003. 8

[176] Antoni Rogalski. Infrared detectors: an overview. *Infrared Physics & Technology*, 43(3):187–210, 2002. 10

[177] Erik Rosen, Helena Stigson, and Ulrich Sander. Literature review of pedestrian fatality risk as a function of car impact speed. *Accident analysis and prevention*, 43(1):25–33, 2011. 8

[178] Payam Sabzmeydani and G Mori. Detecting Pedestrians by Learning Shapelet Features. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007. 23

[179] R.E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336, 1999. 70, 73

[180] Cordelia Schmid and Roger Mohr. Local grayvalue invariants for image retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(5):530–535, 1997. 31

[181] E Seemann, B Leibe, K Mikolajczyk, and Bernt Schiele. An evaluation of local shape-based features for pedestrian detection. *Proc. BMVC*, 2005. 23, 27

[182] A Shashua, Y Gdalyahu, and G Hayun. Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. *Intelligent Vehicles Symposium, 2004 IEEE*, pages 13–18, 2004. 26, 28, 31, 35

[183] Ding Shumin, Liu Zhoufeng, and Li Chunlei. AdaBoost learning for fabric defect detection based on HOG and SVM. In *Multimedia Technology (ICMT), 2011 International Conference on*, pages 2903–2906, 2011. 26

[184] N Siebel and S Maybank. Fusion of Multiple Tracking Algorithms for Robust People Tracking. *Lecture Notes On Computer Science*, 2002. 33

[185] Nils T Siebel and SJ Maybank. Real-time tracking of pedestrians and vehicles. In *IEEE Workshop on PETS*, 2001. 33

[186] D Simonnet, SA Velastin, E Turkbeyler, and J Orwell. Backgroundless detection of pedestrians in cluttered conditions based on monocular images: a review. *Computer Vision, IET*, 6(6):540–550, 2012. 36

[187] Parag Singla and Pedro Domingos. Discriminative training of markov logic networks. In *AAAI*, volume 5, pages 868–873, 2005. 108

[188] Parag Singla and Pedro Domingos. Lifted first-order belief propagation. In *AAAI*, volume 8, pages 1094–1099, 2008. 109

[189] L St-Laurent, X Maldague, and D Prévost. Combination of colour and thermal sensors for enhanced object detection. *Information Fusion, 2007 10th International Conference on*, pages 1–8, 2007. 10, 22, 35

[190] F Suard, A Rakotomamonjy, A Bensrhair, and A Broggi. Pedestrian Detection using Infrared images and Histograms of Oriented Gradients. *Intelligent Vehicles Symposium, 2006 IEEE*, pages 206–212, May 2006. 27

[191] Hao Sun, Cheng Wang, and Boliang Wang. Night Vision Pedestrian Detection Using a Forward-Looking Infrared Camera. *Multi-Platform/Multi-Sensor Remote Sensing and Mapping (M2RSM), 2011 International Workshop on*, pages 1–4, 2011. 24

[192] Zehang Sun, George Bebis, and Ronald Miller. On-road vehicle detection: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(5):694–711, 2006. 1, 2, 15

[193] K.K. Sung and T. Poggio. Example-based learning for view-based human face detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(1):39–51, 1998. 39

[194] Aditya Tatu, Francois Lauze, Mads Nielsen, and Benjamin Kimia. Exploring the representation capabilities of the HOG descriptor. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1410–1417, 2011. 27

[195] Carlo Tomasi and Takeo Kanade. *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ., 1991. 18

[196] H Torresan, B Turgeon, and C Ibarra-Castanedo. Advanced surveillance systems: combining video and thermal imagery for pedestrian detection. *Proc. of SPIE*, 2004. 22, 35

[197] Duan Tran and David A Forsyth. Configuration estimates improve pedestrian finding. In *Advances in neural information processing systems*, pages 1529–1536, 2007. 29

[198] O Tsimhoni. Pedestrian detection with near and far infrared night vision enhancement. *Transportation Research Institute (UMTRI)*, 2004. 10, 22

[199] Oncel Tuzel, Fatih Porikli, and Peter Meer. Human Detection via Classification on Riemannian Manifolds. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 30

[200] Oncel Tuzel, Fatih Porikli, and Peter Meer. Pedestrian Detection via Classification on Riemannian Manifolds. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(10):1713–1727, 2008. 30

[201] M.S. Uddin and T. Shioyama. Detection of pedestrian crossing and measurement of crossing length-an image-based navigational aid for blind people. In *Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE*, pages 331–336. IEEE, 2005. 8

[202] Paul Viola. Robust real-time face detection. *International Journal of Computer Vision*, 2004. 22

[203] Paul Viola and Michael J Jones. Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, 1:I–511–I–518 vol. 1, 2001. 24, 30, 66

[204] Paul Viola, Michael J Jones, and Daniel Snow. Detecting Pedestrians Using Patterns of Motion and Appearance. *International Journal of Computer Vision*, 63(2):153–161, February 2005. 24

[205] Stefan Walk, N Majer, Konrad Schindler, and Bernt Schiele. New features and insights for pedestrian detection. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1030–1037, 2010. 29

[206] Meng Wan and Jean-Yves Herve. Adaptive Target Detection and Matching for a Pedestrian Tracking System. In *2006 IEEE International Conference on Systems, Man and Cybernetics*, pages 5173–5178. IEEE, September 2006. 32, 121

[207] Qing Jun Wang and Ru Bo Zhang. LPP-HOG: A New Local Image Descriptor for Fast Human Detection. *Knowledge Acquisition and Modeling Workshop, 2008. KAM Workshop 2008. IEEE International Symposium on*, pages 640–643, 2008. 26

[208] Xiaoyu Wang, Tony X Han, and Shuicheng Yan. An HOG-LBP human detector with partial occlusion handling. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 32–39, 2009. 29

[209] Zhen-Rui Wang, Yu-Lan Jia, Hua Huang, and Shu-Ming Tang. Pedestrian Detection Using Boosted HOG Features. In *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*, pages 1155–1160, 2008. 30

[210] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. *Computer Vision-ECCV 2000*, pages 18–32, 2000. 28

[211] Chen Wei-Gang. Simultaneous object tracking and pedestrian detection using HOGs on contour. In *Signal Processing (ICSP), 2010 IEEE 10th International Conference on*, pages 813–816, 2010. 26

[212] Greg Welch and Gary Bishop. An introduction to the kalman filter, 1995. 117, 141

[213] D Williams and M Shah. A fast algorithm for active contours and curvature estimation. *CVGIP: Image Understanding*, 1992. 24

[214] Phil Williams, Karl Norris, et al. *Near-infrared technology in the agricultural and food industries.* American Association of Cereal Chemists, Inc., 1987. 10

[215] C. Wojek, G. Dorkó, A. Schulz, and B. Schiele. Sliding-windows for rapid object class localization: A parallel technique. *Pattern Recognition*, pages 71–81, 2008. 23, 26

[216] Christian Wojek, Stefan Roth, Konrad Schindler, and Bernt Schiele. Monocular 3D scene modeling and inference: understanding multi-object traffic scenes. In *ECCV'10: Proceedings of the 11th European conference on Computer vision: Part IV*. Springer-Verlag, September 2010. 35

[217] Bo Wu. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 2007. 28, 29

[218] Bo Wu and Ram Nevatia. Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 90–97 Vol. 1. IEEE, 2005. 23, 28

[219] Bo Wu and Ram Nevatia. Tracking of Multiple, Partially Occluded Humans based on Static Body Part Detection. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06)*, pages 951–958. IEEE, 2006. 33

[220] Bo Wu and Ram Nevatia. Cluster Boosted Tree Classifier for Multi-View, Multi-Pose Object Detection. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, 2007. 23

[221] Bo Wu and Ram Nevatia. Simultaneous object detection and segmentation by boosting local shape feature based classifier. *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8, 2007. 23

[222] Bo Wu and Ram Nevatia. Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008. 29

[223] Dong Xia, Hao Sun, and Zhenkang Shen. Real-time infrared pedestrian detection based on multi-block LBP. In *Computer Application and System Modeling (ICCASM), 2010 International Conference on*, 2010. 28

[224] Fen Xu, X Liu, and K Fujimura. Pedestrian Detection and Tracking With Night Vision. *IEEE Transactions On Intelligent Transportation Systems*, 2005. 32, 33

[225] Fengliang Xu, X Liu, and Kikuo Fujimura. Pedestrian Detection and Tracking With Night Vision. *IEEE Transactions On Intelligent Transportation Systems*, 6(1):63–71, 2002. 17

[226] Yuji Yamauchi, Chika Matsushima, Takayoshi Yamashita, and Hironobu Fujiyoshi. Relational HOG feature with wild-card for object detection. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1785–1792, 2011. 26

[227] Chen Yan-ping, Li Shao-zi, and Lin Xian-ming. Fast hog feature computation based on CUDA. In *Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on*, pages 748–751, 2011. 26

[228] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Computing Surveys (CSUR)*, 38(4), December 2006. 33

[229] Sang Min Yoon and Arjan Kuijper. 3D model retrieval using the histogram of orientation of suggestive contours. In *ISVC'11: Proceedings of the 7th international conference on Advances in visual computing*. Springer-Verlag, September 2011. 26

[230] Xiaoyan Yuan and Pengfei Shi. Iris feature extraction using 2d phase congruency. In *Information Technology and Applications, 2005. ICITA 2005. Third International Conference on*, volume 2, pages 437–441. IEEE, 2005. 48

[231] T.J. Yun, Y.C. Guo, and G. Chao. Human detection in far-infrared images based on histograms of maximal oriented energy map. In *Wavelet Analysis and Pattern Recognition, 2007. ICWAPR'07. International Conference on*, volume 2, pages 933–938. IEEE, 2007. 10

[232] F Zaklouta and B Stanciulescu. Real-time traffic sign recognition using spatially weighted HOG trees. In *Advanced Robotics (ICAR), 2011 15th International Conference on*, pages 61–66, 2011. 26

[233] Chengbin Zeng and Huadong Ma. Robust Head-Shoulder Detection by PCA-Based Multilevel HOG-LBP Detector for People Counting. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 2069–2072, 2010. 8, 26

[234] Li Zhang and Ramakant Nevatia. Efficient scan-window based object detection using gpgpu. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–7. IEEE, 2008. 26

[235] Li Zhang, Bo Wu, and Ram Nevatia. Pedestrian Detection in Infrared Images based on Local Shape Features. *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, May 2007. 23, 27

[236] Wei Zhang, Gregory Zelinsky, and Dimitris Samaras. Real-time accurate object detection using multiple resolutions. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007. 23

[237] L Zhao and C Thorpe. Stereo-and neural network-based pedestrian detection. *Intelligent Transportation Systems*, 2000. 19

[238] Q Zhu, M.C. Yeh, K.T. Cheng, and S. Avidan. Fast human detection using a cascade of histograms of oriented gradients. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2:1491–1498, 2006. 23, 26