Multi-modal Pedestrian Detection

A dissertation submitted by **Alejandro González Alzate** at Universitat Autònoma de Barcelona to fulfil the degree of **Doctor of Philosophy**.

Bellaterra, September 27, 2015

Director	**Dr. David Vázquez Bermúdez** Dept. Ciències de la Computació & Centre de Visió per Computador Universitat Autònoma de Barcelona
Co-Director	**Dr. Antonio López Peña** Dept. Ciències de la Computació & Centre de Visió per Computador Universitat Autònoma de Barcelona
Thesis committee	**Sergio Escalera Guerrero** Dept. de Matemàtica Aplicada i Anàlisi Universitat Barcelona **Ernest Valveny** Dept. Ciències de la Computació & Centre de Visió per Computador Universitat Autònoma de Barcelona **Luis M. Bergasa Pascual** Dept. de Electrónica Universidad de Alcalá

This document was typeset by the author using LaTeX 2$_\varepsilon$.

The research described in this book was carried out at the Centre de Visió per Computador, Universitat Autònoma de Barcelona.

ISBN: 978-84-943427-7-6

Printed by Ediciones Gráficas Rey, S.L.

To my parents and sister

Life is too short to wake up with regrets. So love the people who treat you right. Forget about those who don't. Believe everything happens for a reason. If you get a chance, take it. If it changes your life, let it. Nobody said life would be easy, they just promised it would most likely be worth it

Harvey MacKay (1932)

Acknowledgements

I would like to dedicate some lines to the people and institutions that have supported me during these years.

First, to the Computer Vision Center, the Universitat Autònoma de Barcelona, the *Personal Investigador en Formació* grant and TRA2010-21371-C03-01 project for supporting my research.

I would like to thank the members of the tribunal, Dr. Sergio Escalera, Dr. Ernest Valveny, and Dr. Luis M. Bergasa. I also want to thank the anonymous journal and conference reviewers for their enriching comments during my thesis.

During my Ph.D. I was fortunate to have several collaborations. I would like to thank D. Vazquez, J. Marín, and Y. Socarrás for their good collaborations in research for enriching the quality of the work. Also to the people involved in the ADAS car project: G. Ros, J. Xu, G. Villalonga and L. Sellart. Finally, to the remaining members of the ADAS group: Dr. J. Serrat, Dr. F. Lumbreras, Dra. K. Diaz, Dr. D. Ponsa, Dr. A. Sappa, P. Ganesh and Z. Fang.

I am truly indebted and thankful to my advisors, Dr. D. Vazquez and Dr. A. M. López. Thanks for the invested time, the extensive writing, the support, the patience, the advises. Thanks for the support and guidance they showed me throughout my dissertation writing. I am sure it would have not been possible without their help. I would like also to thank to my co-advisor J. Amores.

To all my friends from the CVC, especially, to the ones who started this Ph.D. with me: Camp, Joan, Carles, and Yainuvis. Also to: Ivet, Lluis, Fran, Toni, Jon, David, Monica, Marc and Anjan with whom I have shared so many interesting talks, lunches, coffees, parties and now I consider part of my best friends, and like my family here in BCN.

To all "5to 3ra" friends, especially, to Johanna and Andrea, with whom I have laugh a lot, had great talks and eternal philosophical discussions during more than 3 years living together. To all my friends from Colombia, for supporting me and for being part of my life all these years, specially the ones in Europe for the great trips during this adventure far away home.

Finally and more important thanks to all my family for being there always present in the distance, supporting me as nobody, believing in me always without doubts, for all the love and care messages. Mom, dad, and sister you are always in my mind giving me strength in the worst moments.

Abstract

Pedestrian detection continues to be an extremely challenging problem in real scenarios, in which situations like illumination changes, noisy images, unexpected objects, uncontrolled scenarios and variant appearance of objects occur constantly. All these problems force the development of more robust detectors for relevant applications like vision-based autonomous vehicles, intelligent surveillance, and pedestrian tracking for behavior analysis. Most reliable vision-based pedestrian detectors base their decision on features extracted using a single sensor capturing complementary features, *e.g.*, appearance, and texture. These features usually are extracted from the current frame, ignoring temporal information, or including it in a post process step *e.g.*, tracking or temporal coherence. Taking into account these issues we formulate the following question: *can we generate more robust pedestrian detectors by introducing new information sources in the feature extraction step?*

In order to answer this question we develop different approaches for introducing new information sources to well-known pedestrian detectors. We start by the inclusion of temporal information following the *Stacked Sequential Learning (SSL)* paradigm which suggests that information extracted from the neighboring samples in a sequence can improve the accuracy of a base classifier.

We then focus on the inclusion of complementary information from different sensors like 3D point clouds (LIDAR - depth), far infrared images (FIR), or disparity maps (stereo pair cameras). For this end we develop a multi-modal framework in which information from different sensors is used for increasing detection accuracy (by increasing information redundancy). Finally we propose a multi-view pedestrian detector, this multi-view approach splits the detection problem in n sub-problems. Each sub-problem will detect objects in a given specific view reducing in that way the variability problem faced when a single detectors is used for the whole problem. We show that these approaches obtain competitive results with other state-of-the-art methods but instead of design new features, we reuse existing ones boosting their performance.

Resumen

La detección de peatones continua siendo un problema muy difícil en escenarios reales, donde diferentes situaciones como cambios en iluminación, imágenes ruidosas, objetos inesperados, escenarios sin control y la variabilidad en la apariencia de los objetos ocurren constantemente. Todos estos problemas fuerzan el desarrollo de detectores más robustos para aplicaciones relevantes como lo son los vehículos autónomos basados en visión, vigilancia inteligente y el seguimiento de peatones para el análisis del comportamiento. Los detectores de peatones basados en visión más confiables deciden basándose en descriptores extraídos usando un único sensor y capturando características complementarias, *e.g.*, apariencia y textura. Estas características son extraídas de una única imagen, ignorando la información temporal, o incluyendo esta información en un paso de post procesamiento *e.g.*, seguimiento o coherencia temporal. Teniendo en cuenta estos hechos, nos formulamos la siguiente pregunta: *¿Podemos generar detectores de peatones más robustos mediante la introducción de nuevas fuentes de información en el paso de extracción de características?*

Para responder a esta pregunta desarrollamos diferentes propuestas para introducir nuevas fuentes de información a detectores de peatones bien conocidos. Empezamos por la inclusión de información temporal siguiendo el paradigma del *aprendizaje secuencial apilado (SSL siglas en inglés)*, el cual sugiere que la información extraída de las muestras vecinas en una secuencia pueden mejorar la exactitud de un clasificador base.

Después nos enfocamos en la inclusión de información complementaria proveniente de sensores diferentes como nubes de puntos 3D (LIDAR - profundidad), imágenes infrarrojas (FIR) o mapas de disparidad (par estéreo de cámaras). Para tal fin desarrollamos un marco multimodal en el cual información proveniente de diferentes sensores es usada para incrementar la exactitud en la detección (aumentando la redundancia de la información). Finalmente proponemos un detector multi-vista, esta propuesta multi-vista divide el problema de detección en n sub-problemas. Cada uno de estos sub-problemas detectara objetos en una vista específica dada, reduciendo así el problema de la variabilidad que se tiene cuando un único detector es usado para todo el problema. Demostramos que estas propuestas obtienen resultados competitivos con otros métodos en el estado del arte, pero envés de diseñar nuevas características, reutilizamos las existentes para mejorar el desempeño.

Resum

La detecció de vianants continua essent un problema molt difícil en escenaris reals, on diferents situacions com canvis en il·luminació, imatges sorolloses, objectes inesperats, escenaris sense control i la variabilitat en l'aparença dels objectes ocorren constantment. Tots aquests problemes forcen el desenvolupament de detectors més robustos per a aplicacions rellevants com poden ser els vehicles autònoms basats en visió, la vigilància intel·ligent i el seguiment de vianants per l'anàlisi del comportament. Els detectors de vianants basats en visió més fiables decideixen basant-se en descriptors extrets utilitzant un únic sensor i capturant característiques complementàries, com poden ser l'aparença i la textura. Aquestes característiques són extretes d'una única imatge, ignorant la informació temporal, o incloent aquesta informació en un pas de post processament. Tenint en compte aquests fets, ens formulem la següent pregunta: Podem generar detectors de vianants més robustos mitjançant la introducció de noves fonts d'informació en el pas d'extracció de característiques?

Per respondre a aquesta pregunta desenvolupem diferents propostes per introduir noves fonts d'informació a detectors de vianants ben coneguts. Comencem per la inclusió d'informació temporal seguint el paradigma de *l'aprenentatge seqüencial apilat (SSL sigles en anglès)*, el qual suggereix que la informació extreta de les mostres veïnes en una seqüència poden millorar l'exactitud d'un classificador base.

Després ens enfoquem en la inclusió d'informació complementària provinent de sensors diferents com núvols de punts 3D (LIDAR - profunditat), imatges infraroges (FIR) o mapes de disparitat (parell estèreo de càmeres). Per aquest fi desenvolupem un marc multimodal en el qual informació provinent de diferents sensors és usada per incrementar l'exactitud en la detecció (augmentant la redundància de la informació). Finalment proposem un detector multi-vista, aquesta proposta multi-vista divideix el problema de detecció en n sub-problemes. Cada un d'aquests sub-problemes detecta objectes en una vista específica, reduint així el problema de la variabilitat que es té quan un únic detector es fa servir per a tot el problema. Vam demostrar que aquestes propostes obtenen resultats competitius amb altres mètodes en l'estat de l'art, però en lloc de dissenyar noves característiques, reutilitzem les existents per millorar el rendiment.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Nowadays, due to increasing number of inhabitants in urban scenarios and consequent increasing number of vehicles in the cities, accidents caused by vehicle-to-human collisions are one of the principal mortal causes in urban scenarios. The World Health Organization (WHO) reports [65] that 1.24 million people dead due to traffic causalities, predicting that for 2030 traffic injuries will be the $5th$ cause of dead in the world. In order to reduce fatalities in traffic accidents authorities, universities, and media have elaborated education campaigns, new rules, and research funding to develop intelligent systems that reduce these accidents.

In particular, automotive companies are continuously introducing smarter Advanced Driver Assistance Systems (ADAS). ADAS aim to improve mobility and traffic safety by providing warnings and performing maneuvers in dangerous real life driving situations. Following this line one of the most dangerous situations is the collision vehicle-to-pedestrian. Pedestrian Protection Systems (PPS) try to avoid vehicle-to-pedestrian collisions by detecting accurately the presence of pedestrians in the vehicle path in order to warn the driver, perform braking actions, or even evasive maneuvers. Accordingly, since vision is the main sense in human driving, vision-based PPS have attracted a lot of attention. Vision-based PPS are based on image acquisition and a processing system able to detect pedestrians in real-time, always subject to an extremely low number of false alarms and missdetections. Pedestrian detection is a hard challenge not fully solved nowadays, because pedestrians present a very high appearance variability including: clothes, pose, accessories, point of view, and size. Also, they have to be detected on-board in real urban scenarios with problems like cluttered background, different weather and illumination conditions, and partial occlusions generated by other objects. During the last decade research on vision-based pedestrian detection for PPS has been a very relevant topic in the computer vision community, as is revealed in different comprehensive state-of-the-art reviews [24, 25, 32, 35, 42, 87].

The goal of a vision-based pedestrian detector is to localize all pedestrians in a given image, providing as output the 3D position relative to the vehicle of each of them. Usually vision-based pedestrian detection systems follow a common pipeline

which includes the following steps:*(i) candidate generation*, where given an image it provides windows that could contain pedestrians; *(ii) candidate description*, where features describing each candidate are extracted; *(iii) candidate classification*, where a label/confidence of containing or not a pedestrian is given to each candidate based on its features; *(iv) detection fusion*, where in case that two or more overlapped windows result from the same pedestrian they are merged into a single detection. To complement this detection phase for still images, detections are tracked over time for assuring temporal coherence, removing spurious detections, and obtaining pedestrian motion information such as speed or motion direction.

All the previous steps are important in the pedestrian detection pipeline, and can influence the final result for getting a more reliable approach. However, in this process we can identify two key steps, description and classification, which may affect in a deep way the general performance of the detector. In the description step the main goal is to capture the information that better represents the pedestrians. In the literature we find different types of descriptors that try to capture this information. For instance, based on appearance [17], texture [89] or movement [87]. While in classification the main goal is to assign a score/probability/label to a given candidate (window descriptor). This assignment is based on a previous learned model (classifier), which is trained with pedestrian samples together with background ones, and defining rules for separating both classes (classifier). In the literature we find different approaches that perform this classification step, some taking the pedestrian as a whole object (holistic models) [17,87,89], others defining the object as a set of parts [28,50,73], and other base its decision looking in random patches in the candidate (patch-based) [59].

As we mentioned before, there are diverse factors challenging pedestrians detection in real on-board scenarios, for instance there are temporal problems that appear in some instances modifying the objects perception in the scene. These type of temporal problems could be defined as those due to motion, *i.e.*, objects movement, vehicle movement (egomotion), noisy movement (car vibration). Those due to illumination changes that can produce saturated or hard shadowed regions where the detection process is much harder. Finally due to cluttered backgrounds in which objects (or bunch of objects) could *"look"* like pedestrians in a frame. These problems need to be faced in order to develop more robust detectors. Usually these temporal problems are addressed by the community in post-detection steps like using tracking or temporal coherence techniques. In this Thesis we will propose a novel way to introduce the temporal information at description level, obtaining significant improvements in different well-known pedestrian datasets.

In any PPS it is imperative to work during the whole day and under different weather conditions. Unfortunately visible spectrum cameras are affected by all these condition, perturbing the image acquisition. When visible spectrum cameras are exposed to sudden illumination changes, they acquire images with saturated/hard-shadowed areas. Moreover, when they acquire images during low illumination (night time), they produce images of low quality where pedestrians are seen only if they are well illuminated by street/car lights, which is not always the case. In fact, each type of sensor (Camera, RADAR, LIDAR, Ultrasound) has its own pros and cons; therefor,

it is of great importance including sensors which provide alternative information invariant to illumination and time conditions. Laser sensors acquiring 3D information, or far infrared (FIR) cameras acquiring thermal information; both provide information invariant to illumination conditions and complementary to the visual spectrum. Accordingly, in this Thesis we propose multi-modal detectors that outperform those based on a single modality. In particular, we asses the combination of dense LIDAR with the visual spectrum, as well as this spectrum with the FIR one. The different combinations are compared with the use of the corresponding isolated modalities.

1.1 Objectives

In summary, the objectives of this Ph.D. dissertation consist of addressing the following questions:

- *How to introduce temporal information in the classification stage of a pedestrian detector?*
- *The combination of depth with visual information does improve the use of these modalities in isolation?*
- *The combination of FIR with visual spectrum information does improve the use of these modalities in isolation?*

In the long term, our goal is to build a pedestrian detection system robust enough for operating under different time conditions, using information coming from different sensors.

1.2 Contributions

Answering these questions lead to several novel contributions:

- The use of Stacked Sequential learning for incorporating both spatial and temporal information at the classification stage.
- Development of a Multi-view, Multi-modal Random Forest of Local Experts based on RGB-D information. In this case, The "D" information stands for either depth or disparity. In the former case the data is obtained from dense laser data. In the latter, a stereo rig is used.
- Development of different Multi-modal pedestrian detectors based on different types of models that incorporate FIR information as well as RGB. In this case, we provide a comprehensive study with special focus on assessing the differences between day and night time.

Along this Ph.D. dissertation all our experiment are based on well established protocols and publicly available datasets. In fact, as an additional contribution of this Thesis we have acquired and annotated different new pedestrian datasets which will be described along the corresponding chapters. It is worth to mention that one of such datasets include RGB and FIR images acquired at the same place at the same time, covering day and night time.

1.3 Outline

The rest of the Thesis is organized as follows. In Chapt. 2 we review the literature related to our proposals. Chapt. 3 presents and discusses the results obtained by using our proposal based on stacked sequential learning. In Chapt. 4 we present the results obtained using our above mentioned multi-view RGB-D approach. In Chapt. 5 we present the study based on RGB and FIR data. Finally, Chapt. 6 draws the main conclusions of the presented work.

Chapter 2

State of the art

In this chapter we review the literature in order to provide insights of the state-of-the-art to our proposals. This review include works related to pedestrian detection general scheme (Section 2.1) including: Candidate generation, detectors (description and classification), post-detection process; detectors that extract spatiotemporal information (Section 2.2), including context information; finally review related detectors based on other sensors: LIDAR or FIR, and multi-modal approaches (Section 2.3).

2.1 Pedestrian Detectors Scheme

Pedestrian detection methods usually follow a general pipeline [35] (Figure 2.1). This pipeline starts after the image acquisition (input image). The first step is the *preprocessing*, where the input image is processed in order to prepare it for the next steps. The second step is the *foreground segmentation*, where the areas of interest of the image are identified, by defining pedestrian candidates (window/ROI) covering those areas (Subsection 2.1.1). The third step is the *classification*, which takes as input the candidates of the previous step for describing and classifying them (Subsection 2.1.2). The fourth step is the *refinement and verification*, where overlapped detections are merged in a single one for final verification (Subsection 2.1.3). The next steps are optional like tracking (Subsection 2.1.3) or the application. In following subsections we will review the different methods related with each of the detection steps, providing a wide view about current pedestrian detection approaches and their impact in the community.

2.1.1 Foreground Segmentation - Candidate Generation

In this section we discuss different approaches for candidate generation. The most common candidate generation technique is the sliding window approach. Most suc-

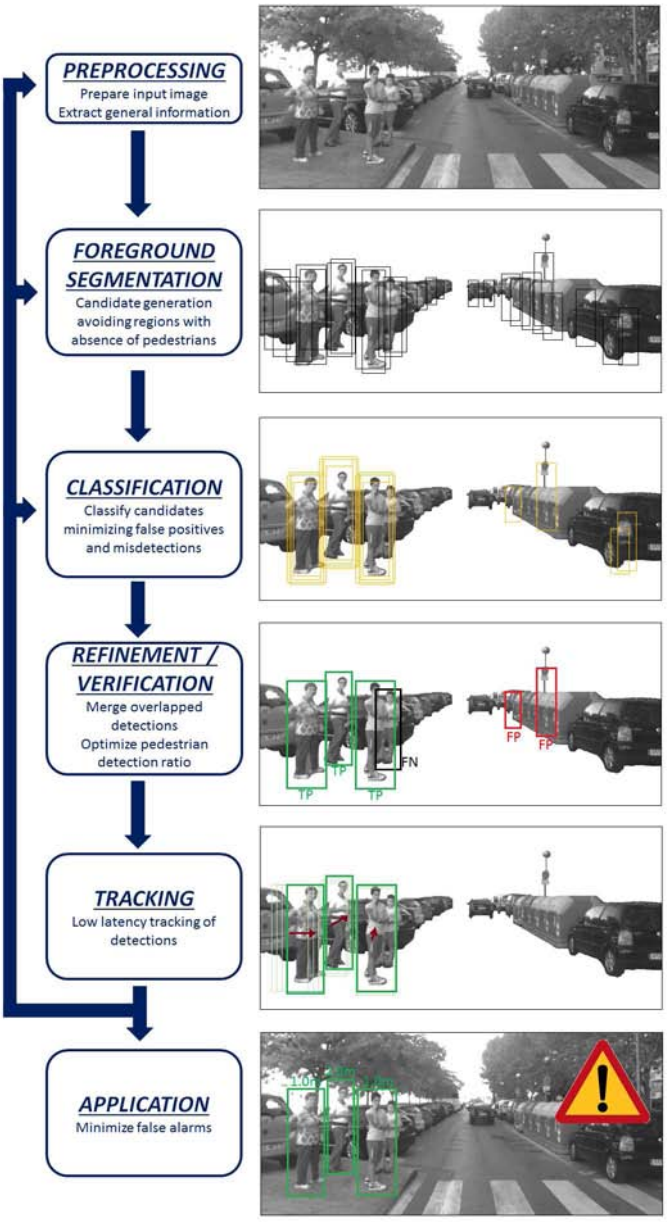


Figure 2.1: Pedestrian detection system general scheme proposed in [35].

cessful pedestrian detection methods in the literature base their detection on a sliding window strategy. One of the first authors that applied this technique for detection was *Papageorgiou et al.* in [68]. Then in [23] *Dollár et al.* remark that non-sliding

window approaches such as segmentation [38], or key point [53, 77], usually fail for low to medium resolution pedestrians. Sliding window based methods usually use an image pyramid in order to handle different detection scales; few other methods propose instead of rescaling the image to apply multi-scale classifiers [7] over the image for efficiency purposes.

Sliding window approach, by doing an exhaustive scanning, ends up with a set of regularly spaced candidates to be sent to the classification stage (See figure 2.2 (a)). As advantages this technique allows us to scan the full image without excluding any possible region in it, but brings two main disadvantages: 1) it generates a large number of possible candidates (usually thousands of them), making it unfeasible to achieve a real-time performance, and 2) irrelevant regions are also scanned, which may increase the number of false positives. Taking these facts into account, techniques, which purpose is to reduce the number of candidates and avoid irrelevant regions of the image, have seen the light.

(a) (b)

Figure 2.2: Basic candidate generation methods for pedestrian description. (a) sliding window, and (b) linear to road approach.

In this line, when stereo/3D information is available geometric constraints can be applied [4, 37] in the candidate generation process. Assuming that pedestrians must be standing at the ground plane, in [37] a candidate generation based on detecting the road plane from the 3D information, and then uniformly distributing the candidate windows over the road (See figure 2.2 (b)) and projecting them to the image plane (e.g. to the left image of the stereo pair). In [4], candidates are generated according to a clustering based on 3D point density. Then a set of candidates is generated for each cluster.

In order to avoid exhaustive sliding window search across images methods based on 2D information has born. In [41] *Hosang et al.* perform a comparison from many of these candidate generation methods. In this comparison there are methods based on segmentation (superpixels [29]) [74,84]; others based on saliency information [3,10,30]; others based on the graph cut method [9,54]; and finally other authors start by using sliding window and then filtering the candidates [13,38].

2.1.2 Detectors

Computer vision researchers have been following different research lines for improving the localization of humans in images. This is a challenging task with more than a decade of history by now and as a result, a plethora of features, models, and learning algorithms have been proposed to develop the pedestrian classifiers which are at the core of pedestrian detectors [36].

One of the researching lines for boosting the accuracy of pedestrian classifiers is in developing image descriptors well-suited for pedestrians. These descriptors are designed for capture different features, which differentiate between pedestrians and background. In this line descriptors based on appearance, meaning contours and shape, have come out. The Histogram of Oriented Gradient (**HOG**) descriptor presented by *Dalal et al.* in [17] captures the object appearance based on the idea that the human body has a characteristic shape (vertical contours in both sizes, low contours rate in the central part). In order to capture this shape, HOG descriptor uses histograms of the gradient orientations; these histograms are computed over overlapped blocks regular distributed across the window; in this way pedestrian contours are captures (See figure 2.3 top). *Papageorgiou et al.* in [68] propose to use Haar-wavelets for obtain structural information by filtering the image with them, based in this method another appearance based descriptor is the Speeded Up Robust Features (**SURF**) descriptor presented by *Bay et al.* in [5]. This descriptor is based on the responses of Haar wavelets, these responses are regular spatially distributed, and provide us with information of changes in intensity around the region of interest; capturing in this way the human body shape (See figure 2.3 bottom). Other descriptors try to capture texture information. This is the case of the Local Binary Pattern (**LBP**) presented by *Ahonen et al* in [2]. LBP captures object texture by defining unique labels to each different texture pattern in a $3X3$ pixel neighborhood, then a histogram of these labels describes a given block (See figure 2.3 middle). *Wang et al* in [89] combine HOG and LBP features for final description capturing complementary information. *Gerónimo et al.* in [34] combine the Edge Orientation Histograms (**EOH**) with Haar wavelets, resulting a robust and fast pedestrian detector; *Walk et al.* combine appearance features with color self similarity histograms (**CSS**) [87]. Other authors focus on fast and robust features like Integral Channels [22] or Macrofeatures [60].

The second direction that the researchers have taken, is focused on design pedestrian models. Most of the above descriptors deals with pedestrians as a whole objects, learning a *holistic* model by fitting a Support Vector Machine (**SVM**) or Adaptive Boosting (**AdaBoost**. SVM learns the hyperplane that better split the samples (positives and negatives) in the feature space (HOG, LBP, EOH, etc). AdaBoost learns *"weak"* classifiers (decision trees) and combining their responses represents the final output of the boosted classifier. Other authors define pedestrians as a set of parts, Deformable Multi-component Part-based Models (**DPM**) [28, 50, 73], and bases the final decision in the individual parts detection and in a deformation cost for each part (based in the relative position in the window). Other authors define pedestrian as a set of diferent resolution models, each model detect pedestrians in a given scale [7,70]). Other authors recently have developed detectors based on deep learning paradigm;

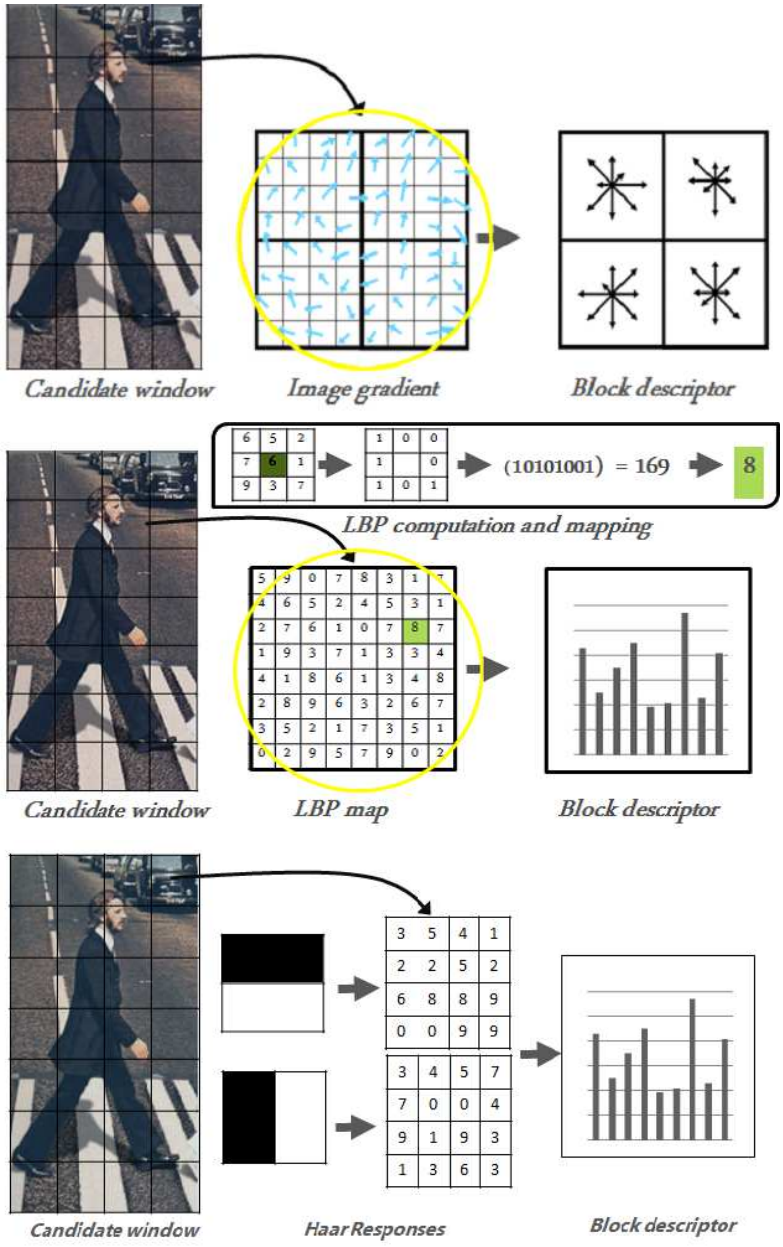
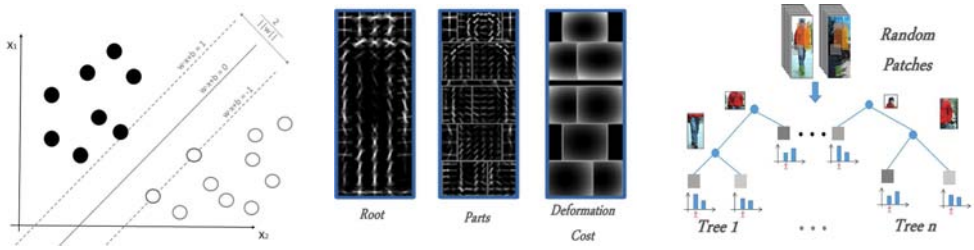


Figure 2.3: Basicfeatures for pedestrian description. *Top* HOG descriptor, *middle* LBP descriptor, and *bottom* Haar descriptor.

providing frameworks where Convolutional Neural Network (CNN) are included for pedestrian detection [40, 66, 94]

A third research line is define by the the classification architecture proposed in order to get robust pedestrian detectors. In this line is worth to mention, HOG-SVM/LRF-MLP cascades [63], Haar+EOH-AdaBoost cascades with meta-stages [12], distributed detections HOG/DOD [67], and ensemble of trees [96]. *Marin et al.* define the Random Forest of Local Experts (**RF**) in [59], which based on HOG and LBP features it learn SVM as local experts at each node in the trees. These SVM are learned over random patches defined by the HOG and LBP description cells.



(a) (b) (c)

Figure 2.4: Basic models for pedestrian description. (a) Holistic SVM, (b) part-based DPM, and (c) patch-based RF.

A fourth line followed by researchers is looking for the collection of *"good"* samples for training. *Enzweiler et al.* propose a generative approach method in [27], *Abramson el al.* in [1] propose an active learning method called SEVILLE (SEmi-atomatic VIsual LEarning) for sample selection, finally *Vazquez el al.* in [85] propose to use virtual-world data with domain adaptation for avoiding manual annotation of sequences.

2.1.3 Post-detection Methods

After the detection step with the multi-scale sliding window framework we will obtain several overlapped detection with different sizes for each real object. In order to obtain a clean result it is necessary to deal with these overlapped detections, and find a way to fuse them in a single detection per object. In order to fuse the overlapped detections it is commonly used a non-maximum suppression algorithm. Dalal et al. pesent in [16] a non-maximun suppression algorithm that generates the fusion of multiple detection in a single one. This algorithm implements a mean shift framework, representing each detection in a 3D space (position and scale). Other authors propose a iterative process in which detection are fused with its overlapped ones keeping the detection with higher probability/score.

2.1.4 Evaluation Protocols

In order to compare the accuracy of different detectors applied to the same images it has been defined different measurements. False Positives per Window (**FPPW**) capture the performance of the detector when the detector runs over independent

crops (positives and negatives). FPPW curve is calculated by varying the operation point of the detector (threshold or probability for defining a positive detection) and counting missdetections and false positives. Then pairs of (misdetection ratio; false positives ratio) are calculated for each operation point. This measure was used when no large sequences of images were available and detection was evaluated over few images or crops. When large sequences were annotated and created for pedestrian detection tasks, the measure changes to False Positives per Image (**FPPI**) following the same principle that FPPW, missdetections and false positives are counted in each image and then average miss rate and false positive rate per image is computed. This measurement is extended for pedestrian detection in [24, 25, 35]. This FPPI measurement is used in order to compare methods. The average miss rate (AMR) in the range of 10^{-2} to 10^{0} FPPI is taken as indicative of each detector accuracy, *i.e.*, the lower the better. There is other measurement recently used in pedestrian detection benchmarks, in which the precision-recall curve is computed and use the average precision (AP) as accuracy measurement, *i.e.*, the higher the better.

2.2 Spatiotemporal information

By now we only present descriptors based in information extracted from a single frame. However, in the community researchers have done improvements by adding extra information extracted from the context of the window. This context can be defined as a spatial context, which extracts information of the surroundings of the window for enhancing the detection, or as temporal information where features from neighboring frames are extracted.

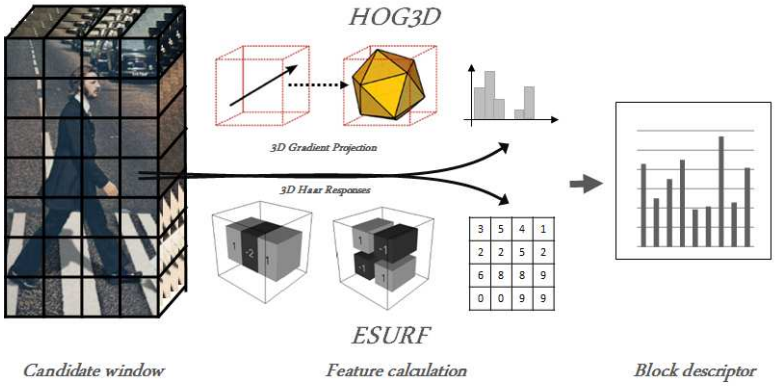


Figure 2.5: Spatiotemporal descriptor. *Top* HOG3D descriptor, and *bottom* ESURF descriptor.

In this line the histogram of 3D gradients (**HOG3D**) descriptor is an extension of the HOG descriptor introduced by *Kläser et al.* in [48]. The main idea is expanding the gradient definition in the temporal dimension. It extracts the features in the same

way HOG descriptor does, but defining the orientation bins based on the projection of the gradients over a 3D regular polyhedron (See Fig. 2.5 top). Another descriptor born from the idea of including temporal information is the Extended SURF descriptor (**ESURF**) proposed by *Willems et al.* in [91]. ESURF is an extension in the time-space of the SURF descriptor. In order to expand the descriptor they generate Haar masks in 3D, and use them for the descriptor calculations. This change allows them to obtain not only the spatial illumination changes but also the illumination changes presented over the time (See Fig. 2.5 bottom); other authors also propose temporal extension to haar-like features [15, 44, 45, 86]. Popular HOG descriptor was also extended to encode temporal information for detecting humans [18]. In this case using optical flow to compensate motion. In the same spirit the histograms of flow (**HOF**) were also introduced for detecting pedestrians [87]. In all cases motion information was complemented with appearance information (*i.e.*, Haar/HOG for luminance and/or color channels).

Regarding context information focusing on single frames, it has been recently shown how pedestrian detection accuracy can be boosted by analyzing the image area surrounding potential pedestrian detections. In particular, [11, 21] propose an iterative method in which responses obtained in neighboring areas are merged to enhance spatial coherent detections, while spurious ones are vanished.

2.3 Detection under different modalities

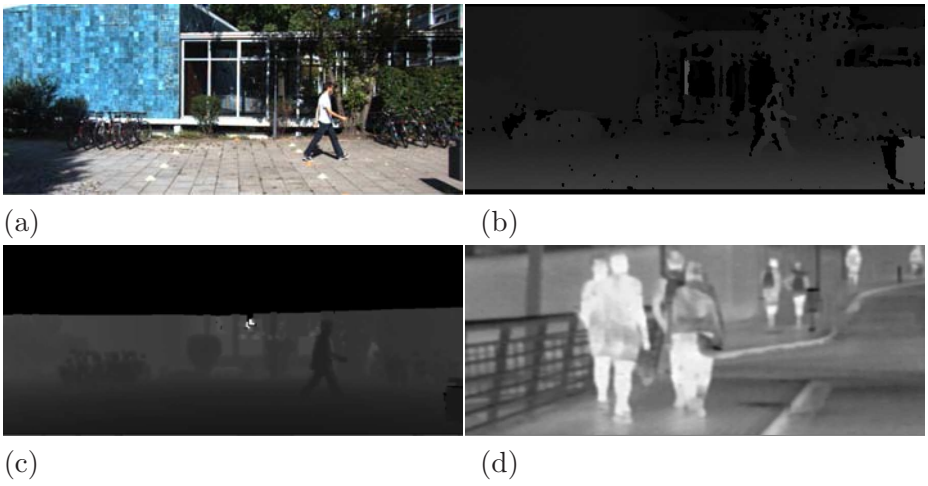


Figure 2.6: Different image modalities. (a) RGB image, (b) Disparity from stereo pair, (c) dense depth map from LIDAR data, and (d) FIR image.

Up to now all the mentioned methods base their detection in a single color/grey image acquired by a normal visible spectrum camera. By using more than one image researchers have developed detectors that extract information from different modali-

ties. These extra modalities provide complementary information like the mentioned motion features. In [92] *Wojek el al.* propose a variation of HOF features combining appearance and motion. Others extract information from dense stereo reconstruction. In [26] *Enzweiler et al.* propose a detector that combines appearance, depth, and motion.

Going further researchers propose to extract information from alternative sensors. These sensors try to solve problems of visible spectrum cameras, like problems with illumination changes or acquiring *"good"* images in low light conditions. In order to fulfill the information lost in these scenarios sensors based on laser beams or thermal/far-infrared cameras are taking relevance in the computer vision community for object detection.

Recently authors are starting to study the impact of high-definition 3D LIDAR [6, 46, 47, 61, 72, 79, 95] in pedestrian detection. Most of these works propose specific descriptors for extracting information directly from the 3D cloud of points [6, 46, 47, 61, 79, 95], but these methods work well in static controlled scenarios in which few objects appear. A common approach is to detect objects independently in the 3D cloud of points and in the visible spectrum images, and then combining the detections using an appropriate strategy [46, 47, 95]. Most relevant approach, looking for multi-modal detectors, is the one presented by *Premebida et al.* in [72], in this approach they propose to densify the 3D cloud of points to obtain a dense depth maps; first registering the 3D cloud of points captured by a Velodyne sensor with the RGB image captured with the camera, and then interpolating the resulting sparse set of pixels to obtain a dense map where each pixel has an associated depth value. Given this map the authors perform detection over both images separately for then merge detection.

Looking in other direction some authors try to solve the problem of acquiring image during night time with enough information for detecting pedestrians. Taking into account that thermal information is invariant to illumination conditions, and with the increasing resolution of far infrared (**FIR**) cameras, researchers have starting to use FIR cameras in detection problems. There are applications relying on video surveillance [19, 20] using static cameras (zenital position) and tracking objects [71]. All these approaches work in controlled scenarios where cameras are in a fixed position, and objects to detect are the only non fixed objects in the scene. Recently [43, 49, 64] proposed methods for extracting information in non-controlled scenarios for pedestrian detection. *Hwang et al.* in [43] propose a multi-model approach inspired in [23] adding FIR images as new channels in the description process.

Chapter 3

Stacked Sequential Learning

Pedestrian detectors base the detection on the responses obtained by aplying a classifier to decide which image windows contain a pedestrian. These responses usually provide with relatively high response to neighboring windows overlapped with a real pedestrian, while the responses around potential false positives are expected to be lower. Applying a non-maximun suppression algorithm turn these overlapped high responses windows in a single detection, but false positives remain without changes. Same coherence is expected for image sequences. If there is a pedestrian located within a frame, the same pedestrian is expected to appear close to the same location in neighbor frames. This location has chances of receiving high response during several frames, while false positives are expected to be more spurious. Following this expected behavior in this chapter we propose a method to exploit such correlations for improving the accuracy of base pedestrian classifiers. To this end we propose a method that introduce information of this spatiotemporal behavior at description level. In order to validate our proposal it will be tested over different well-known pedestrian detection datasets.

3.1 Introduction

The outcome of a pedestrian classifier, termed here as *base classifier*, determines if a given image window contains a pedestrian or background by assigned to it a score/probability. In practice, such classifiers provide a relatively high response at neighbor windows overlapping a pedestrian, while the responses surrounding non-pedestrian windows are expected to be lower. In fact, non-maximum suppression (NMS) is usually performed as last detection stage in order to reduce multiple detections arising from the same pedestrian to a single one. The same reasoning applies when we detect in image sequences. If in a given location is detected a pedestrian, high classification scores are expected in the same location of neighboring frames, while false positives are expected to be more spurious. Usually, these spourious de-

tection may be removed by the use of a tracking algorithm. In this chapter we propose to exploit such expected *response correlations* for improving the accuracy of the classification stage itself. Instead of only exploiting spatiotemporal coherence by means of general post-classification stages like NMS and tracking, we propose to add such a type of reasoning in the classification stage itself as well. In particular, we propose to use two-stage classification strategy which not only relies on the image descriptors required by the base classifiers, but also on the response of the own base classifiers in a given spatiotemporal neighborhood. More specifically, we train pedestrian classifiers using a stacked sequential learning (SSL) paradigm [14].

Temporal SSL involves the analysis of window temporal volumes. The different types of temporal volumes can be potentially useful for different applications depending on the motion of the camera and the targets of interest, as well as the working frame rate. As we are specially interested in on-board pedestrian detection within urban scenarios, we will face camera and targets movements. Accordingly, we test our SSL approach for a fixed neighborhood (*i.e.*, fixed spatial window coordinates across frames) and for an scheme relying on an ego-motion compensation approximation (*i.e.*, varying spatial window coordinates across frames). Moreover, in order to assess the dependency of the results with respect to the frame rate, we acquired our own pedestrian dataset at 30fps (**CVC08 dataset**) by normal driving in an urban scenario. This new dataset is used as main guide for our experiments, but we also complement our study with other challenging datasets publicly available.

To perform this study, we start by using a competitive baseline in pedestrian detection [24], namely a holistic base classifier based on HOG+LBP features and linear SVM. Note that HOG/linear-SVM is the core of more sophisticated pedestrian detectors as the popular deformable part-based model (DPM) [28]. Moreover, HOG with LBP are also used as base descriptors of multi-modal multi-view pedestrian models [26], and HOG+LBP/linear-SVM has been used for classifiers with occlusion handling [58,89], as well as for acting as node experts in random forest ensembles [59]. In addition, it has recently been shown that HOG+LBP/linear-SVM approaches are well suited for domain adaptation [85]. Altogether, we think that HOG+LBP/linear-SVM is a proper baseline to start assessing our proposal. Moreover we have extended this baseline with the HOF [87] motion descriptor that complements the appearance and texture features of the baseline.

In this chapter we will conduct experiments over the new CVC08 dataset together with the well known datasets: Caltech, Daimler, CVC02, KITTI. The obtained results show that our SSL proposal boosts detection accuracy significantly with a minimal impact on the computational cost. Interestingly, SSL improves more the accuracy at the most dangerous situations, *i.e.* when a pedestrian is close to the camera.

This chapter is organized as follows. In section 3.2 we review some works related to our proposal. Section 3.3 briefly introduces the SSL paradigm. In section 3.4 we develop our proposal. Section 3.5 presents the experiments carried out to assess our spatiotemporal SSL, and discuss the obtained results. Finally, section 3.6 draws our main conclusions.

3.2 Related Work

The use of motion patterns as image descriptors was already proposed as an extension of spatial Haar-like filters for video surveillance applications (static zenital camera) [15, 44, 86] and for detecting human visual events [45]. In these cases, original spatial Haar-like filters were extended with a temporal dimension. Popular HOG descriptor was also extended to encode temporal information for detecting humans [18], in this case using optical flow to compensate motion. In the same spirit the histograms of flow (HOF) were also introduced for detecting pedestrians [87]. In all cases motion information was complemented with appearance information (*i.e.*, Haar/HOG for luminance and/or color channels).

In contrast with these approaches, our proposal does not involve to compute new temporal image descriptors as new features for the classification process. As we will see, we use the responses of a given base classifier in neighbor frames as new features for our SSL classifier. In fact, our proposal can be also applied to base classifiers that already incorporate motion features. Therefore, the reviewed literature and our proposal are complementary strategies.

Focusing on single frames, it has been recently shown how pedestrian detection accuracy can be boosted by analyzing the image area surrounding potential pedestrian detections. In particular, [11, 21] follow an iterative process that uses contextual features of several orders (*e.g.*, involving co-occurences) for progressively enhancing the response of base classifiers for true pedestrians and lowering it for hallucinatory ones. Our SSL proposal does not require new image descriptors of pedestrian surroundings and is not iterative, which makes it inherently faster. Moreover, we treat equally spatial and temporal response correlations, *i.e.*, under the SSL paradigm, giving rise to a more straightforward method.

Finally, we would like to clarify that our SSL proposal is not a substitute for NMS and tracking post-classification stages. What we expect is to allow these stages to produce more accurate results by increasing the accuracy of the classification stage. For instance, tracking must be used for predicting pedestrian intentions [76], thus, if less false positives reach the tracker, we can reasonably expect to obtain more reliable pedestrian trajectories and so guessing intentions in the very short time this information is required (*i.e.*, around a quarter of second before a potential collision).

3.3 Stacked sequential learning (SSL)

Stacked sequential learning (SSL) was introduced by Cohen *et al.* [14] with the aim of improving base classifiers when the data to be processed has some sort of sequential order. In particular, given a data sample to be classified, the core intuition is to consider not only the features describing the sample but also the response of the base classifier in its neighbor samples. Figure 3.1 summarizes the SSL learning process that we explain in more detail in the rest of this section.

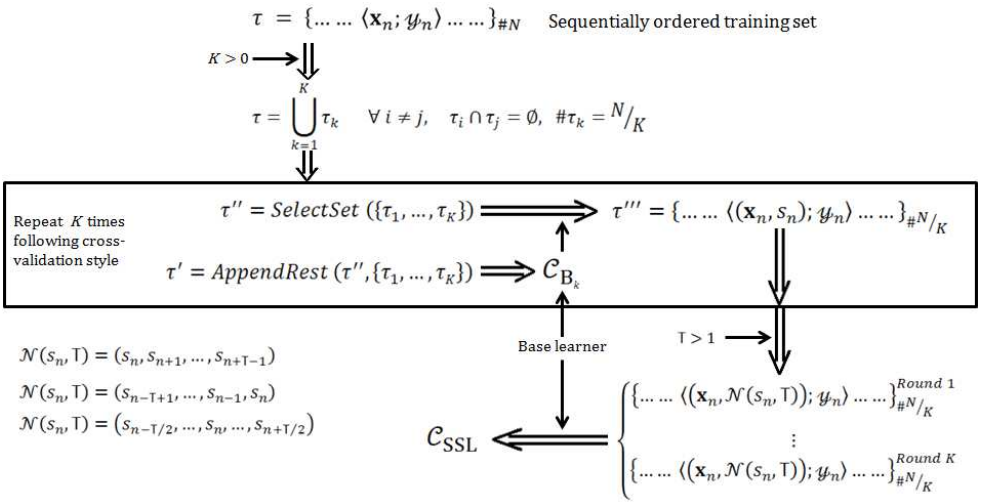


Figure 3.1: SSL learning. See main text in Sect. 3.3 for details

Let τ be an ordered training sequence of cardinality N. In order to avoid overfitting, the SSL approach involves to select a sub-sequence for training a base classifier, $\mathcal{C}_B$, and the rest to apply $\mathcal{C}_B$ and so training the SSL classifier, $\mathcal{C}_{SSL}$. If this is done once, then the final classifier $\mathcal{C}_{SSL}$ would be trained with less than N samples. Thus, to avoid this, it is followed a cross-validation style were τ is divided in $K > 0$ disjoint sub-sequences, $\tau = \cup_{k=1}^{K} \tau_k \wedge i \neq j \Rightarrow \tau_i \cap \tau_j = \emptyset$, and K rounds are performed by using a different subset each round to test the $\mathcal{C}_{B_k}$ and the rest of subsets for training this $\mathcal{C}_{B_k}$. At the end of the process, joining the K sub-sequences processed by the corresponding $\mathcal{C}_{B_k}$, we can have N *augmented* training samples for learning $\mathcal{C}_{SSL}$. $K = 1$ means to train the C_B and C_{SSL} on the same training set, without actually doing partitions.

Let us explain what means *augmented* training samples. The elements of τ, *i.e.*, the initial training samples, are of the form $< \mathbf{x_n}; y_n >$, where $\mathbf{x_n}$ is a vector of features with associated label y_n. Therefore, the elements of each sub-sequence τ_k are of the same form. As we have mentioned before, during each round k of the cross-validation-style process, a sub-sequence τ'' is selected among $\{\tau_1, \ldots, \tau_K\}$, while the rest are appended together to form a sub-sequence τ'. From τ' it is learned $\mathcal{C}_{B_k}$ and applied to τ'' to obtain a new τ'''. The elements of τ''' are of the form $< (\mathbf{x_n}, s_n); y_n >$, where we have augmented the feature $\mathbf{x_n}$ with the classifier score $s_n = \mathcal{C}_{B_k}(\mathbf{x_n})$. Therefore, after the K rounds, we have a training set of N samples of the form $< (\mathbf{x_n}, s_n); y_n >$. It is at this point when we can introduce the concept of neighbor scores into the learning process. In particular, the final training samples are of the form $< (\mathbf{x_n}, \mathcal{N}(s_n, T)); y_n >$, where $\mathcal{N}(s_n, T)$ denotes a neighborhood of size $T > 1$ anchored to the sample n. For instance, $\mathcal{N}(s_n, T) = (s_{n-T+1}, \ldots, s_{n-1}, s_n)$ is a *past* neighborhood, $\mathcal{N}(s_n, T) = (s_n, s_{n+1}, \ldots, s_{n+T-1})$ is a *future* neighborhood, and $\mathcal{N}(s_n, T) = (s_{n-\frac{T}{2}}, \ldots, s_n, \ldots, s_{n+\frac{T}{2}})$ is a *centered* neighborhood, which are analogous

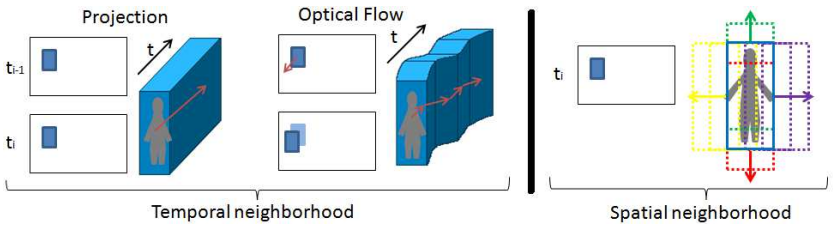


Figure 3.2: Different types of neighborhood for SSL. See main text in Sect. 3.4.1 for details.

concepts to the ones of filtering, extrapolation and smoothing, resp., used in the classical tracking literature.

3.4 SSL for pedestrian detection

In this section, without loosing generality, we will assume the use of the *past neighborhood* (Sect. 3.3) to illustrate and explain our SSL approach (use previous images to do detection in the current one). Actually there is no need to save the previous images, saving the already computed scores is enough to compute the current SSL descriptor making the computation of SSL very computational efficient.

3.4.1 Spatiotemporal neighborhoods for SSL

For object detection in general and for pedestrian detection in particular, applying SSL starts by defining which are the neighbors of a given window under analysis. In learning time, such a window will correspond either to the bounding box of a labeled pedestrian or to a rectangular chunk of the background. In operation time (*i.e.*, testing), such a window will correspond to a candidate generated by a pyramidal sliding window scheme or any other candidate selection method. In this paper we assume the processing of image sequences and, consequently, we propose the use of a spatiotemporal neighborhood.

Temporal SSL involves the analysis of window volumes. Therefore, there are several possibilities to consider (see Fig. 3.2). Let us term as W_f the set of coordinates defining an image window in frame f, and $\mathbf{V}_f = \text{vol}(\cup_{t=0}^{T-1} W_{f-t})$ the window volume defined by a temporal neighbor of T frames. The simplest volume is obtained by assuming fixed locations across frames, which we term as *projection* approach. In other words, $W_f = W_{f-1} = ... = W_{f-(T-1)}$. Another possibility consists in building volumes taking into account motion information. For instance, $W_f = W_{f-1} + t_{OF(W_{f-1})}$, where $t_{OF(W_{f-1})}$ is a 2D translation defined by considering the *optical flow* contained in W_{f-1}, and $'+'$ stands for summation to all coordinates defining W_{f-1}.

Spatial SSL involves the analysis of windows spatially overlapping the window of interest (see Fig. 3.2). For instance, we can fix a 2D displacement $\Delta = (\delta_x, \delta_y)$ and

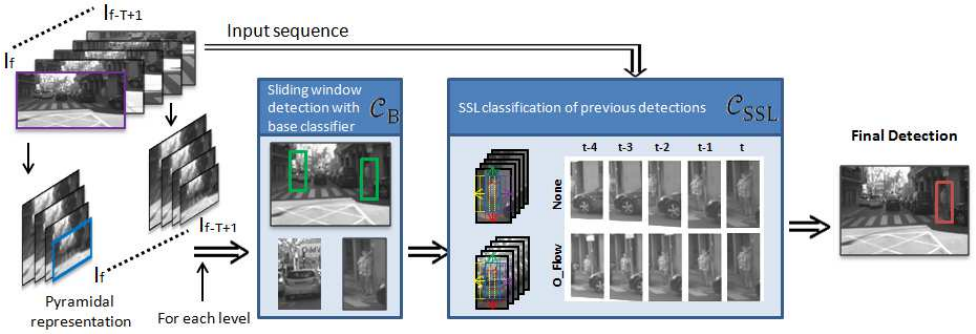


Figure 3.3: Two-stage pedestrian detection based on SSL. See main text in Sect. 3.4.3 for details.

n_x displacements in the x axis, to the left and to the right, an analogously for the y axis given a n_y number of up and down displacements.

Our proposal combines both ideas, *i.e.*, the temporal volumes and the spatial overlapping windows, in order to define the spatiotemporal neighborhood required by SSL (Sect. 3.3).

3.4.2 SSL training

As usual, we assume an image sequence with labeled pedestrians (*i.e.*, using bounding boxes) for training. Negative samples for training are obtained by random sampling of the same images, of course, these samples cannot highly overlap labeled pedestrians. The cross-validation-style rounds of SSL (Sect. 3.3) are performer with respect to the images of the sequence, not with respected to the set of labeled pedestrians and negative samples as it may suggest the straightforward application of SSL (note that pedestrian/negative labels are for individual windows not for full images). Moreover, as we have seen in Sect. 3.4.1, the neighborhood relationship is not only temporal but spatial too. The training process is divided in two stages. First, we train the auxiliary classifiers ($\mathcal{C}_{B_k}$) as usual, using three bootstraping rounds. Then we train the SSL classifier (using final $\mathcal{C}_{B_k}$ as auxiliary), again we run three bootstrapping rounds for obtaining the final classifier ($\mathcal{C}_{SSL}$).

Using the full training dataset, we also assume the training of a base classifier $\mathcal{C}_B$. Another possibility is to understand the different $\mathcal{C}_{B_k}$ as the result of a bagging procedure and ensemble them to obtain $\mathcal{C}_B$. Without loosing generality, in this paper we have focused on the former approach.

3.4.3 SSL detector

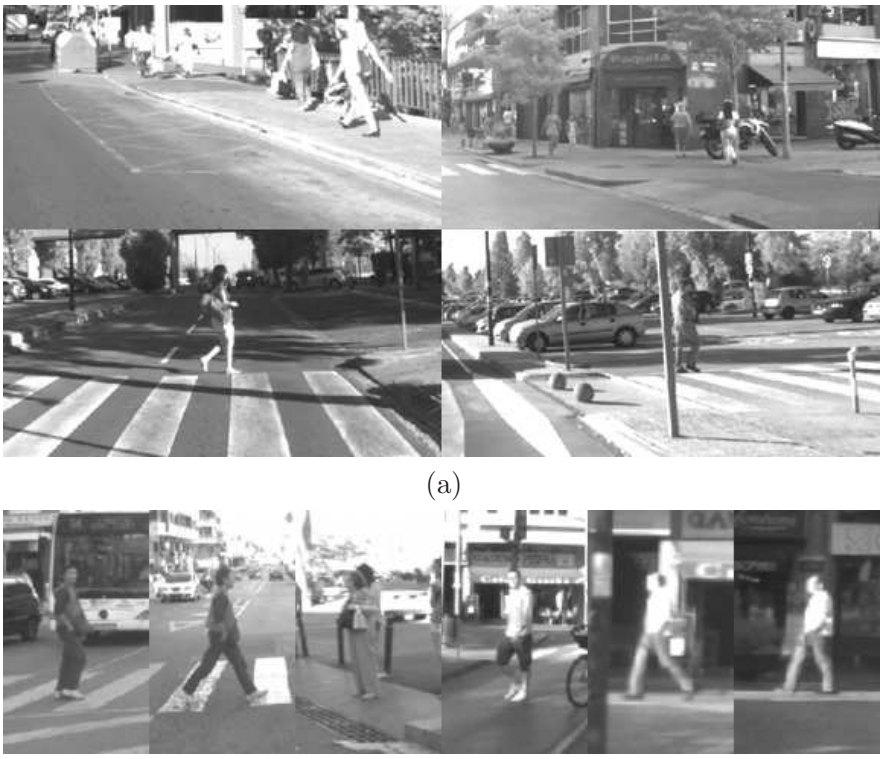
The proposed pedestrian detection pipeline is shown in Fig. 3.3. As we can see there are two main stages. The first stage basically consists in a classical pedestrian detection method relying on the learned base classifier $\mathcal{C}_B$. In Fig. 3.3 we have illustrated the idea for a pyramidal sliding window approach, but using other candidate selection approaches is also possible. Detections at this stage are just considered as potential ones. Then, the second stage applies the spatiotemporal SSL classifier, $\mathcal{C}_{SSL}$, to such potential detections in order to reject or keep them as final detections.

There are some details worth to mention. First, the usual non-maximum suppression (NMS) step included in pedestrian detectors is not performed for the output of the first stage, but it is done for the output of the second stage. Second, for ensuring that true pedestrians reach the second stage, we apply a threshold on $\mathcal{C}_B$ such that it guarantees a very high detection rate even having a very high rate of false positives. In our experiments this usually implies that while the $\mathcal{C}_B$ processes hundred of thousands windows (for pyramidal sliding window), $\mathcal{C}_{SSL}$ only process a few thousands. Third, although in Fig. 3.3 we show pyramids of images for a temporal neighborhood of T frames, what we actually keep from frame to frame are the already computed features, so that we compute them only once. However, this depends on the type of temporal neighborhood we use (Sect. 3.4.1). For instance, using projection style no feature re-computing is required (*i.e.*, keeping the scores would be sufficient). However, if we use either optical flow style, we may need to compute features in previous frames if the window under consideration does not map to a location where they were already computed.

3.5 Experimental results

Protocol As evaluation methodology we follow the de-facto Caltech standard for pedestrian detection [24], *i.e.* we plot curves of false positives per image (FPPI) *vs* miss rate. The miss rate average in the range of 10^{-2} to 10^{0} FPPI is taken as indicative of each detector accuracy, *i.e.*. the lower the better. Moreover, during testing we consider three different subset: *Near* subset includes pedestrians with height equal or higher than 75 pixels, *medium* subset includes pedestrian between 50 and 75 pixel height. Finally we group the two previous subset in the *reasonable* subset (height >= 50 pixels).

CVC08 On-board Sequence (CVC08) Since the temporal axis is important for the SSL classifier, we acquired our own dataset to be sure we have stable 30 fps sequences. The sequences were acquired on-board under normal urban driving conditions. The images are monochrome and of 480×960 pixels. We used a 4mm focal length lens, so providing a wide field of view. We drove during 30 minutes approximately, giving rise to a sequence of around 60,000 frames. Then, using steps



(a)

(b)

Figure 3.4: (a) Four images samples of our new dataset (CVC08). (b) Six pedestrian crops samples extracted from CVC08 dataset.

of 10 frames we annotated all the pedestrians[1]. This turns out in 7,900 annotated pedestrians, 5,400 reasonable and non occluded (See images and pedestrian crops examples in figure 3.4). We have divided the video sequence into three sequential parts, the first one for training (3,600 pedestrians), the last one for testing (1,300 pedestrians), in the middle we have leaved a gap for avoiding testing and training with the same persons.

Other publicly available datasets We have also used other three popular datasets acquired on-board. The Caltech dataset [24], which contain 3,700 reasonable pedestrians for training. The KITTI object detection dataset [33], which contains 7,481 training images, we split it in two sets (3,740 images for training and testing each) due to the absence of annotations in the original testing images. Finally the 15 sequences of the CVC02 dataset [37]. In this case, we took 10 first sequences for training and five last ones for testing. Overall, there are 5,090 mandatory pedestrians

[1]Publicly available in: http://www.cvc.uab.es/adas/site/?q=node/7

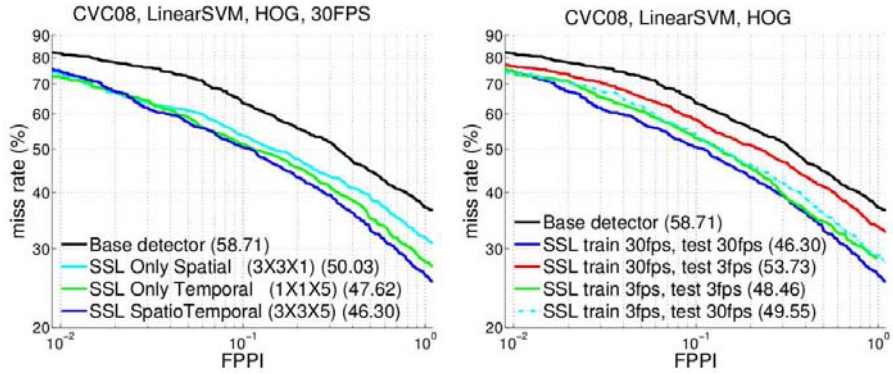


Figure 3.5: Left - Results using different neighborhoods. ($\Delta x \times \Delta y \times \Delta f$) stands for the spatial (Δx, Δy) displacements in HOG/LBP cell units and the temporal (Δf) window in frames (past window style). The projection approach is assumed here. Right - Results using SSL experiments training and testing under different frame rates.

for training and 2,900 for testing. It is worth to mention that, Caltech and KITTI images were acquired at 25 fps and CVC02 at 10 fps.

Base detectors For the experiments presented in this section we use our own implementation of HOG and LBP features [85], using TV-L1 [98] for computing optical flow, we obtain HOF features [87] as well. We call Base to the HOG+LBP/Linear-SVM and Base+HOF to the HOG+LBP+HOF/Linear-SVM.

Experiment 1 In Fig. 3.5 we show the results corresponding to only using spatial neighbors, only temporal neighbors, and both. Note how in all cases there is a large accuracy improvement, of even 12 perceptual points with respect to the base classifier for the spatiotemporal case. The rest of experiments will be based on the spatiotemporal SSL (with past temporal window style) and settings $(\Delta x, \Delta y, \Delta f) = (3, 3, 5)$. Also in Fig. 3.5 we observe that the SSL descriptor trained at 3 fps keeps its accuracy when it is tested on 30 fps, while the opposite is not true.

Experiment 2 In table 3.1 we show results for the spatiotemporal projection approach as well as for the spatiotemporal one based on optical flow computation over different subsets. Again, we observe large accuracy improvements for all the tested frame rates (30 fps, 10 fps, 3 fps) in CVC08 dataset and for the different evaluated datasets (Caltech, CVC02 and KITTI datasets). However, no significant difference is observed between the projection and optical flow cases. Also in figures 3.6, and 3.7, confirms the SSL improvement for all datasets, over different testing subsets. In this case, a relevant improvement is observed for so-called *near* pedestrians ($>$ 75 pixels

of height). Figure 3.8 shows some qualitative results from CVC08 for the projection case.

Discussion SSL approach outperforms its baseline in almost all the tested configurations. However, the improvement is more clear for near pedestrians at high frame rates. If we generate the *past neighborhood* over the far away pedestrians, we should expect a *past neighborhood* with pedestrians smaller than the minimum pedestrian size that the base detector can detect. That is why the SSL improvement is not so clear for the medium subset. However, in near pedestrians *past neighborhood* is more probable to find a history of confident responses. This is a very relevant improvement since for close pedestrians the detection system has less time to take decisions like braking or doing any other manoeuvre. Regarding the neighborhood generation approaches, the optical flow slightly improves the projection.

3.6 Conclusion

In this chapter we have presented a new method for improving pedestrian detection based on spatiotemporal SSL. We have shown how even simple projection windows can boost the detection accuracy in different datasets acquired on-board. We have shown that our approach is effective for different frame rates and using different pedestrian base classifiers as: HOG+LBP/Linear-SVM and HOG+LBP+HOF/Linear-SVM. Thus, looking at the promising obtained results we propose as future work to focus on testing the same approach for other base classifiers of the pedestrian detection state-of-the-art. Also, regarding the improvement obtained using optical flow neighborhood, we propose to further explore different approaches for dealing with the neighborhood generation for moving pedestrians, for instance the application of a affine transform based on optical flow that adapt not only the spatial position but also the size of the neighboring window in the temporal axis. Also weighted approaches that capture better the object movements ignoring the scene egomotion.

Table 3.1: Evaluation of SSL over different datasets, frame rates and pedestrian sizes. For FPPI $\in [0.01, 1]$, the miss rate average % is indicated.

Dataset	FPS	Experiment	Near	Medium	Reasonable
CVC08	N/A	Base: HOG+LBP	39.71	50.83	45.91
	3	SSL(Base) Proj. - OptFl.	**36.03** - 36.72	**50.01 - 50.04**	44.40 - **44.02**
		Base+HOF	47.98	56.65	50.88
		SSL(Base+HOF) Proj.	**37.62**	**52.21**	**45.47**
	10	SSL(Base) Proj. - OptFl.	35.49 - **34.79**	50.22 - **49.42**	43.56 - **42.10**
		Base+HOF	39.24	52.37	42.43
		SSL(Base+HOF) Proj.	**29.42**	**44.62**	**37.13**
	30	SSL(Base) Proj. - OptFl.	34.18 - **34.01**	49.84 - **48.04**	42.90 - **41.73**
		Base+HOF	37.81	53.39	38.78
		SSL(Base+HOF) Proj.	**27.37**	**46.53**	**35.85**
Caltech	25	Base	45.4	82.3	59.4
		SSL(Base) Proj. - OptFl.	40.6 - **38.9**	81.2 - **80.4**	59.4 - **57.6**
		Base+HOF	33.8	78.4	52.9
		SSL(Base+HOF) Proj.	**32.0**	**77.1**	**51.6**
CVC02	10	Base	30.19	51.86	40.18
		SSL(Base) Proj. - OptFl.	**25.43 - 25.31**	**48.97 - 48.63**	**35.06** - 35.54
KITTI	25	Base Pedestrian	27.02	51.92	35.99
		SSL(Base Pedestrian) Proj. - OptFl.	**20.75** - 21.39	49.27 - **46.02**	**30.50 - 30.62**
		Base Cyclist	20.28	42.22	31.40
		SSL(Base Cyclist) Proj.	**17.60**	**35.01**	**27.27**

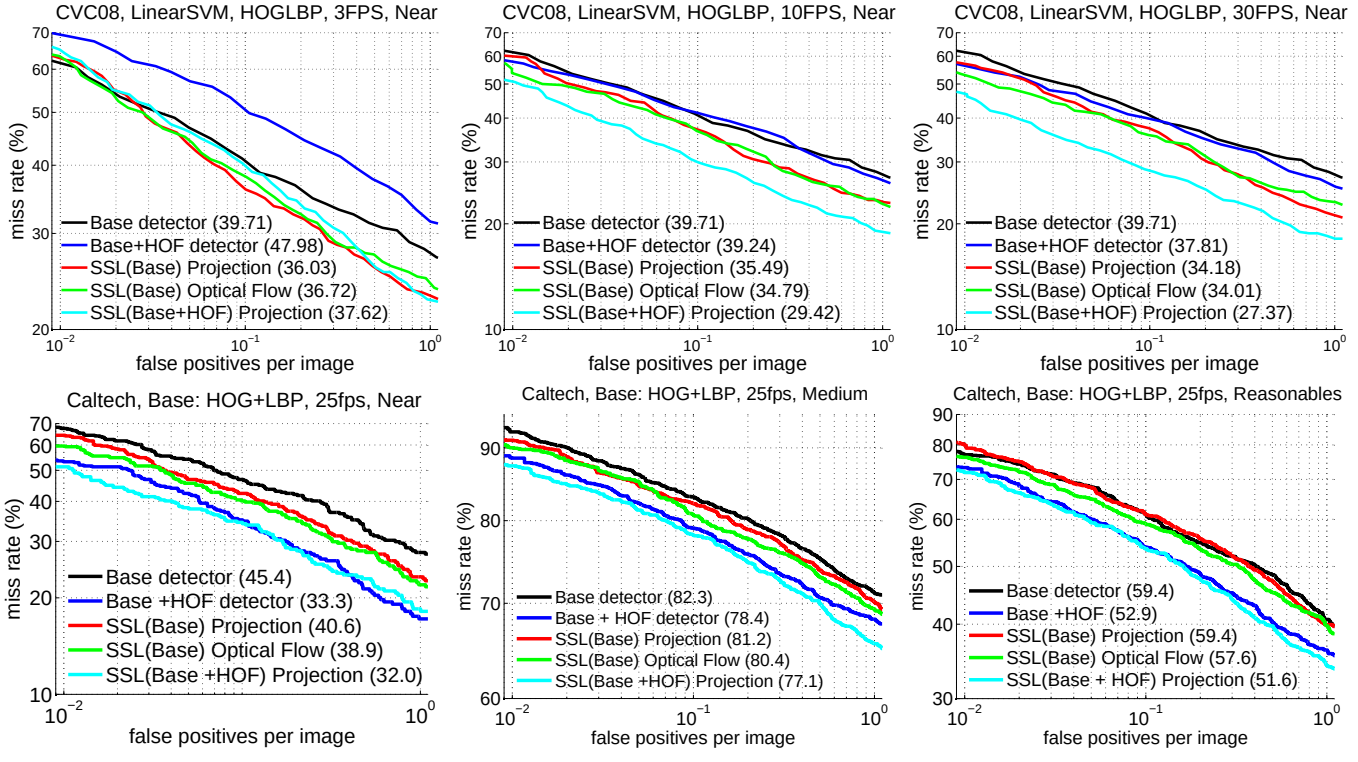


Figure 3.6: Results for CVC08, and Caltech datasets. At the top row there are the 30fps, 10fps and 3fps cases of CVC08 using the *near* testing subset. The last two cases are obtained by sub-sampling the video sequence, but always keeping the same training and testing pedestrians. At the bottom row there are the experiments over the *near*, *medium* and *reasonable* testing of Caltech.

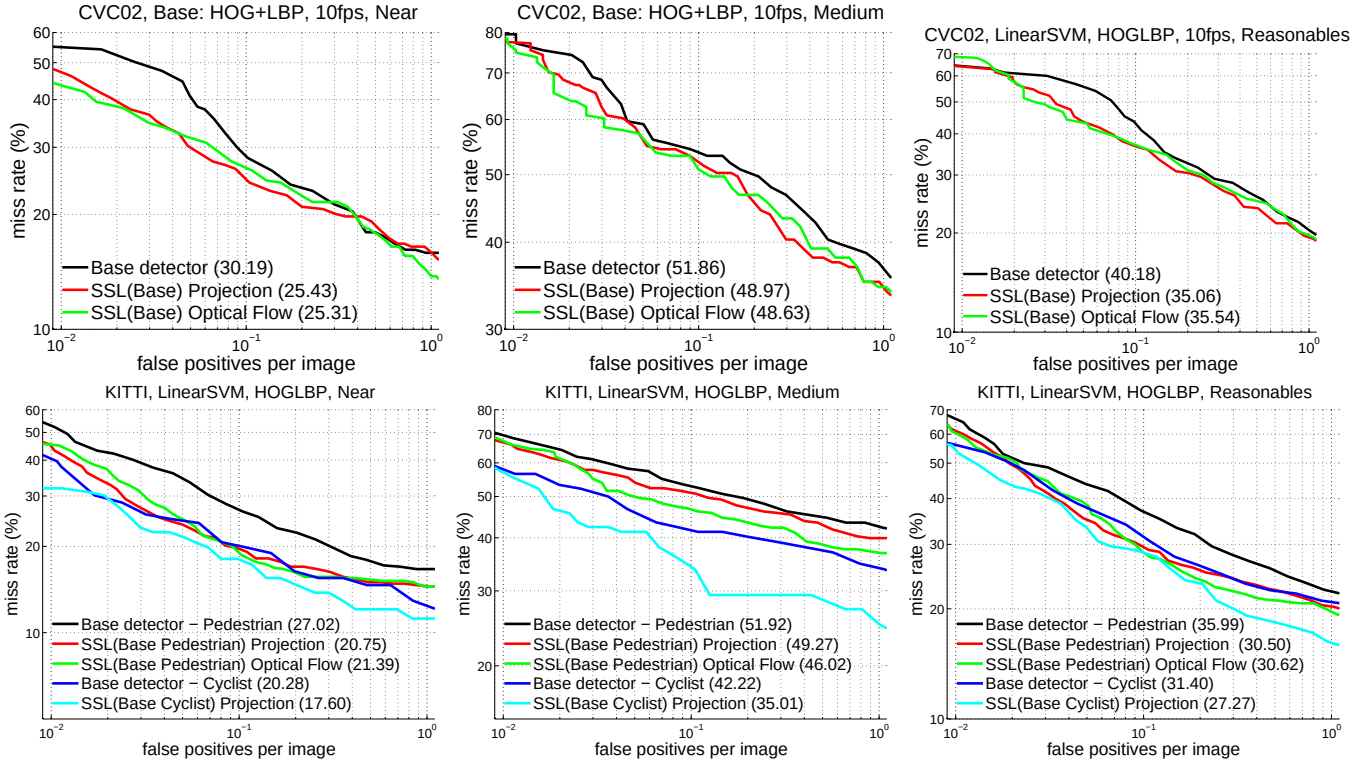


Figure 3.7: Results for CVC02 and KITTI datasets. At the top row there are experiments over CVC02 dataset. At the bottom there are experiments over KITTI dataset. Both rows contain experiments performed over *near*, *medium* and *reasonable* subsets.

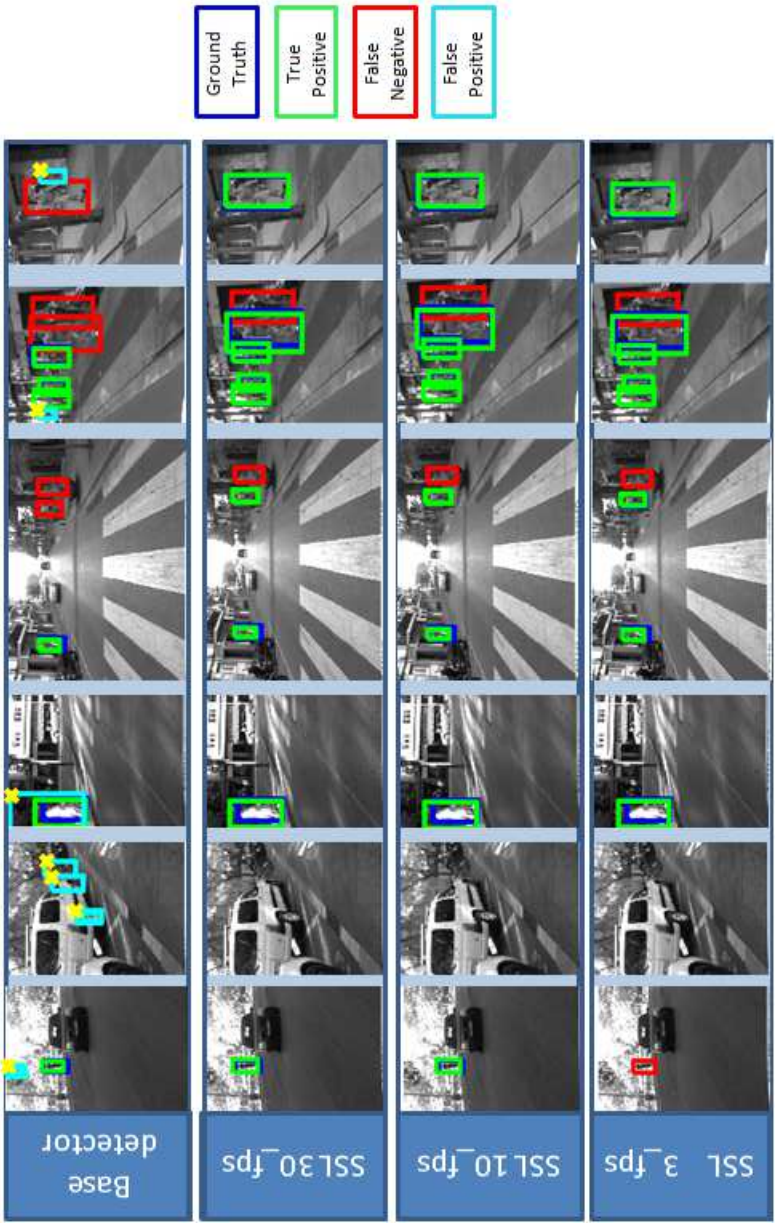


Figure 3.8: Qualitative results from the CVC08 dataset comparing the base classifier and the SSL for 3, 10 and 30 fps. The first three columns focus on improvements regarding false positives rejection, while the rest focus on examples where SSL avoids missing pedestrians. The non-detected pedestrians with the SSL approach (last two columns) correspond to occluded pedestrians.

Chapter 4

Multi-view, Multi-modal Random Forest of Local Experts

Despite recent significant advances, object detection continues to be an extremely challenging problem in real scenarios. In order to develop a detector that successfully operates under these conditions, it becomes critical to leverage upon multiple cues, multiple imaging modalities and a strong multi-view classifier that accounts for different object views and poses. In this chapter we provide an extensive evaluation that gives insight into how each of these aspects (multi-cue, multi-modality and strong multi-view classifier) affect accuracy both individually and when integrated together. In the multi-modality component we explore the fusion of RGB images with depth maps obtained by high-definition LIDAR, and by a stereo-pair reconstruction. In the multy-view component we extend the evaluation to other objects relevant to autonomous vehicles: cyclists, and cars. As our analysis reveals, although all the aforementioned aspects significantly help in improving the accuracy, the fusion of visible spectrum and depth information allows to boost the accuracy by a much larger margin. The resulting detector not only ranks among the top best performers in the challenging KITTI benchmark, but it is built upon very simple blocks that are easy to implement and computationally efficient.

4.1 Introduction

Developing a reliable object detector enables a vast range of applications such as video surveillance and the practical deployment of autonomous and semi-autonomous vehicles. In order to obtain a detector that successfully operates under realistic conditions, it becomes critical to exploit sources of information along different orthogonal axis like: i) the integration of multiple feature cues (contours, texture, etc.), ii) the fusion of multiple image modalities (color, depth, etc.), and iii) the use of multiple views (frontal, lateral, etc.) of the object by learning a strong classifier that accommodates

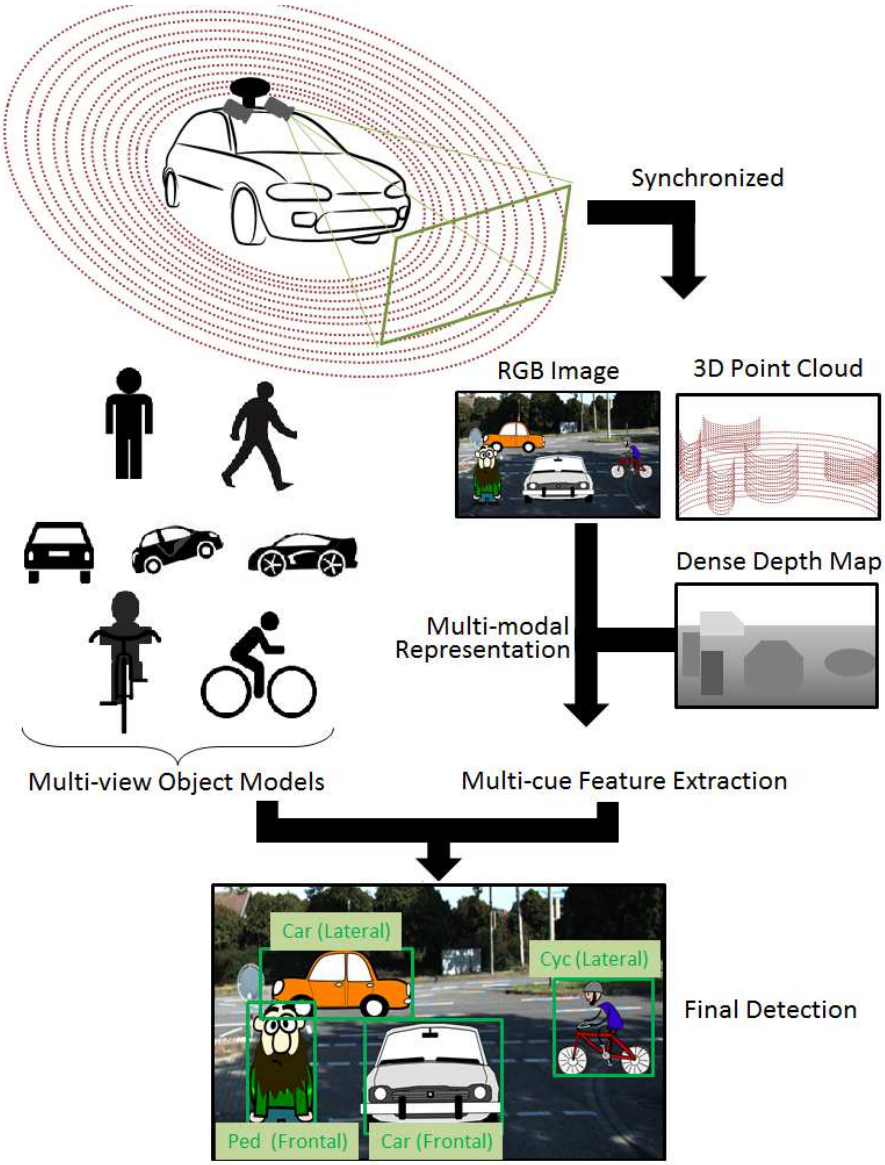


Figure 4.1: General scheme: From RGB images and LIDAR data to object detection. RGB images and LIDAR data synchronized for multi-modal representation. Multi-modal representation based on RGB images and dense depth maps. Multi-cue feature extraction over the multi-modal representation. Multi-view detection of different objects.

for both different 3D points of view and multiple flexible articulations (See general scheme in Fig. 4.1). These three axis allow us to increase robustness of detectors. This increase is due to: i) capture complementary features that increase description of objects; ii) include redundant information that compensate bad performance of one modality with information from the other one; and iii) reducing intra-class variability by spliting the pedestrian detection problem in $n - views$ sub-problems.

In order to integrate different cues we use HOG [17], that provides a good description of the object contours, and LBP [2] as texture-based feature. These two types of features provide complementary information and the fusion of both types of features has been seen to boost the performance of either feature separately [26, 58, 89]. Both types of features are extracted for the different image modalities. We show that by appropriately choosing the parameters used in the computation of these features for each modality we can obtain an important gain in accuracy.

In order to integrate multiple image modalities, we considered the fusion of depth maps with visible spectrum images (Multi-modal scheme specification in Fig. 4.2). The use of depth information has gained attention thanks to the appearance of cheap sensors such as the one in Kinect, which provides a dense depth map registered with an RGB image (RGB-D). However, the sensor of Kinect has a maximum range of approximately 4 meters and is not very reliable in outdoor scenes, thus having limited applicability for objects detection in on-board sequences. On the other hand, Light Detection and Ranging (LIDAR) sensors such as the Velodyne HDL-64E have a maximum range of up to 50 meters and are appropriate for outdoor scenarios. Although they produce a sparse cloud of points and they are only recently starting to receive attention for application to object detection. Also depth information coming from 3D stereo reconstruction has received low attention for pedestrian detection. Even it produce a dense depth map, author usually do not pay attention in this information source. In this Chapter we explore the fusion of dense depth maps (obtained based on the sparse cloud of points or by 3D stereo reconstruction) with RGB images. Following [39], the information provided by each modality can be fused using either an early-fusion scheme, *i.e.* at the feature level, or a late-fusion scheme, *i.e.* at the decision level.

Learning a model flexible enough for dealing with multiple views and multiple positions of an articulated object is a hard task for a holistic classifier. In order to fulfill this aspect we make use of Random Forests (RF) of local experts [59], which has a similar expressive power than the popular Deformable Part Models (DPM) [28] and less computational complexity. In this method, each tree of the forest provides a different configuration of local experts, where each local expert takes the role of a part model. At learning time, each tree learns one of the characteristic configurations of local patches, thus accommodating for different flexible articulations occurring in the training set. In [59] the RF approach consistently outperformed DPM. An advantage of the RF method is that only a single descriptor needs to be extracted for the whole window, and each local expert re-uses the part of the descriptor that corresponds to the spatial region assigned to it. Its computational cost is further significantly reduced by applying a soft cascade, operating in close to real time. Contrary to the

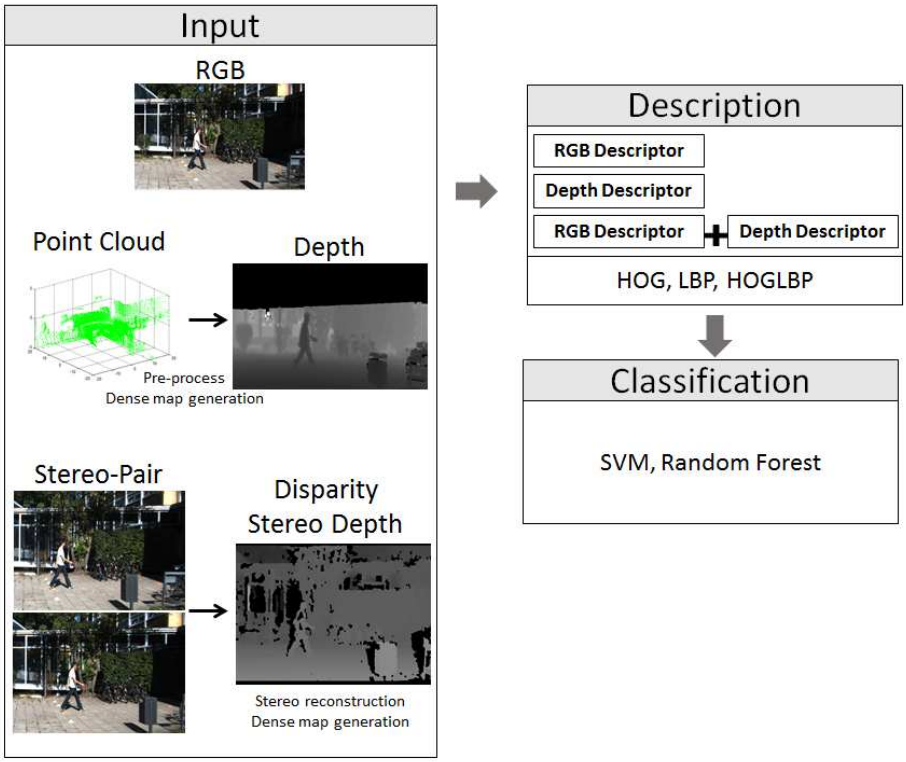


Figure 4.2: Multi-cue, Multi-modal detector scheme. 1) Generate a multi-modal representation using RGB and depth. 3) Extract multi-cue features. 4) Train multi-modal, multi-cue classifier.

DPM, the original RF method learns a single model, thus not considering different viewpoints separately. In this work, we extend this method to learn multiple models, one for each 3D pose, and evaluate both the original single model approach and the multi model approach. Several authors have proposed methods for combining local detectors [28, 93] and multiple local patches [31, 52, 81]. The method in [97] also makes use of RF with local classifiers at the node level, although it requires to extract many complex region-based descriptors, making it computationally more demanding than [59].

In this chapter we perform an extensive evaluation providing insights about how each of these three aspects affect accuracy, both individually and when integrated together. The proposed method (General scheme in Fig. 4.1) will be evaluated in well-known KITTI pedestrian dataset pedestrians, under different base classifiers. As our analysis reveals, the fusion of visible spectrum and depth information allows to boost the accuracy by a much larger margin.

The rest of the chapter is organized as follows.In Sect. 4.2 we present the related

work to our proposal. In Sect. 4.3 we develop our proposal. Section 4.4 presents the experiments carried out to assess our proposal step by step, and discuss the obtained results. Finally, section 4.5 draws our main conclusions.

4.2 Related Work

From the seminal work of Dalal and Triggs [17] it has been seen that using different types of gradient-based features and their spatial distribution, such as in the HOG descriptor [17] provides a distinctive representation of both humans and other objects classes. However, there exist in the literature other approaches such the integral channel features proposed by Dollar et al. [22] that allows to integrate multiple kinds of low-level features such as the gradient orientation over the intensity and LUV images, extracted from a large number of local windows of different sizes and at multiple positions, allowing for a flexible representation of the spatial distribution. In [8], [75] it has been seen that including color boosts the performance significantly, being this type of feature complementary to the ones we used in this study. Context features have also been seen to aid [88], [11] and could be easily integrated as well. Exploring alternative types of spatial pooling of the local features is also beneficial as shown in [90] and is also complementary to the approach used in this paper.

Object detection based on data coming from multiple modalities has been a relatively active topic of study [36], and in particular the use of 2D laser scanners and visible spectrum images has been studied in several works, for instance [72, 79]. Only recently authors are starting to study the impact of high-definition 3D LIDAR [6, 46, 47, 61, 72, 79, 95]. Most of these works propose specific descriptors for extracting information directly from the 3D cloud of points [6, 46, 47, 61, 79, 95]. A common approach is to detect objects independently in the 3D cloud of points and in the visible spectrum images, and then combining the detections using an appropriate strategy [46,47,95]. Following the steps of [72], dense depth maps are obtained by first registering the 3D cloud of points captured by a Velodyne sensor with the RGB image captured with the camera, and then interpolating the resulting sparse set of pixels to obtain a dense map where each pixel has an associated depth value. Given this map, 2D descriptors in the literature can be extracted in order to obtain a highly distinctive object representation. Our work differ from [72] in that we use multiple descriptors and adapt them to have a good performance in dense depth images. While [72] employs a late fusion scheme, in our experimental analysis we evaluate both early and late fusion approaches in the given multi-cue, multi-modality framework.

Most relevant to our approach is the presented in [26] where the authors combine multiple views (front, left, back, right), modalities (luminance, depth based on stereo, and optical flow), and features (HOG and LBP). The main differences between [26] and our work are as follows: i) in order to complement RGB information, we make use of a sensor modality, high-definition 3D LIDAR, which has received relatively little attention in pedestrian detection until now, but it is being used for autonomous driving, and ii) while [26] makes use of an holistic classifier, we make use of a more

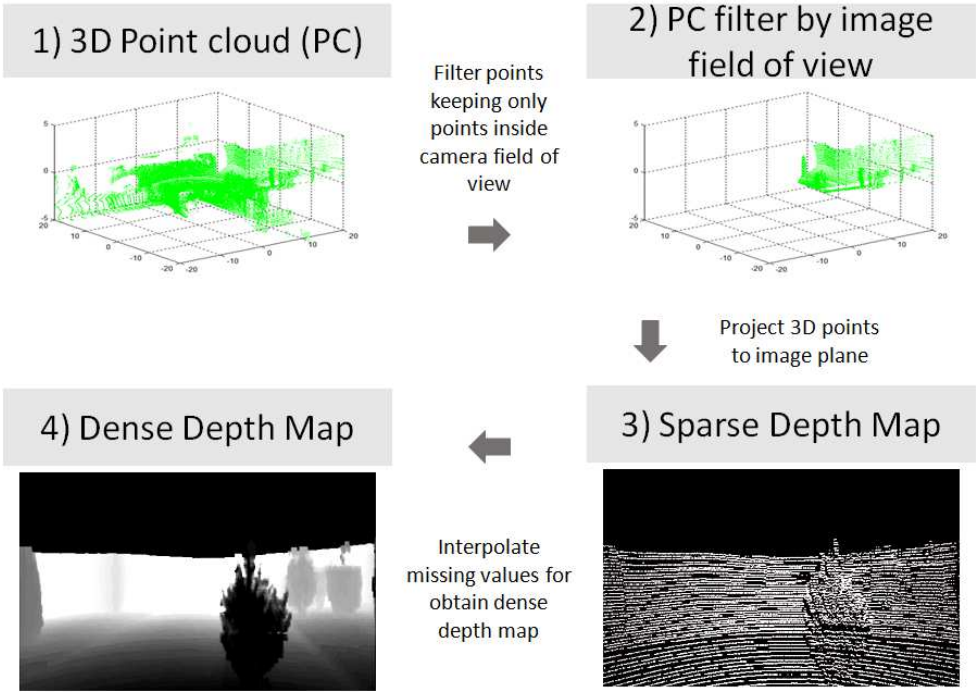


Figure 4.3: Dense depth map generation scheme. From a cloud of points to a dense depth map: Filter cloud of points for synchronize with view field of image. Projection of 3D points into 2D image coordinates. Interpolate depths for getting a dense depth map.

expressive patch-based model, and iii) in [26] multiples cues are combined following late-fusion style, while we consider also early-fusion, which, in fact, gives better results in our framework.

Our analysis reveals that, although all the aforementioned components (the use of multiple feature cues, multiple modalities and a strong multi-view classifier) are important, the fusion of visible spectrum and depth information allows to boost the accuracy significantly by a large margin in pedestrian class, but the multi-view axis in the one with high impact in cyclists, and cars. The resulting detector not only ranks among the top best performers in the challenging KITTI benchmark, but it is built upon very simple blocks that are easy to implement and computationally efficient.

4.3 Multi-modal Detector for Pedestrian Detection

We propose a complete framework in which our final model incorporates the multi-cue characteristic by extracting HOG and LBP descriptors. Also the multi-modal characteristic by extracting information from RGB and depth modalities, which will be combined at feature level (early fusion) or at decision level (late fusion). In addition we will model objects both holistically and as a set of relevant patches. In the former case the model will be learnt with linear SVM; and in the latter with a Random Forest of Local Experts [59].

4.3.1 Multi-cue feature representation

In order to improve the pedestrian detection accuracy it is widely used the incorporation of different cues or features. This incorporation looks for complementarity by using different cues for describing the same object. In order to incorporate different cues in our framework we use the HOG [17] descriptor (shape) and the LBP [62] descriptor (texture). Both descriptors are combined using an early fusion technique, concatenating them, obtaining a robust descriptor with complementary information (HOGLBP). HOG descriptor is composed by a histogram of gradient orientations. Given a candidate window the histograms are calculated on overlapped blocks inside it. LBP descriptor calculates histograms of texture patterns over the same overlapped blocks than HOG. This texture patterns are based on value differences between the central pixel and the surrounding ones in a 3×3 neighborhood. We use our own implementation that includes some modifications that improve the final detection rate. The first modification is included in the image pyramid construction. The image re-size process is done by bilinear interpolation with antialiasing, which helps the gradient calculation and thereby the HOG descriptor classification accuracy. The second modification is included in the LBP descriptor. When the value differences are calculated we accept as equal values the ones included in a defined range, this range (defined as $ClipTh$) allows that small noises (small value changes) do not affect the texture pattern.

4.3.2 Multi-modal image fusion

Keeping in mind that more complementarity is better for pedestrian detection, we want to explore the integration of different modalities. Usually information is extracted from a single-modal sensor (RGB camera), but we combine this visual information with Depth information. This depth information can be acquired by to different sources.

First depth source is based on 3D information extracted from a LIDAR sensor. Lidar sensor provide a sparse point cloud. In order to transform the point cloud obtained using the LIDAR into a dense depth map, we follow the approach presented by Premediba *et al.* [72]. In this method, the 360° 3D point cloud from the LIDAR sensor is filtered in order to take only those points included in the viewfield of the RGB

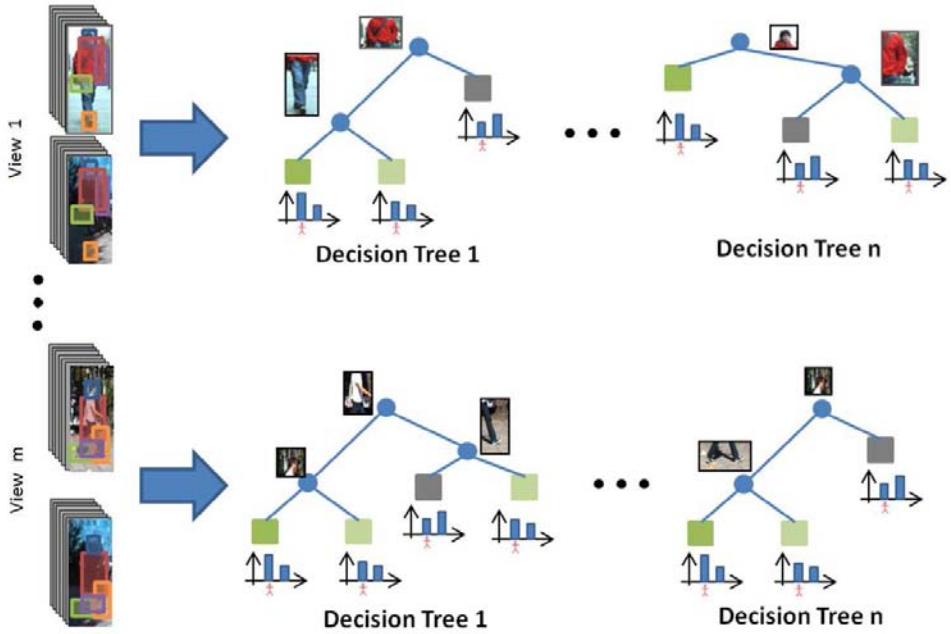


Figure 4.4: Multi- View Random Forest scheme. For each view is learnt a different random forest, and each tree has different configuration of random patches.

camera. In order to do this each point P_i is projected into the image plane using the calibration and projection matrices provided in the dataset, using $TM = P2 \times R0 \times VtC$, where, $P2$ is the projection matrix from camera coordinate system to left color image coordinate system, $R0$ is the rectification matrix, and VtC is the projection matrix from velodyne coordinate system to camera coordinate system. Once we have the transformation matrix (TM) we can project any 3D point (defined by its 3D coordinates $[x_{3D}, y_{3D}, z_{3D}]$) to its correspondent point in the image plane (defined by its 2D coordinates $[x_{2D}, y_{2D}]$) by applying $[x_{2D}, y_{2D}, 1] = TM * [x_{3D}, y_{3D}, z_{3D}, 1]$. Then the points that fall inside the image borders are selected, while the others are rejected, ending up with points that form a sparse depth image, time and space synchronized with the visual image. At this step by defining a neighborhood (N) for each valid pixel of the depth map we interpolate the information for filling the missing values. In order to calculate the missing values we use the bilateral filtering formalism [69]: $D_p = \frac{1}{W_p} * \sum_{q \epsilon N} G_d(\|p - q\|) * G_i(|I_q|) * I_q$ where I_q is the depth value of the point q, G_d weights points q inversely to their distance to position p, G_i penalizes as function of their range values, and W_p is a normalization factor. After this process, the pixels without depth information will be filled, ending up in a dense depth map (see Fig. 4.3).

Second depth information is based on 3D stereo reconstruction. The input of

this system is a stereo-pair calibrated camera. This calibration provides for each camera the intrinsic (internal parameters such lens distortion and focal length) and extrinsic (external parameters such relative position and orientation) parameters. The calibration process needs point correspondences between the two images to compute the calibration matrix. For stereo processing both cameras should have exactly the same internal parameters (sensor size, focal length, lens distortion) and be totally parallel (no rotation, and same position except for the baseline distance). In practice this is is very difficult to achieve due to camera manufacturing errors and mounting imprecision. In order to make a correct 3D reconstruction first a rectification of the images must be done. This rectification transforms both images as if they would be acquired from two ideal cameras (same internal parameters, totally parallels and with no lens distortion). After the rectification process matching points from one image to the other lie at the same horizontal line. This fact reduces the time of the disparity computation. The disparity image is a gray level image that indicate for each pixel the distance (in pixels) to its corresponding pixel in the other image (both pixels represent the same 3D world point). From disparity images, it can be easily computed the depth map using the following formula: $Depth = (b * f)/Disparity$, where b is the baseline and f the focal length.

4.3.3 Multi-view classifier

In general, reducing intra-class variability is a good way to better discriminate a class from potential false positives (background). One of the biggest causes of the large variability in object detection, is the pose and orientation of the object. In order to solve this problem we propose to use a multi-view approach (Fig. 4.5). Given a set of annotated pedestrians for training a detector, we propose to separate them into n different views depending on its orientation and aspect ratio. For this goal we cluster the training set samples using regular-spaced seeds in orientation using K-Means algorithm [56]. By spliting in this way the samples we can adjust the canonical size of the detection window for each subset, selecting the mean size of the samples in the partition set. Thus, allowing the final detector to deal with objects in different orientation having each orientation its own aspect ratio (*e.g.* it is not the same bounding box for a frontal-viewed pedestrian than for a side-viewed pedestrian). In figure 4.6, 4.7, and 4.8 it is shown the training samples and their views definition based on the clustering process for the pedestrians, cyclists, and cars classes respectively. In order to cluster we use the the orientation angle (α) and the aspect ratio (AR) of the sample.

4.3.4 Object model

In our study we focus on two different models: one holistic, where the whole object view is considered as the model; and a patch-based one where only a subset of object patches are considered as model. As holistic model we use the $descriptor/SVM$ with a linear kernel ($linSVM$) which has a good compromise between computation

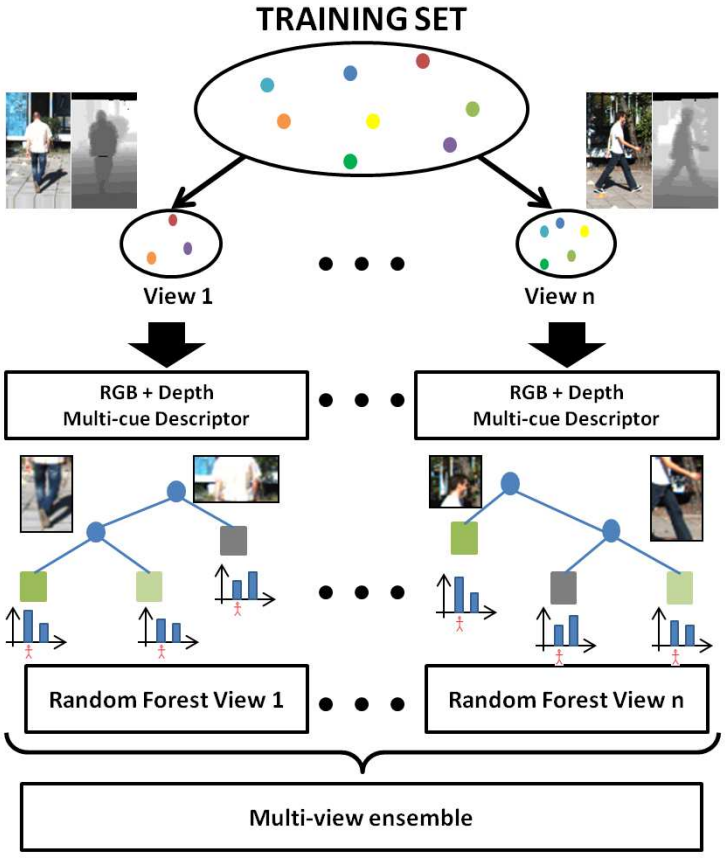


Figure 4.5: Multi-view, Multi-cue, Multi-modal detector scheme. 1) Split training set samples in different views. 2) Generate a multi-modal representation using RGB and depth. 3) Extract multi-cue features. 4) Train a random forest of local experts for each view. 5) Ensemble different views detection.

time and accuracy. As patch based model we use $descriptor/RandomForest(RF)$. Following the implementation in [59], each node in the tree learns a classifier based on a random patch inside the candidate window, obtaining a RF in which each tree has different configuration of patches (see Fig. 4.4), and the classification decision is made by taking into account the configuration learned in each tree of the forest. We will use the RF formed by 100 trees, 7 levels as maximum depth and each node in a tree will be a $linSVM$ local expert (see [59] for details).

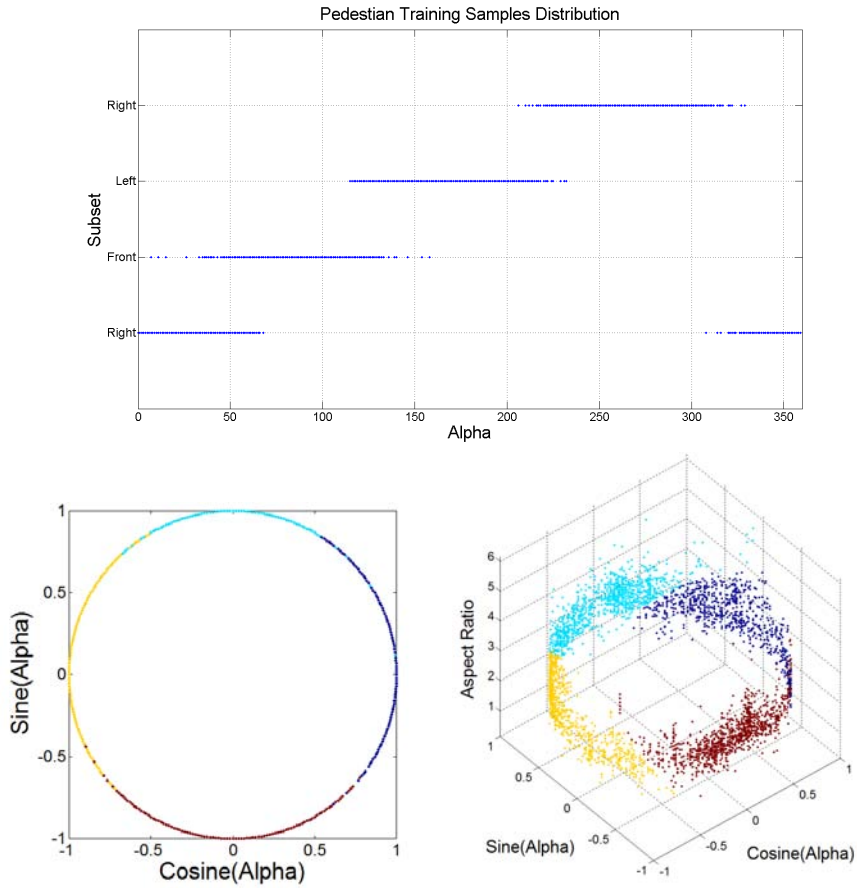


Figure 4.6: Pedestrian Orientation Histogram and Distribution. Upper image shows the assigned views against the angle for each sample. Down images show the samples distribution in the clustering space (angle and aspect ratio).

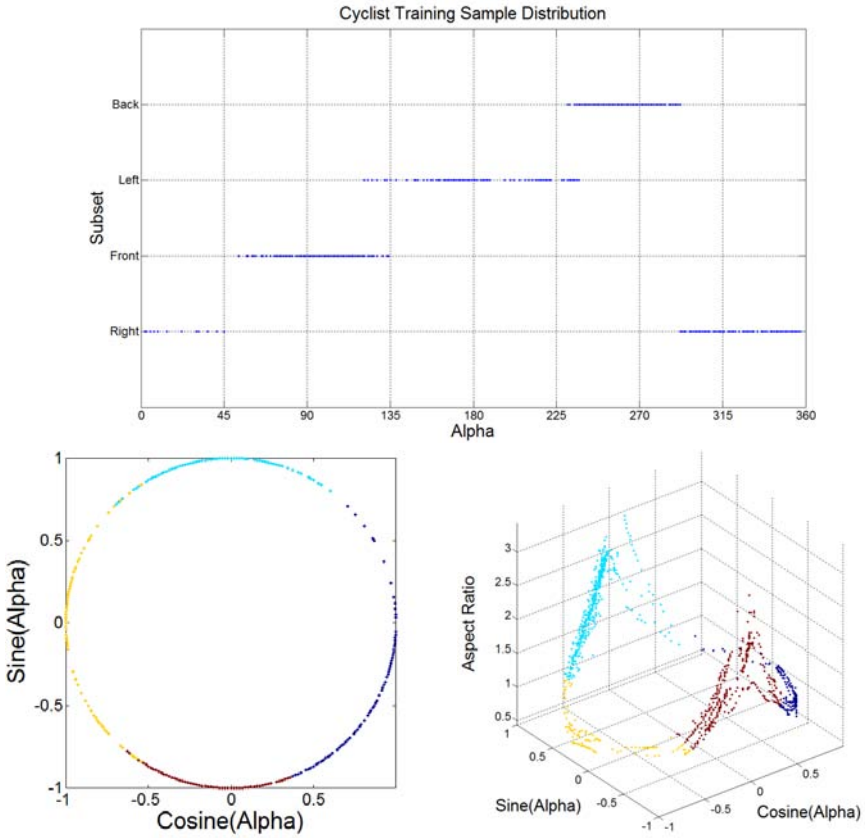


Figure 4.7: Cyclist Orientation Histogram and Distribution. Upper image shows the assigned views against the angle for each sample. Down images show the samples distribution in the clustering space (angle and aspect ratio).

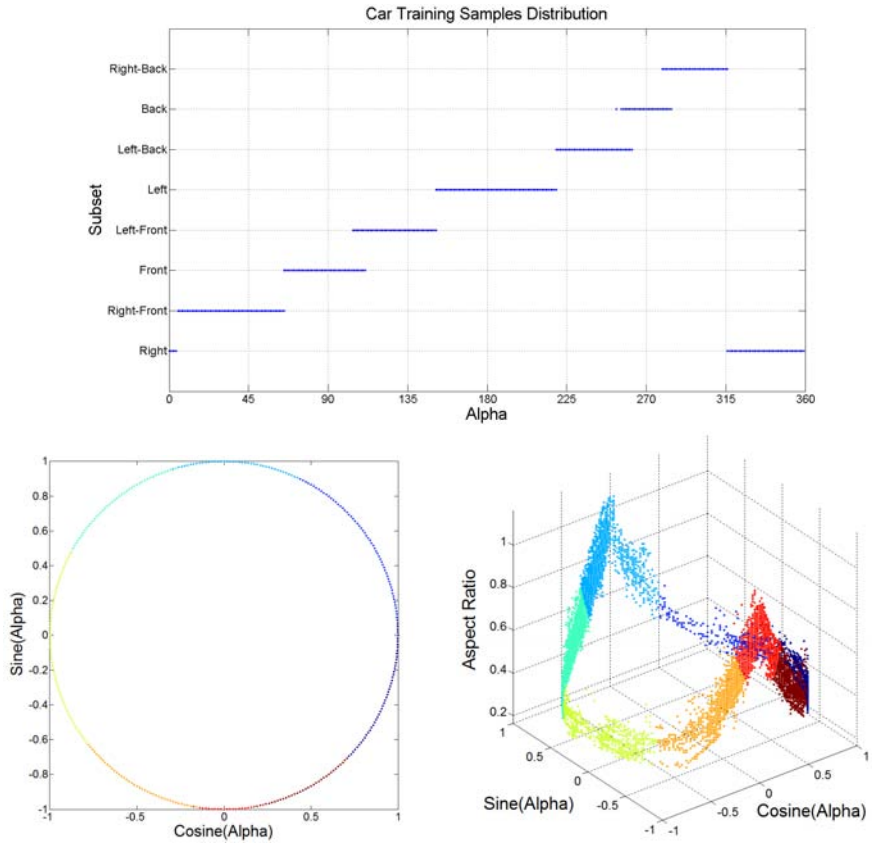


Figure 4.8: Car Orientation Histogram and Distribution. Upper image shows the assigned views against the angle for each sample. Down images show the samples distribution in the clustering space (angle and aspect ratio).

Table 4.1: Results for PEDESTRIAN detection using different modalities, and detectors; tested over the validation set. For each detector AMR for caltech evaluation protocol is shown . Best AMR for each detector across the different modalities is indicated in bold

Evaluation	Detector	RGB	Disp.	Depth	RGB/Disp.	RGB/Depth
AMR (Reasonable)	HOG/SVM	50.29	66.30	42.17	44.26	**40.04**
	LBP/SVM	48.91	72.72	41.42	42.37	**38.59**
	HOGLBP/SVM	38.66	58.63	38.05	**32.30**	33.15

4.4 Experimental results

In this section we will evaluate each step of the proposed approach: multi-cue, multi-modal, and multi-view as we have described in previous sections, in order to fulfill this evaluation we will use HOG and LBP features and as classifier the SVM with linear kernel, and the Random Forest of local experts. Letting us with a bunch of possibles detectors: HOG/linSVM, LBP/linSVM, HOGLBP/linSVM, HOGLBP/RF. We will use as baseline for comparing the different steps the HOG/linSVM detector which was the first milestone in pedestrian detection.

KITTI Dataset We use the KITTI dataset since it provides synchronized stereo-pair camera and LIDAR data. KITTI dataset for object detection includes 7,481 training images and 7,518 test images, comprising a total of 80,256 labeled objects. Annotations are provided only for the training set. For this reason we split the training set into a training set (the first 3,740 images) and a validation set (the last 3,741 images) as in [72], these subsets are used for the evaluation of each step of our approach. The original training and testing set will be used for training and testing the optimal configuration of the detector. During training we consider pedestrians higher than 25 pixels and not occluded (Reasonable subset).

Evaluation protocol As evaluation methodology we follow the de-facto Caltech standard for pedestrian detection [24], *i.e.*, we plot curves of false positives per image (FPPI) *vs* miss rate. The average miss rate (AMR) in the range of 10^{-2} to 10^{0} FPPI is taken as indicative of each detector accuracy, *i.e.*. the lower the better. Also we will evaluate using the KITTI evaluation framework in which the precision-recall curve is calculated for ranking the methods by the average precision (AP), *i.e.*, the higher the better. For testing we use the reasonable subset in the caltech evaluation and the KITTI evaluation is performed over three different subsets depending on height and occlusion level: *easy subset* (Min height: 40 px; max occlusion level: fully visible; max truncation: 15%), *moderate subset* (Min height: 25 px; max occlusion level: partly occluded; max truncation: 30%), *hard subset* (Min height: 25 px; max occlusion level: difficult to see; max truncation: 50%). This KITTI evaluation will be performed in the validation set.

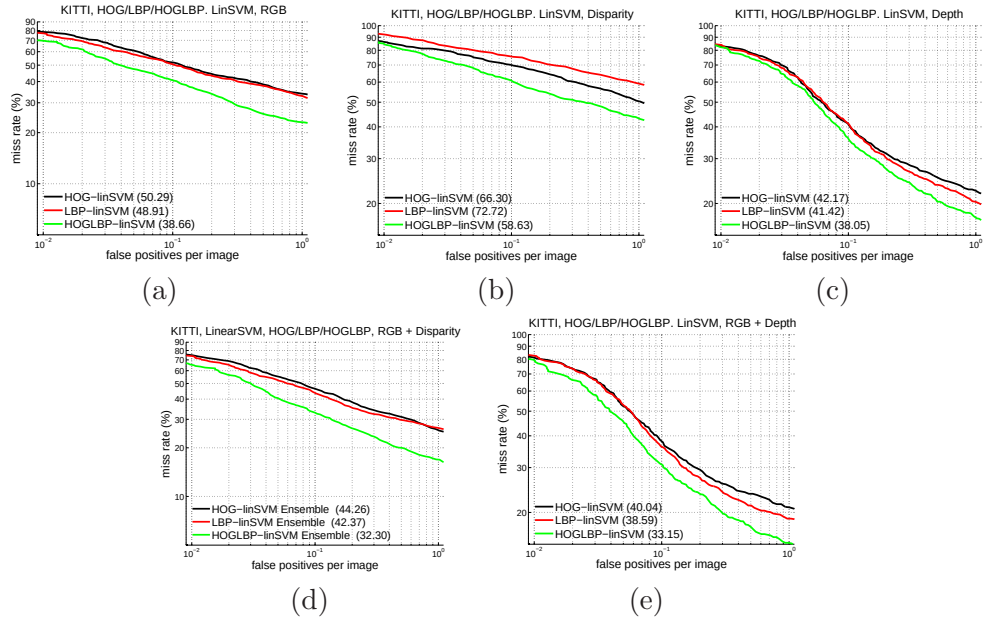


Figure 4.9: Results over Validation Set for HOG/LinSVM, LBP/LinSVM, and HOGLBP/LinSVM, using RGB disparity (stereo), and depth (LIDAR).Experiments using: (a) RGB modality, (b) Disparity (Stereo) modality, (c) Depth (LIDAR) modality, (d) multi-modal RGB+Disparity, (e) multi-modal RGB+Depth.

Multi-cue We start by evaluating the gain obtained by using multiple cues, for that reason we start by comparing detectors over different modalities isolated, comparing single-cues based detectors (HOG and LBP) against multi-cue based one (HOGLBP). However, first of all we have tuned the LBP parameters, getting $ClipTh_{RGB} = 4$, $ClipTh_{Disparity} = 6$, and $ClipTh_{Depth} = 0.2$. These parameters mean that, for calculating the texture pattern, we will treat as the same value those in the range on 4 luminance units for the RGB modality, 6 pixels in disparity, and 0.2 meters in depth. As it is usual for pedestrian detection to use the Caltech standar evaluation method, in table 4.1 we tabulate the AMR, and in Fig. 4.9 we plot the FPPI curves for different detectors. Comparing the AMR in HOG/linSVM detector against HOGLBP/linSVM one, we can see that the gain in AMR is around 12 points with RGB modalities, around 8 with disparity, and around 4 with depth. The same behavior can be seen also if we compare the LBP/linSVM detector against the HOGLBP/linSVM one where we obtain improvements of around 10 , 14 and 3 respectively.

Multi-modal Regarding the evaluation of the multi-modal approach, we compare the HOG/linSVM detector over RGB, Disparity, and Depth against its combinations

Table 4.2: Results for PEDESTRIAN detection using different modalities, and detectors, tested over the validation set. For each detector AP for KITTI evaluation is shown . Best AP for each detector across the different modalities in bold

Evaluation	Detector	RGB	Depth	Early Fusion	Late Fusion
AP (Easy)	HOG/SVM	0.50	0.59	**0.71**	0.63
	LBP/SVM	0.52	0.62	**0.69**	0.66
	HOGLBP/SVM	0.64	0.65	**0.74**	0.70
	HOGLBP/RF	0.73	0.74	**0.79**	0.77
AP (Moderate)	HOG/SVM	0.38	0.46	**0.57**	0.49
	LBP/SVM	0.41	0.48	**0.57**	0.52
	HOGLBP/SVM	0.50	0.51	**0.61**	0.56
	HOGLBP/RF	0.59	0.58	**0.65**	0.62
AP (Hard)	HOG/SVM	0.33	0.40	**0.50**	0.43
	LBP/SVM	0.36	0.42	**0.50**	0.45
	HOGLBP/SVM	0.43	0.45	**0.53**	0.49
	HOGLBP-RF	0.51	0.50	**0.56**	0.54

RGB+Depth, and RGB+Disparity. We start this multi-modal study by applying a naive late fusion technique, which merge the detections over each of the different modalities in a single multi-modal detection. In Fig. 4.1 are tabulated the AMR for the different detectors over the different modalities and their combination. We can see that the late fusion experiment outperform the single-modality one for all the proposed detectors. Regarding the RGB-based detectors against RGB+Disparity based ones we obtain a gain in AMR of around 16 point over HOG/SVM detector, 6 over LBP/SVM detector, and 6 over HOGLBP/SVM. Similar results are obtained when we compare RGB-based detectors aging the RGB+Depth-based ones; Disparity-based against RGB+Disparity-based ones; and Depth-based against RGB+Depth-based ones. Regarding the results obtained in these previous experiments, we can observe the better performance of detector based on depth information coming from a LIDAR sensor. Therefore, we decide that from now ahead we will use only depth information provided from a LIDAR, over depth information provided from 3D stereo reconstruction.

Multi-cue, Multi-modal Random Forest of Local Experts Concluding that multi-modal approaches outperform single-modal ones. we propose to extend the study to random forest of local experts (**RF**) [59]. In order to perform a deeper study, we exploit the implementation of an early fusion technique, which fuse the descriptors from both modalities in a single one. To this end we test the detectors used in previous experiments adding to this study the implementation of a multi-modal RF. In figure 4.10 are shown the FPPI curves obtained for the different configurations of detectors and multi-modal techniques. Regarding the results ovserved in figure 4.10 we conclude that early fusion technique outperform late fusion one in all the tested detectors. In table 4.2 we resume the results obtained over the different testing subset (*easy, moderate, and hard*) in KITTI evaluation protocol. Regarding *moderate* subset in

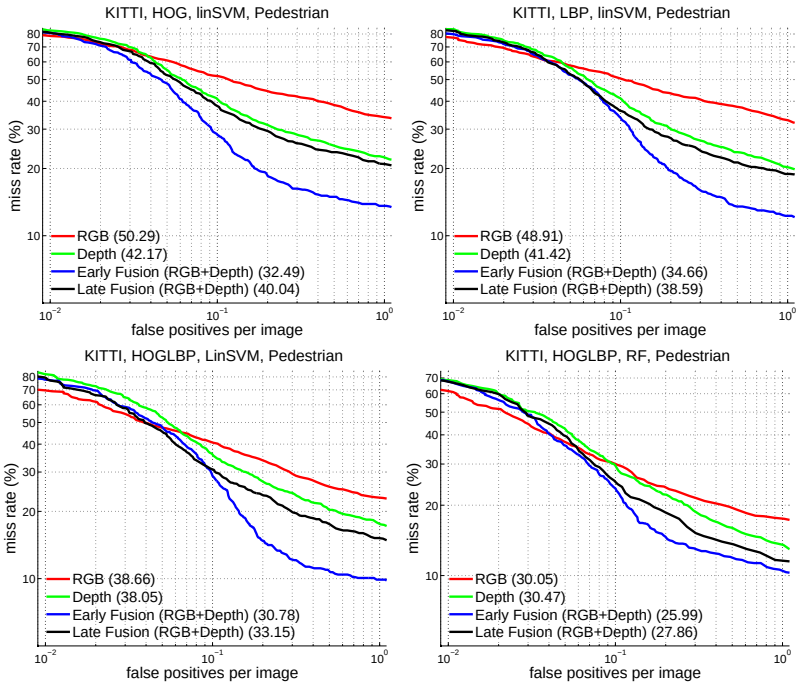


Figure 4.10: Results over validation set of detectors using early and late fusion. Descriptors: HOG/linSVM, LBP/linSVM, HOGLBP/linSVM, HOGLBP/RF under the different sources: RGB, Depth and RGB+Depth (late and early).

table 4.2 we observe that multi-modal (**MM**) approach using a early fusion techniche has a gain in AP against the single-modal (**SM**) detectors. MM-HOG/LinSVM has a gain of 19 points against SM-HOG/LinSVM(RGB), and a gain of 11 points against SM-HOG/LinSVM(Depth). Same behavior is observed for all detectors.

Multi-view detector In order to evaluate the multi-view (MV) axis improvements introduced in a single-view (SV) model, we will compare the SV-detector against the MV one. In table 4.3 are tabulated the different views relevant values, min and max angle, aspect ratio and number of samples. We define a two-view approach for pedestrians and cyclist grouping the left and right views (Lateral-view) and the front and back views (Frontal-view), for cars we define eight-view approach using a detector for each one of these orientations: right, right-front, left-front, left, left-back, back and back-right. Looking in Table 4.4, 4.5, and 4.6 where the results for the different classes are tabulated, and comparing the SV-HOG/SVM against its MV counterpart, for pedestrian detection (table 4.4) we obtain an AP improvement of $\sim$ 4 (RGB), $\sim$ 3 (Depth) and $\sim$ 2 (Early Fusion) and $\sim$ 2% (Late Fusion). The same behavior is obtained by comparing the other SV pedestrian models against its

Table 4.3: Multi-view partition specification for pedestrians, cyclists, and cars.

Class	View	Angle		Aspect Ratio	Num. Samples
		min	max		
Pedestrian	Left	136	219	2.50	940
	Right	-42	37		
	Front	37	136	2.69	1415
	Back	219	318		
Cyclist	Left	127	234	1.85	328
	Right	-68	49		
	Front	49	127	0.97	760
	Back	234	292		
Car	Right	-44	4	0.37	902
	Right-Front	4	65	0.36	274
	Front	65	107	0.74	1713
	Front-Left	107	151	0.50	4170
	Left	151	219	0.37	1194
	Left-Back	219	257	0.57	1850
	Back	257	284	0.84	4061
	Back-Right	284	316	0.55	1542

MV counterpart: LBP/linSVM (~ 4 / ~ 3 / ~ 6 / ~ 2), HOGLBP/linSVM (~ 4 / ~ 2 / ~ 2 / ~ 3) and HOGLBP/RF (~ 2 / ~ 1 / ~ 0 / ~ 2). Following the same analysis in cyclists (table 4.5) and cars detection (table 4.6), we observe a similar behavior getting improvements in each one of the proposed detectors.

Discussion Each of the mentioned detectors in this section is developed using RGB, Depth and Early Fusion and Late Fusion information sources in order to compare the accuracy under the different conditions. Also for evaluating the multi-view performance the experiments are carried out using a single-view (all samples) and a multi-view (samples divided in different views). In Table 4.4, 4.5, and 4.6 there are the accuracy measurements over the validation set. The measurements include the KITTI evaluation methodology for *easy*, *moderate* and *hard* pedestrian subset. Regarding the obtained results it is easy to see the accuracy improvements at each step of the proposed method. First we can see the improvement introduced by the RF over the other detectors. Comparing the results obtained in each column (training subset and information source) we obtain always the best accuracy in the HOGLBP/RF detector. The second improvement is introduced by the multi-view proposed method, comparing each row (detector) we obtain the best performance for each of the information sources (RGB, Depth, Early Fusion and Late Fusion) when we perform the multi-view ensemble classifier. The third improvement is introduced by the early fusion of information sources, in this case for each detector and given a training subset we obtain the best performance in the Early Fusion experiment.

Finally if we compare the baseline method SM-HOG/linSVM against our proposed MM-RF/LinSVM we obtain an AP gain of ~ 29, ~ 27, and ~ 23 in pedestrians

Table 4.4: Results for PEDESTRIAN detection using different subsets for training (Single-view (SV), Multi-view (MV)), modalities, and detectors, tested over the validation set. For each detector *AP* for KITTI evaluation is shown . Best *AP* for each detector in each modality is indicated in bold, while the best detector across the different modalities in red

Evaluation	Detector	RGB		Depth		Early Fusion		Late Fusion	
		SV	MV	SV	MV	SV	MV	SV	MV
AP (Easy)	HOG/SVM	0.50	**0.54**	0.59	**0.62**	0.71	**0.73**	0.63	**0.65**
	LBP/SVM	0.52	**0.56**	0.62	**0.65**	0.69	**0.75**	0.66	**0.68**
	HOGLBP/SVM	0.64	**0.68**	0.65	**0.67**	0.74	**0.76**	0.70	**0.73**
	HOGLBP/RF	0.73	**0.75**	0.74	**0.75**	0.79	**0.79**	0.77	**0.79**
AP (Moderate)	HOG/SVM	0.38	**0.41**	0.46	**0.47**	0.57	**0.58**	0.49	**0.51**
	LBP/SVM	0.41	**0.44**	0.48	**0.50**	0.57	**0.61**	0.52	**0.54**
	HOGLBP/SVM	0.50	**0.54**	0.51	**0.52**	0.61	**0.62**	0.56	**0.58**
	HOGLBP/RF	0.59	**0.60**	0.58	**0.58**	0.65	**0.66**	0.62	**0.63**
AP (Hard)	HOG/SVM	0.33	**0.35**	0.40	**0.41**	0.50	**0.51**	0.43	**0.44**
	LBP/SVM	0.36	**0.38**	0.42	**0.43**	0.50	**0.53**	0.45	**0.47**
	HOGLBP/SVM	0.43	**0.47**	0.45	**0.46**	0.53	**0.55**	0.49	**0.51**
	HOGLBP-RF	0.51	**0.52**	0.50	**0.50**	0.56	**0.57**	0.54	**0.55**

Table 4.5: Results for CYCLIST detection using different subsets for training (Single-view (SV), Multi-view (MV)), modalities, and detectors, tested over the validation set. For each detector AP for KITTI evaluation is shown . Best AP for each detector in each modality is indicated in bold, while the best detector across the different modalities in red

Evaluation	Detector	RGB		Depth		Early Fusion		Late Fusion	
		SV	MV	SV	MV	SV	MV	SV	MV
AP (Easy)	HOG/SVM	0.43	**0.52**	**0.44**	0.42	0.62	**0.66**	0.48	**0.51**
	LBP/SVM	0.34	**0.48**	**0.48**	0.46	0.62	**0.62**	0.50	**0.53**
	HOGLBP/SVM	0.49	**0.60**	0.48	**0.49**	0.69	**0.69**	0.55	**0.59**
	HOGLBP/RF	0.64	**0.70**	0.49	**0.49**	0.72	**0.73**	0.54	**0.57**
AP (Moderate)	HOG/SVM	0.31	**0.41**	**0.30**	0.29	0.44	**0.49**	0.34	**0.39**
	LBP/SVM	0.29	**0.41**	**0.34**	0.33	0.48	**0.50**	0.38	**0.43**
	HOGLBP/SVM	0.39	**0.50**	0.34	**0.35**	0.52	**0.54**	0.42	**0.48**
	HOGLBP/RF	0.50	**0.57**	0.33	**0.35**	0.52	**0.55**	0.41	**0.45**
AP (Hard)	HOG/SVM	0.28	**0.38**	**0.28**	0.27	0.41	**0.45**	0.32	**0.36**
	LBP/SVM	0.26	**0.38**	**0.31**	0.30	0.45	**0.46**	0.35	**0.39**
	HOGLBP/SVM	0.35	**0.46**	0.32	**0.33**	0.48	**0.50**	0.38	**0.44**
	HOGLBP/RF	0.45	**0.52**	0.31	**0.32**	0.47	**0.50**	0.38	**0.41**

Table 4.6: Results for CAR detection using different subsets for training (Single-view (SV), Multi-view (MV)), modalities, and detectors, tested over the validation set. For each detector *AP* for KITTI evaluation is shown . Best *AP* for each detector in each modality is indicated in bold, while the best detector across the different modalities in red

Evaluation	Detector	RGB		Depth		Early Fusion		Late Fusion	
		SV	MV	SV	MV	SV	MV	SV	MV
AP (Easy)	HOG/SVM	0.26	**0.72**	0.22	**0.78**	0.29	**0.77**	0.17	**0.78**
	LBP/SVM	0.11	**0.62**	0.04	**0.70**	0.11	**0.71**	0.10	**0.71**
	HOGLBP/SVM	0.16	**0.66**	0.18	**0.70**	0.21	**0.72**	0.06	**0.72**
	HOGLBP/RF	0.29	**0.81**	0.38	**0.81**	0.37	**0.82**	0.24	**0.82**
AP (Moderate)	HOG/SVM	0.21	**0.67**	0.17	**0.56**	0.24	**0.69**	0.18	**0.71**
	LBP/SVM	0.11	**0.60**	0.03	**0.61**	0.11	**0.65**	0.12	**0.67**
	HOGLBP/SVM	0.14	**0.65**	0.16	**0.63**	0.19	**0.67**	0.11	**0.68**
	HOGLBP/RF	0.26	**0.75**	0.28	**0.61**	0.29	**0.76**	0.24	**0.75**
AP (Hard)	HOG/SVM	0.17	**0.52**	0.14	**0.44**	0.19	**0.54**	0.15	**0.57**
	LBP/SVM	0.10	**0.48**	0.03	**0.49**	0.09	**0.52**	0.09	**0.54**
	HOGLBP/SVM	0.11	**0.52**	0.13	**0.50**	0.14	**0.54**	0.09	**0.55**
	HOGLBP/RF	0.21	**0.61**	0.22	**0.48**	0.23	**0.62**	0.20	**0.62**

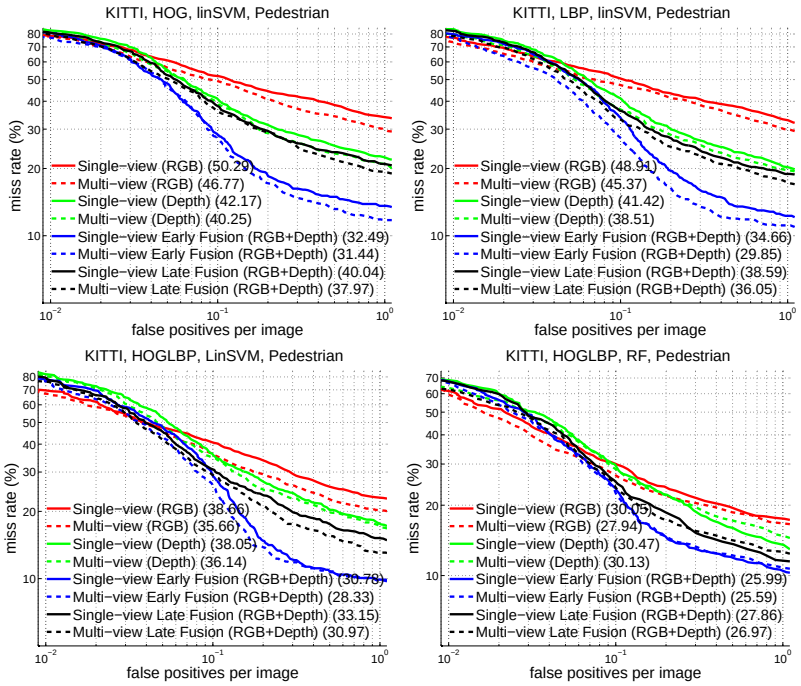


Figure 4.11: Results over validation set using HOG/linSVM, LBP/linSVM, HOGLBP/linSVM, HOGLBP/RF under the different sources: RGB, Depth and RGB+Depth.

detection over the validation set for *easy, moderate, and hard* respectively (Table 4.4).

Finally if we compare the baseline method SV-HOG/linSVM against our proposed multi-cue, multi-modal and multi-view Random Forest of Local Experts we obtain an AP gain of ~ 29 in pedestrians detection, ~ 30 in cyclists detection, and ~ 50 in cars detection, in the validation set (Table 4.4, 4.5, and 4.6 respectively).

Regarding the final approach MV-HOGLBP/RF early fusion of RGB and Depth in table 4.7 and comparing against the methods with an associated paper in the object detection competition, we obtain an AP of 73.30%, 56.59%, 49.63% for the *easy*, *moderate* and *hard* subset respectively, ranking the second best pedestrian detector in the challenge. In table 4.8, we obtain an AP of 53.97%, 42.61%, 37.42% for the *easy*, *moderate* and *hard* subset respectively, ranking the second best cyclist detector in the challenge. Finally in table 4.9, we obtain an AP of 70.40%, 69.92%, 57.47% for the *easy*, *moderate* and *hard* subset respectively, ranking the fifth best car detector in the challenge. Fig. **??**, shows the precision-recall curves obtained over each subset using the final approach.

It is worth to mention that one of the first ranked method, *i.e.*Regionlets [57], appeared posterior to our random forest of local experts but has common key ideas

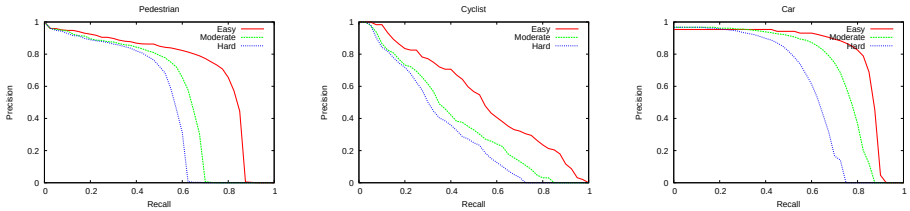


Figure 4.12: Precision-recall curve of the testing set for each subset: *easy*, *moderate* and *hard*, for pedestrian, cyclist and car classes.

Table 4.7: Evaluation and comparison of Multi-view RGBD RF detector using the final test set for PEDESTRIAN detection

Rank	Method	Moderate	Easy	Hard
1	Regionlets	61.15 %	73.14 %	55.21 %
2	**MV-RGBD-RF**	**56.59 %**	**73.30 %**	**49.63 %**
3	pAUCEnsT	54.49 %	65.26 %	48.60 %

such as using HOG and LBP as features, and being patch-based. Thus we think our conclusions also apply for it.

4.5 Conclusions

In this chapter we develop a complete multi-cue, multi-modal and multi-view framework for object detection. We have shown the applicability to different models (holistic, patch-based), obtaining significant accuracy improvements. In this chapter we focus on object detection using HOG/linSVM as baseline applying the different proposed method: different cues (HOG and LBP), different modalities (RGB and Depth) and different views (Frontal, Lateral, etc.). As future work we propose to focus on detection using more complex features (motion, context), classification algorithms (CNN), and modalities (far infrared). Also the candidate generation and re-localization based on segmentation as in [57] could be integrate in this pipeline improving the obtained results.

Table 4.8: Evaluation and comparison of Multi-view RGBD RF detector using the final test set for CYCLIST Detection

Rank	Method	Moderate	Easy	Hard
1	Regionlets	58.72 %	70.41 %	51.83 %
2	**MV-RGBD-RF**	**42.61 %**	**52.97 %**	**37.42 %**
3	pAUCEnsT	38.03 %	51.62 %	33.38 %

Table 4.9: Evaluation and comparison of Multi-view RGBD RF detector using the final test set for Car Detection

Rank	Method	Moderate	Easy	Hard
1	Regionlets	76.45 %	84.75 %	59.70 %
2	3DVP	75.77 %	87.46 %	65.38 %
3	SubCat	75.46 %	84.14 %	59.71 %
4	AOG	71.88 %	84.36 %	59.27 %
5	**MV-RGBD-RF**	**69.92 %**	**76.40 %**	**57.47 %**

Chapter 5

Combining the Visible and Far Infrared Spectrum

Regarding the results obtained in previous chapters, pedestrian detection remains as an extremely challenging problem in real scenarios. In order to develop a detector that successfully operates under different lighting conditions, it becomes necessary to combine multiple imaging modalities including visible and far infrared (FIR). FIR cameras capture energy emitted in the far infrared spectrum not visible to the human eye. FIR cameras capture to thermal information, which is quite invariant to illumination conditions. In this chapter we want to compare the accuracy of pedestrian detectors based on visible images with detectors based on FIR images. Specially we want to study how the accuracy is affected during day and night time. In order to do so we have created a multi-modal dataset containing several on-board sequences simultaneous acquired with visible and FIR cameras during day and night time. Then we trained and evaluated different state-of-the-art detectors over those sequences. Results show that accuracy for FIR sequences is not affected by time conditions. However on visible sequences, as expected, detector accuracy is similar to its FIR counterpart during day time, while during nigh time its accuracy drops dramatically. Finally we tested the multi-modal approach presented in chapter 4 adapted to deal with visible/FIR images. we performed this test in KAIST dataset [43] obtaining competitive results in this benchmark.

5.1 Introduction

Visual pedestrian detection has been receiving attention for more than a decade [37] in the computer vision community due to its multiple applications like Advance Driver Assistance System (ADAS) [23, 25], autonomous vehicles [33], and video surveillance [71], [82], [83]. Pedestrian detection continues being a challenging task still waiting for better solutions. Although different research lines have been proposed,

the accuracy/performance of pedestrian detection methods remains limited in hard scenarios like occlusions, cluttered backgrounds, bad visibility conditions, illumination changes, etc. In this chapter we focus on improving the performance in the latter case. The use of infrared images will allow us to operate under different illumination scenarios during day and night time, addressing the problem of visible spectrum cameras for operating under hard visibility conditions.

Figure 5.1: Setup for dataset acquisition: Stereo-pair FIR/Visible, images with different resolution and field of view.

The infrared/thermal cameras for autonomous vehicles exibit two types of sensors, typically: near infrared ($0.75 \sim 1.3\mu m$), and far infrared ($7.5 \sim 13\mu m$). In [80] it is shown that human body radiates in far infrared (FIR) spectrum ($9.3\mu m$). Accordingly, FIR images have been successfully employed for pedestrian detection in [78], [43], taking advantage of its invariant to different illumination conditions allowing to detect under day/night time (24 hours) without image acquisition problems.

In this chapter we will carry out a comparison of detectors trained with visual and/or FIR images acquired during day and night time. We want to check which modality or combination of tem perform the best in each lighting conditions. At the time of starting this study there was no dataset publicly available, so we acquired a new dataset with sequences at day and night. These sequences are acquired on-board simultaneously with a visible and a FIR cameras in normal driving sessions through urban scenarios. Both cameras are facing forward recording the same scene at the same frame rate but not synchronized (due to hardware limitations of the FIR camera). Our hypothesis is that at day time detectors based on visible spectrum images will outperform FIR-based ones, and the opposit behavior at night. to test it we will evaluate different state-of-the-art features as HOG, LBP and their combination HOGLBP, input to linear SVM, Random Forest and DPM classifiers. The obtained results proved the hypothesis correct during night time when visible-based detectors drop their accuracy drastically, but we found similar performance at day time. Facing this new fact, we propose to use during the day a multi-modal detector which extracts information from both modalities simultaneously as explained in chapter 4. In order to test this approach we required a time and space synchronized dataset that allows us to extract the multi-modal information. Given that our dataset is not synchronized we test the multi-modal approach in the KAIST dataset [43]. Finally our multi-modal results outperform the state-of-the-art methods reported in this dataset.

The rest of the chapter is organized as follow. In section 5.2 we will review the state-of-the-art in FIR detection. Section 5.3 will explain the methodology used to acquire our dataset and the multi-modal approach. Section 5.4 will present the dataset and the experiments carried out to perform our comparison. Finally, section 5.5 draws our main conclusions and future work.

5.2 FIR Detection

Recently FIR images have acquired relevance in object detection applications but very few studies have been carried out about pedestrian detection under FIR images on different times (Day/Night).

There are applications relying on video surveillance [19], [20] and objects tracking [71]. All these approaches work in controlled scenarios in which cameras are in a fixed position.

More related works in on-board detection based in FIR images propose different features, setups, etc. Hwang *et al.* in [43] introduce a new on-board dataset with FIR-visible image pair acquired using a beam splitter hardware, ending up with a space/time synchronized image pair. In order to test their new dataset they propose a multispectral ACF [23] detector adding as a new channels the FIR image intensity and the HOG descriptor calculated over the FIR image. Olmeda *et al.* [64] present a new descriptor, HOPE – *Histograms of Oriented Phase Energy*, specifically targeted for infrared images, and a new adaptation of the latent variable SVM approach to FIR images. HOPE is a contrast invariant descriptor that encodes a grid of local

oriented histograms extracted from the phase congruency of the images computed from a joint of Gabor filters. In [49] is presented an analysis of color-, infrared-, and multimodal-stereo pedestrian detection approaches, using a four-camera experimental setup consisting of two color and two infrared cameras. In order to detect pedestrians they use a detector based on [17] using HOG descriptor and SVM with a radial basis kernel. This detector is evaluated over different modalities and a combination of them. All these approaches mentioned above try to cover different aspects of FIR detection, such as new features designed for enhance FIR images characteristics [64], or advantages of a multimodal-stereo approach [49], or a new multispectral descriptor which extracts information from both sources simultaneously [43]. In contrast to these approaches we test state-of-the-art methods and perform a fair comparison of them under different time/modality conditions. Providing in this way a solid baseline which can be used as starting point for designing new multi-modal approaches getting the maximum advantages of both sensors during each time condition.

5.3 Methodology

In this section we will start by introduce our new dataset setup acquisition method (Subsection 5.3.1). Finally the multi-modal approach combining visible and FIR images in presented in subsection 5.3.2.

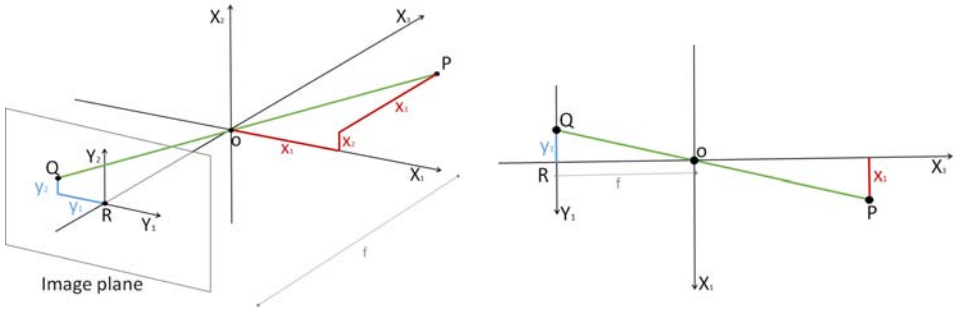


Figure 5.2: Geometry of a pinhole camera model. (a) shows a 3D view of the model and (b) a 2D view seen from the X2 axis.

5.3.1 Dataset Acquisition Setup

In order to get a dataset that allows us to evaluate the performance under different conditions we recorded two sequences using a couple of cameras one visible the other FIR (see Fig. 5.1), one during the day time and the other at night. We used UI-3240CP camera (Visible images) and FLIR Tau 2 camera (FIR images), see specifications of mentioned cameras in Table 5.1. Notice that neither resolution nor

Table 5.1: FLIR Tau 2 and UI-3240CP camera specifications.

Specifications	**FLIR Tau 2**	**IDS UI-3240CP**
Resolution	640 x 512 pixels	1280 x 1024 pixels
Pixel size	17 μm	5.3 μm
Lens length	13mm	Adjustable
Sensitive area	10.88mm x 8.7mm	6.784 mm x 5.427 mm
Frame rate	30/25 Hz (NTSC/PAL)	60 fps

field-of-view match, capturing images of different size and covering different area of the scenes. The visible camera produces a wide field of view with high resolution, whereas the FIR camera resolution is very limited. However, we expect that FIR camera resolution will be increasing in the next years.

For the purpose of this work, we must have images with same field of view ans resolution, because it is mandatory that every target object have similar number of pixels assigned in both images, visible and FIR. Since the FIR camera specifications can not be changed, we need to adjust the visual camera settings. A plausible solution would be to calculate the size of the lens that the camera must have to produce the same image as the FIR camera in terms of field of view and resolution. To do so, it is necessary to follow the *pinhole camera model*.

Pinhole camera model is a geometric model which states that a 3D point (real world coordinates) can be mapped to a 2D point (image plane coordinates) by geometric calculation, as shown in Figure 5.2. Following this model and taking into account that the visible camera geometry can be modified, we adjust the focal length to assure that a given object in the real world is projected to both cameras with the same height in pixels. In order to calculate the focal length of the visible camera we follow the following procedure. The point O represents the origin in the $< X1, X2, X3 >$ world coordinate system, where P and Q are a real world point and its projection in the image plane, respectively. The X_3 represents the distance from P to the $(X1; X2)$ image plane and f is the distance from O to the image plane.

If we apply trigonometry basics to this problem we can compute the distance from the $X2$ axis to the points P and Q. According to the similar triangles principle we realize that y_1 (Figure 5.2b) can be calculated as follows:

$$|y_1| = \frac{f * X_1}{X_3} \tag{5.1}$$

Since the aim of this work is to have similar pixel distribution in both FIR and visual images, it is necessary to compute the pixel distribution for the FIR camera and adjust the visual camera settings to the FIR output. Suppose having a target object of $1.5m$ tall that is located $3.0m$ away from the camera. Here, it is necessary to find out the number of pixels of this object in the image. Hence, equation 5.1 can be applied

to solve the problem. Considering the FIR camera parameters we have $f = 13mm$ and using the object height ($X_1 = 1500mm$) and distance ($X_3 = 3000mm$), hence, y_1 represents the number of pixels of the target object. According to equation 5.1 $y_1 = 6.5mm$, taking in to account the size of a pixel ($17\mu m$) we obtain 382.35 pixels for this object. For getting the focal length needed for the visual camera and taking into account the pixel size for the visible camera, by using eq. 5.1 we obtain that 382.35 pixels corresponds to $2.027mm$ (y_1) in this camera. Replacing y_1, X_1, and X_3 if equation 5.1 we obtain $f = 4.054mm$. Accordingly we set this value in our visible camera, capturing objects with same height in both images. In Figure 5.3 we have image examples for acquired sequences using this setup, where objects in both images (Visible and FIR) have the same height and recorded images have same field of view.

5.3.2 Multi-modal Approach

Keeping in mind that more complementarity information is better for pedestrian detection, we want to explore the integration of different modalities. In this case visible and FIR images. In order to generate a multi-modal detector we propose to use a similar approach than in previous chapter **??**. For each candidate window we extract HOG and LBP features over each modality (visual and FIR). Then, we combine these features into a single detector (Random Forest of local experts). We combine the features using an early fusion approach; using this approach we train a single model using as descriptor the concatenation of the features computed at each modality. Keeping in mind that the acquired dataset is not synchronized in time (lag) and space (stereo disparity), we propose to test our approach in the KAIST dataset [43]. This dataset is time and space synchronized, meaning that each pixel in a pair of images refers to the same point in the scene, allowing us to extract multi-modal information.

5.4 Experimental Results

In this section we assess the performance comparison between detectors trained under different modalities (FIR/Visible), analyzing the impact in the performance when the same detector operates under different time conditions (Day/Night). For comparison, we have chosen a bunch of detectors starting with features like HOG [17], LBP [89] and the concatenation of both (HOG+LBP), using linear-SVM as classifier. Then evaluating with more complex detectors based on the previous ones but using a random forest of local experts (RF) [59] or a deformable part-based model (DPM) [28].

Evaluation Protocol As evaluation methodology we follow the de-facto Caltech standard for pedestrian detection [24], *i.e.*, we plot curves of false positives per image (FPPI) *vs* miss rate. The miss rate average AMR in the range of 10^{-2} to 10^{0} FPPI is taken as indicative of each detector accuracy, *i.e.*. the lower the better. Moreover,

Figure 5.3: CVC Multispectral FIR/Visible Pedestrian Dataset image examples. During day time both sensors provide useful information, During Night visible camera cant capture details of pedestrians providing poor information.

during testing we consider three different subset: Near, medium, and reasonable. The *near* subset includes pedestrians with height equal or higher than 75 pixels. The *medium* subset includes pedestrian between 50 and 75 pixel height. Finally we group the two previous subset in the *reasonable* subset ($height >= 50px$).

CVC Multispectral FIR/Visible Pedestrian Dataset In order to perform this study we acquire and present a new multispectral pedestrian dataset. In this case, FIR and visible cameras were not hardware synchronized. However, they were running at same frame rate (10 fps) and started at the same time for recording sequences concurrently. Thus even there is a time shift between FIR and visible sequence pair, this is negligible for comparing the performance of the pedestrian detectors (*e.g.*, see image pair in Fig. 5.3).

In Table 5.2 is summarized the number of frames and annotated pedestrian for each one of the subsets: Day/FIR, Night/FIR, Day/Visible, and Night/Visible. There is defined as mandatory pedestrian the ones with height larger than 50 pixels. This new dataset has more than 2000 annotated pedestrian of each subset, where more than 1300 are mandatory. Finally for testing it has more than 1500 annotated, where more than 1300 are mandatory. In Figure 5.3 are shown examples of image pairs Visible-FIR for both time condition scenarios. In Figure 5.4 are shown some examples

Figure 5.4: CVC Multispectral FIR/Visible Pedestrian Dataset crops examples.

of pedestrian crops. During day time we can see that in shadows areas objects are well defined in FIR images, while areas with normal light conditions objects are well defined in visible image (if background is not clutter). During night time while in visible images only objects in areas illuminated by car lights are well defined, in FIR images objects are clearly recognizable no matter their position and illumination conditions.

KAIST Multispectral Pedestrian Dataset The KAIST Multispectral Pedestrian dataset [43] is acquired using a hardware consisting of a color camera, a thermal camera and a beam splitter to capture the aligned multispectral (RGB color + Thermal). The sequences are acquired in on-board traffic scenarios. This dataset consists of $95k$ color-thermal pairs (640x480, 20Hz) taken from a vehicle, and manually annotated (person, people, and cyclist). Ending up 103,128 dense annotations and 1,182 unique pedestrians.

Table 5.2: New dataset resume of images and annotated pedestrian.

Set	Variable	FIR		Visible	
		Day	Night	Day	Night
Training	Positive Frames	2232	1386	2232	1386
	Negative Frames	1463	2004	1463	2004
	Annotated Pedestrians	2769	2222	2672	2007
	Mandatory Pedestrians	1327	1787	1514	1420
Testing	Frames	706	727	706	727
	Annotated Pedestrians	2433	1895	2302	1589
	Mandatory Pedestrians	2184	1541	2079	1333

Experiments over CVC dataset In order to evaluate the detection accuracy over different modalities, we train/test a bunch of detectors under all possible combinations of modalities and time conditions. In Figures 5.5, and 5.6 we can see the plots obtained using the caltech standard evaluation protocol for the different

proposed detectors: HOG/LinSVM, LBP/LinSVM, HOG+LBP/LinSVM, HOG/RF, HOG+LBP/RF, HOG/DPM, and HOG+LBP/DPM. *AMR* from all experiments in Figures 5.5, and 5.6 are summarized in Table 5.3. Obtained results show that the same detector using FIR images as information source outperforms (less miss rate) the one which uses visible spectrum images, no matter the time condition (Day/Night) under the experiments is performed. Also shows the invariance to time conditions when the detector is applied over FIR images, obtaining similar performance for sequences during day and night time, while the experiments over visible images have different performance, due to the lack of information in visible images during the night time.

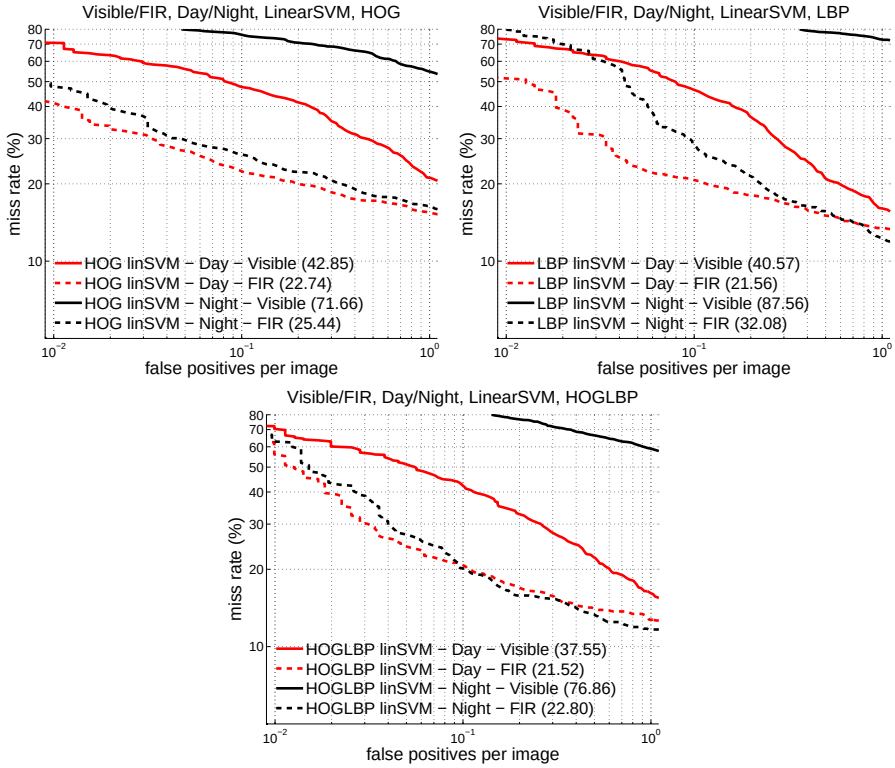


Figure 5.5: Results using SVM detectors over CVC multispectral dataset.

Experiments over KAIST dataset In order to evaluate the multimodal detector in the new benchmark presented in [43] for RGB/FIR images we perform HOG+LBP/RF detector with and early fusion techniques for merge the different modalities. For providing a complete comparison we test the proposed method in different subsets of annotated pedestrians: Reasonables, near, and medium. In Table 5.4 are the average miss rates for the different detectors over the different subsets. In Figure 5.7 is shown the complete curves of some representative results.

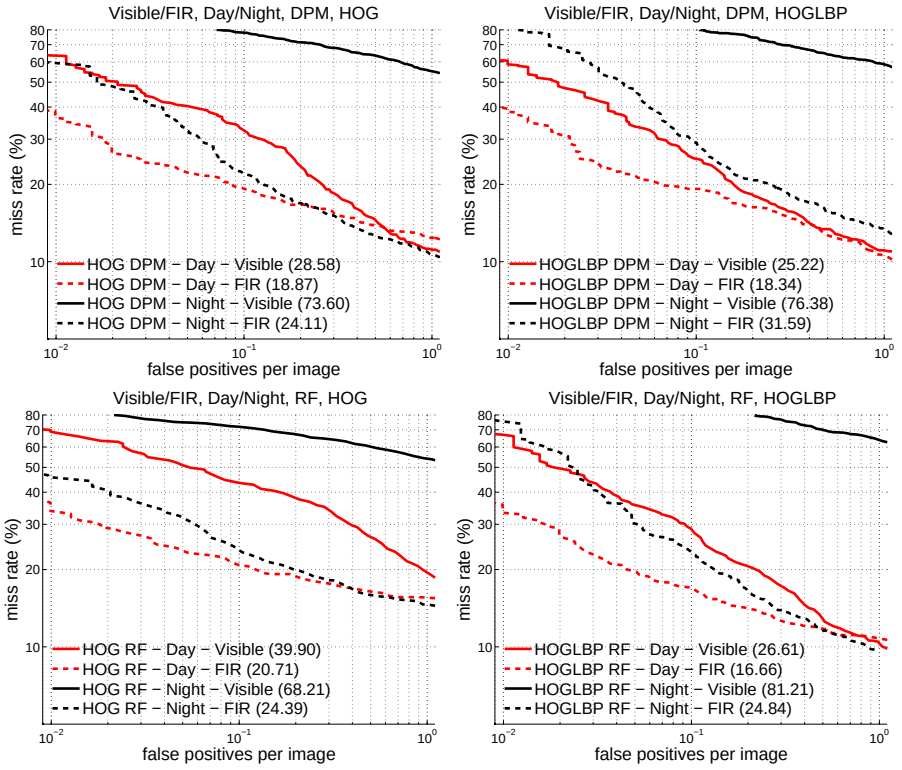


Figure 5.6: Results using different detectors over CVC multispectral dataset. Top row detectors based on DPM, bottom row based on RF.

Discussion Results obtained during the performed study reveal that during day time visible and FIR cameras provide useful information for applying vision-based pedestrian detection algorithm; obtaining similar accuracy in those conditions. While in night time visible cameras can not capture pedestrian details. This problem is solved by a FIR camera; making the detector to perform well obtaining low miss rate in the whole FPPI evaluation range. This can be explained by the temperature difference between humans and the scene, which highlights humans profiles in the scene giving rise to well-defined contours. In consequence, we propose a multi-modal detector and test it on the KAIST dataset. This multi-modal approach outperforms single-modal approaches during the day when both sensors provide useful information. During night due to the bad performance of visible spectrum cameras, the multi-modal solution perform similar to FIR-based one but is affected by the noisy information of the visible spectrum information decreasing its accuracy. In this evaluation we obtain competitive detectors in this benchmark.

Table 5.3: CVC Multispectral FIR/Visible Pedestrian Dataset results for different detectors over different time conditions

Detector	FIR		Visible	
	Day	Night	Day	Night
HOG/LinSVM	22.74	25.44	42.85	71.66
LBP/LinSVM	21.56	32.08	40.57	87.56
HOGLBP/LinSVM	21.52	22.80	37.55	76.86
HOG/DPM	18.87	24.11	28.58	73.60
HOGLBP/DPM	18.34	31.59	25.22	76.38
HOG/RF	20.71	24.39	39.90	68.21
HOGLBP/RF	16.66	24.84	26.61	81.21

5.5 Conclusions

In this chapter we have presented an exhaustive study of pedestrian detection using visible and FIR sensors operating during day and night time. This evaluation is based on well known features HOG and LBP and holistic, patch-based, and part-based models. Taking into account the detector performances under the different time conditions and the sensor used for acquiring the images, we propose a multi-modal approach that extracts information from both sensors during the day time. This approach outperforms the state-of-the-art method in the KAIST dataset [43], showing that a simple early fusion of features extracted from both sensors can generate competitive detectors. Regarding the obtained results we propose the study of convolutional neural networks to pedestrian detection in FIR images, and the design of a network that deals with multi-modal information.

Table 5.4: KAIST Multispectral Pedestrian Dataset multi-modal results for HOG/LinSVM, LBP/LinSVM nad HOGLBP/RF detectors over different subsets: Reasonables, near, medium, partial occluded and heavy occluded.

Detector	Time	Modality	Reasonable	Near	Medium
HOG/LinSVM	**Day**	Visible	84.20	59.53	84.48
		FIR	79.85	**49.44**	80.68
		Multi-modal	**78.23**	52.03	**76.99**
	Night	Visible	96.94	89.57	98.74
		FIR	**59.87**	**35.31**	**62.32**
		Multi-modal	80.94	65.18	74.91
LBP/LinSVM	**Day**	Visible	84.52	57.35	85.25
		FIR	75.90	38.93	77.42
		Multi-modal	**72.47**	**36.52**	**74.10**
	Night	Visible	87.23	66.30	91.13
		FIR	**59.25**	**31.94**	**63.34**
		Multi-modal	61.54	32.57	66.98
HOGLBP/RF	**Day**	Visible	72.66	39.75	74.52
		FIR	70.52	31.52	72.53
		Multi-modal	**65.75**	**28.75**	**66.44**
	Night	Visible	91.43	75.91	93.22
		FIR	**53.51**	**25.33**	**59.96**
		Multi-modal	56.68	29.36	61.69

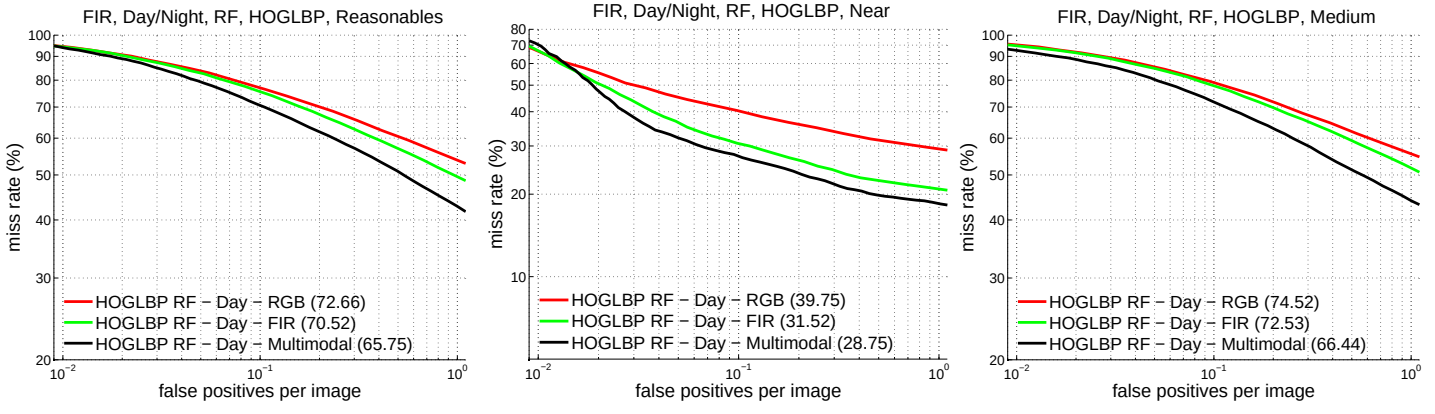


Figure 5.7: Results using different test subsets over KAIST multispectral dataset during day time. Left plot include results tested over reasonable subset, middle plot over near subset, right plot over medium dataset.

Figure 5.8: Qualitative Results comparing HOG/LinSVM detector in different time/sensor conditions. Blue boxes represent correct detections (True Positive), while red boxes represent misdetections (False Negative), number on the top of detection represent the classification score.

Chapter 6

Conclusions

In this Thesis we explore the inclusion of multiple information sources in order to increase the robustness of base pedestrian detectors. During this Thesis we develop approaches that include temporal information, information from different modalities, and develop of a strong multi-view approach.

The work included in this dissertation can be divided in three main topics. First the study and development of a method that includes spatiotemporal information. Second the developments of a multi-view, multi-modal, and multi-cue detector. Finally the study of FIR images as an alternative information source for vision-based pedestrian detectors.

In **chapter 3** we develop a novel method for introducing spatiotemporal information. This information until now was introduced in post-detection stages, instead we propose to include it at the pedestrian description level. We show how even simple projection windows can boost the detection accuracy of different base classifiers in on-board sequences.

In **chapter 4** we perform a full study of a multi-cue, and multi-modal approach, which combines information of visible cameras with depth information. To this end we test two different depth information sources (stereo and LIDAR). Finally we extend the Random Forest of local Experts detector for dealing with multi-modal information. Following this line, we extend the previous approach with the inclusion of a strong multi-view technique. We propose an automatic partition of the training set for views definition. This multi-view extension boost the detectors accuracy by reducing the intra-class variability regarding the samples of each view. Finally this multi-view, multi-modal, and multi-cue approach ranks among the top detectors in the KITTI benchmark for pedestrians, cyclists, and cars detection.

In **chapter 5** we perform a an exhaustive study of applicability for far infrared (FIR) cameras to pedestrian detection during day and night time. This evaluation was performed under different detectors: holistic, patch-based, and part-based models. Obtaining as a result clear evidence about the better performance of FIR-based

detectors against visible-based one at night, while at day time both detectors perform similar. Also in this chapter we adapt the multi-modal approach of previous chapter for introducing FIR/thermal information in the pedestrian detection framework, this multi-modal approach outperform the single-modal ones and we obtain competitive results in the KAIST multi-spectral dataset.

At the end of this Thesis we present two new on-board dataset for pedestrian detection. The first one acquired for spatiotemporal information was record at 30 FPS, and allows comparison at three different frame rates 30 FPS, 10 FPS and 3 FPS. The second acquired with a not synchronized pair visible/FIR during day and night time. Both datasets and their specifications are presented in chapters 3, and 5 respectively.

Therefor, with respect to the questions we set as objectives in chapter 1, we can state that:

- *How to introduce temporal information in the classification stage of a pedestrian detector?*
  Answer. Introducing spatiotemporal information by following SSL paradigm, allow us to boost accuracy of base classifiers. Re-utilizing the descriptors previously computed in previous frames, SSL paradigm does not introduce extra description calculation having same computational cost that base classifiers.
- *The combination of depth with visual information does improve the use of these modalities in isolation?*
  Answer. Multi-modal RGB-D approaches boost accuracy of single-modal ones. Multi-modal approaches capture complementary information obtaining more robust classifiers.
- *The combination of FIR with visual spectrum information does improve the use of these modalities in isolation?*
  Answer. Multi-modal Visible-FIR approaches boost accuracy of single-modal ones during day time, but at night visible information is so poor that detectors using the multi-modal approach have lower accuracy that FIR one.

6.1 Future Work

Pedestrian detection remains as a not-completed solved challenge. In this Thesis we include robustness to base classifiers by introducing alternative information sources, despite of these improvements, detection rates in real urban scenarios is still a big challenge. Regarding the recent big improvements in different computer vision task of techniques based on neural networks, we propose as future research line to adapt these neural network approaches for including our proposed information sources (Temporal, multi-view, and multi-modal).

As a first proposed future work, we propose to research the *Recurrent Convolutional Neural Networks (***R-CNN***)* [51, 55], which include temporal/sequential in-

formation in their architecture. R-CNN define concatenated CNN where outputs of previous stages go to following stages as features. This architecture follows the same paradigm as SSL, so it is expected that a good adaptation of it to spatiotemporal pedestrian detection could boost the accuracy. Continuing with the spatiotemporal information, to research for motion invariant methods for volume definition, will allow a better temporal description of pedestrian. These volume definition method should avoid object motion and egomotion, also should be aware of object size changes during the sequence.

As a second future research line, the definition of a CNN architecture that allows the integration of multiples modalities, should provide with a high accuracy detector.

List of Publications

This dissertation has led to the following communications:

Conference Contributions

- A. González, D. Vázquez, S. Ramos, Antonio M. López, J. Amores. Spatiotemporal Stacked Sequential Learning for Pedestrian Detection. **In Iberian Conf. on Pattern Recognition and Image Analysis**. 2015.

- A. González, G. Villalonga, G. Ros, D. Vázquez, J. Amores, Antonio M. López. 3D-Guided Multiscale Sliding Window for Pedestrian Detection. **In Iberian Conf. on Pattern Recognition and Image Analysis**. 2015.

- A. González, G. Villalonga, J. Xu, D. Vázquez, J. Amores, Antonio M. López. Multiview Random Forest of Local Experts Combining RGB and LIDAR data for Pedestrian Detection. **In IEEE Intelligent Vehicles Symposium**. 2015.

Submitted Journal Papers

- A. González, D. Vázquez, Antonio M. López, J. Amores. Pedestrian Detection: Spatiotemporal Stacked Sequential Learning. **IEEE Trans. on Cybernetics**.

- A. González, D. Vázquez, Antonio M. López, J. Amores. On-Board Object Detection: Multi-cue, Multi-modal and Multi-view Random Forest of Local Experts. **IEEE Trans. on Cybernetics**.

Exhibitions

- A. González, G. Villalonga, G. Ros, D. Vázquez, J. Amores, Antonio M. López. 3D pedestrian detection. **In Iberian Conf. on Pattern Recognition and Image Analysis**. Santiago de Compostela, 2015

- A. Gonzalez, G. Villalonga, G. Ros, D. Vazquez, and A. M. Lopez. 2D pedestrian detection. **In Saló del Ensenyament**. barcelona, 2015

- A. Gonzalez, G. Villalonga, A. Shuvin, L. Sellart, G. Ros, D. Vazquez, and A. M. Lopez. Autonomous Vehicle. **In MEMEnginy/NOVUM**. Barcelona, 2015.

Bibliography

[1] Y. Abramson and Y. Freund. SEmi-automatic VIsuaL LEarning (SEVILLE): a tutorial on active learning for visual object recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005.

[2] T. Ahonen, A. Hadid, and Pietikainen M. Face recognition with local binary patterns. In *European Conf. on Computer Vision*, 2004.

[3] B. Alexe, T. Deselares, and V. Ferrari. Measuring the objectness of image windows. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2012.

[4] I. P. Alonso, D. F. Llorca, M. A. Sotelo, L. M. Bergasa, Pedro Revenga de Toro, J. Nuevo, M. Ocana, and M. A.G. Garrido. Combination of feature extraction methods for SVM pedestrian detection. *IEEE Trans. on Intelligent Transportation Systems*, 2007.

[5] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *In: Proc. Computer Vision and Image Understanding (VIU*.

[6] Jens Behley, Volker Steinhage, and Armin B. Cremers. Laser-based segment classification using a mixture of bag-of-words. In *International Conference on Intelligent Robots and Systems*, 2013.

[7] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool. Pedestrian detection at 100 frames per second. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012.

[8] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool. Seeking the strongest rigid detector. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2013.

[9] J Carreira and C Sminchisescu. CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(7):1312–1328, 2012.

[10] k.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai. Fusing generic objectness and visual saliency for salient object detection. In *Int. Conf. on Computer Vision*, 2011.

[11] Guang Chen, Yuanyuan Ding, Jing Xiao, and Tony Han. Detection evolution with multi-order contextual co-occurrence. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Portland, 2013.

[12] Y.-T. Chen and C.-S. Chen. Fast human detection using a novel boosted cascading structure with meta stages. *IEEE Trans. on Image Processing*, 17(8):1452–1464, 2008.

[13] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip H. S. Torr. BING: Binarized normed gradients for objectness estimation at 300fps. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2014.

[14] W. Cohen and V. de Carvalho. Stacked sequential learning. In *Int. Joint Conferences on Artificial Intelligence*, Edinburgh, Scotland, 2005.

[15] X. Cui, Y. Liu, S. Shan, X. Chen, and W. Gao. 3d haar-like features for pedestrian detection. In *IEEE Int. Conf. on Multimedia & Expo*, Bejing, China, 2007.

[16] N. Dalal. *Finding People in Images and Videos*. PhD thesis, Institut National Polytechnique de Grenoble, 2006. Advisors: Cordelia Schmid and William J. Triggs.

[17] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005.

[18] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *European Conf. on Computer Vision*, Graz, Austria, 2006.

[19] James W. Davis and Mark A. Keck. A two-stage template approach to person detection in thermal imagery. In *IEEE Winter Conference on Applications of Computer Vision*, 2005.

[20] James W. Davis and Vinay Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. In *Computer Vision and Image Understanding*, 2007.

[21] Yuanyuan Ding and Jing Xiao. Contextual boost for pedestrian detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012.

[22] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *British Machine Vision Conference*, London, UK, 2009.

[23] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: a benchmark. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, 2009.

[24] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: an evaluation of the state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012.

[25] M. Enzweiler and D.M. Gavrila. Monocular pedestrian detection: survey and experiments. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(12), 2009.

[26] M. Enzweiler and D.M. Gavrila. A multi-level mixture-of-experts framework for pedestrian classification. *IEEE Trans. on Image Processing*, 20(10), 2011.

[27] M. Enzweiler and D.M. Gavrila. A mixed generative-discriminative framework for pedestrian classification. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008.

[28] P. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[29] P. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. In *Int. Journal on Computer Vision*, 2004.

[30] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun. Salient object detection by composition. In *Int. Conf. on Computer Vision*, 2011.

[31] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

[32] T. Gandhi and M.M. Trivedi. Pedestrian protection systems: issues, survey, and challenges. *IEEE Trans. on Intelligent Transportation Systems*, 8(3):413–430, 2007.

[33] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR'12*, 2012.

[34] D. Gerónimo, A.D. Sappa, A.M. López, and D. Ponsa. Pedestrian detection using adaboost learning of features and vehicle pitch estimation. In *IASTED Int. Conference on Visualization, Imaging and Image Processing*, Palma de Mallorca, Spain, 2006.

[35] D. Gerónimo, A.M. López, A.D. Sappa, and T. Graf. Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(7):1239–1258, 2010.

[36] D. Gerónimo and A.M. López. *Vision-based Pedestrian Protection Systems for Intelligent Vehicles*. Springer, 2013.

[37] D. Gerónimo, A.D. Sappa, D. Ponsa, and A.M. López. 2D-3D based on-board pedestrian detection system. *Computer Vision and Image Understanding*, 114(5):583–595, 2010.

[38] Chunhui Gu, Joseph J. Lim, Pablo Arbeláez, and Jitendra Malik. Recognition using regions. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

[39] D. Hall and J. Llinas. An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1):6–23, 1997.

[40] J. Hosang, R. Benenson, M. Omran, , and B. Schiele. Taking a deeper look at pedestrians. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2015.

[41] J. Hosang, R. Benenson, and B. Schiele. How good are detection proposals, really? In *British Machine Vision Conference*, 2014.

[42] M. Hussein, F. Porikli, and L. Davis. A comprehensive evaluation framework and a comparative study for human detectors. *IEEE Trans. on Intelligent Transportation Systems*, 10(3):417–427, 2009.

[43] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. June 2015.

[44] M. Jones and D. Snow. Pedestrian detection using boosted features over many frames. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008.

[45] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *Int. Conf. on Computer Vision*, Beijing, China, 2005.

[46] Kiyosumi Kidono, Takeo Miyasaka, Akihiro Watanabe, Takashi Naito, and Jun Miura. Pedestrian recognition using high-definition lidar. In *IEEE Intelligent Vehicles Symposium*, 2011.

[47] Kiyosumi Kidono, Takashi Naito, and Jun Miura. Reliable pedestrian recognition combining high-definition lidar and vision data. In *IEEE Int. Conf. on Intelligent Transportation Systems*, 2012.

[48] Alexander Kläser, Marcin Marszalek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, 2008.

[49] Stephen J. Krotosky and Mohan Manubhai Trivedi. On color-, infrared-, and multimodal-stereo approaches to pedestrian detection. In *IEEE Trans. on Intelligent Transportation Systems*, 2007.

[50] J. Lafferty, A. McCallum, and F. Pereira. Real-time pedestrian detection with deformable part models. In *IEEE Intelligent Vehicles Symposium*, Madrid, Spain, 2012.

[51] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *The AAAI Conference on Artificial Intelligence: Physically Grounded AI Track*, 2015.

[52] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *Int. Journal on Computer Vision*, 2008.

[53] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2005.

[54] Fuxin Li, J. Carreira, and C. Sminchisescu. Object recognition as ranking holistic figure-ground hypotheses. In *Computer Vision and Pattern Recognition*, 2010.

[55] Ming Liang and Xiaolin Hu. Recurrent convolutional neural network for object recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2015.

[56] Stuart P Lloyd. Least squares quantization in pcm. *IEEE Trans. on Information Theory*, 28:129–137, 1982.

[57] Chengjiang Long, Xiaoyu Wang, Ming Yang, and Yuanqing Lin. Accurate object detection with location relaxation and regionlets relocalization. In *Asian Conf. on Computer Vision*, 2014.

[58] J. Marin, D. Vázquez, A.M. López, J. Amores, and L.I. Kuncheva. Occlusion handling via random subspace classifiers for human detection. *IEEE Trans. on Systems, Man, and Cybernetics (Part B)*, 2013.

[59] J. Marin, D. Vázquez, A.M. López, J. Amores, and B. Leibe. Random forests of local experts for pedestrian detection. In *Int. Conf. on Computer Vision*, Sydney, Australia, 2013.

[60] W. Nam, B. Han, and J. Han. Improving object localization using macrofeature layout selection. In *Int. Conf. on Computer Vision- Workshop on Visual Surveillance*, Barcelona, Spain, 2013.

[61] Luis E. Navarro-Serment, Christoph Mertz, and Martial Hebert. Pedestrian detection and tracking using three-dimensional ladar data. *International Journal of Robotics Research*, 29(12).

[62] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[63] L. Oliveira, U. Nunes, and P. Peixoto. On exploration of classifier ensemble synergism in pedestrian detection. *IEEE Trans. on Intelligent Transportation Systems*, 11(1):16–27, 2010.

[64] Daniel Olmeda, Cristiano Premebida, Urbano Nunes, Jose Maria Armingol, and Arturo de la Escalera. Pedestrian detection in far infrared images. In *Integrated Computer-Aided Engineering*, 2013.

[65] World Health Organization. Global status report on road safety 2013: Suporting a decade of action. In *World Health Organization*, 2013.

[66] Wanli Ouyang and Xiaogang Wang. Joint deep learning for pedestrian detection. In *Int. Conf. on Computer Vision*, 2013.

[67] Yanwei Pang, Kun Zhang, Yuan Yuan, and Kongqiao Wang. Distributed object detection with linear svms. *IEEE Trans. on Systems, Man, and Cybernetics (Part B)*, 44(11):2122–2133.

[68] C. Papageorgiou and T. Poggio. A trainable system for object detection. *Int. Journal on Computer Vision*, 38(1):15–33, 2000.

[69] S Paris, P Kornprobst, J Tumblin, and F Durand. Bilateral filtering: Theory and applications. *Foundations and Trends in Computer Graphics and Vision*, 4(1):1–73, 2009.

[70] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *European Conf. on Computer Vision*, Crete, Greece, 2010.

[71] J. Portmann, S. Lynen, M. Chli, and R. Siegwart. People detection and tracking from aerial thermal views. 2014.

[72] C. Premebida, J. Carreira, J. Batista, and U. Nunes. Pedestrian detection combining rgb and dense lidar data. In *International Conference on Intelligent Robots and Systems*, 2014.

[73] D. Ramanan. *Part-based Models for Finding People and Estimating Their Pose*. Springer, 2009.

[74] P. Rantalankila, J. Kannala, and E. Rahtu. Generating object segmentation proposals using global and local search. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2014.

[75] M.A. Rao, D. Vázquez, and A.M. López. Color contribution to part-based person detection in different types of scenarios. In *Int. Conf. on Computer Analysis of Images and Patterns*, 2011.

[76] N. Schneider and D.M. Gavrila. Pedestrian path prediction with recursive bayesian filters: A comparative study. In *German Conference on Pattern Recognition*, Saarbrücken, Germany, 2013.

[77] E. Seemann, B. Leibe, K. Mikolajczyk, and B. Schiele. An evaluation of local shape-based features for pedestrian detection. In *British Machine Vision Conference*, 2005.

[78] Yainuvis Socarras, Sebastian Ramos, David Vázquez, Antonio M. López, and Theo Gevers. Adapting pedestrian detection from synthetic to far infrared images. In *Int. Conf. on Computer VisionWorkshop*, 2011.

[79] L. Spinello, K. O. Arras, R. Triebel, and R. Siegwart. A layered approach to people detection in 3d range data. In *The AAAI Conference on Artificial Intelligence: Physically Grounded AI Track*, 2010.

[80] Louis St-Laurent, Xavier Maldague, and Donald Prévost. Combination of colour and thermal sensors for enhanced object detection. In *Information Fusion*, pages 1–8. IEEE, 2007.

[81] D. Tang, Y. Liu, and T.-K. Kim. Fast pedestrian detection by cascaded random forest with dominant orientation templates. In *British Machine Vision Conference*, 2012.

[82] M. Teutsch, T. Mller, M. Huber, and J. Beyerer. Low resolution person detection with a moving thermal infrared camera by hotspot classification. In *IEEE Conf. on Computer Vision and Pattern RecognitionWorkshop*, 2014.

[83] M. Torabi, G. Mass, and G. A. Bilodeau. An interative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for video serveillance applications. In *Computer Vision and Image Understanding*, 2012.

[84] Koen E. A. van de Sande, Jasper R. R. Uijlings, Theo Gevers, and Arnold W. M. Smeulders. Segmentation as selective search for object recognition. In *Int. Conf. on Computer Vision*, 2011.

[85] D. Vázquez, A.M. López, J. Marín, D. Ponsa, and D. Gerónimo. Virtual and real world adaptation for pedestrian detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2013.

[86] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Int. Conf. on Computer Vision*, Nice, France, 2003.

[87] S. Walk, N. Majer, K. Schindler, and B. Schiele. New features and insights for pedestrian detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010.

[88] M. Wang and X. Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2011.

[89] X. Wang, T.X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In *Int. Conf. on Computer Vision*, Kyoto, Japan, 2009.

[90] Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2013.

[91] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Lecture Notes in Computer Science. 363-370*, 2003.

[92] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, 2009.

[93] B. Wu and R. Nevatia. Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. *Int. Journal on Computer Vision*, 2009.

[94] Chen Xiaogang, Wei Pengxu, Ke Wei, Ye Qixiang, and Jiao Jianbin. Pedestrian detection with deep convolutional neural network. In *Asian Conf. on Computer Vision*, 2014.

[95] Jiejun Xu, Kyungnam Kim, Zhiqi Zhang, Hai-wen Chen, and Yuri Owechko. 2d/3d sensor exploitation and fusion for enhanced object detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2014.

[96] Yanwu Xu, Dong Xu, S. Lin, T.X. Han, Xianbin Cao, and Xuelong Li. Detection of sudden pedestrian crossings for driving assistance systems. *IEEE Trans. on Systems, Man, and Cybernetics (Part B)*, 42(3):729–739, 2012.

[97] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2011.

[98] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l1 optical flow. In *German Association for Pattern Recognition (DAGM) Conference*, Heidelberg, Germany, 2007.