UNIVERSITY OF ALCALA

ESCUELA POLITÉCNICA SUPERIOR

Department of Electronics

STEREO VISION BASED PEDESTRIAN DETECTION SYSTEM FOR ASSISTED DRIVING

Author

David Fernández Llorca

Director

Miguel Ángel Sotelo Vázquez

2008

PhD THESIS (Summary)

Abstract

A stereo vision-based pedestrian detection system, in the visible spectrum, onboard intelligent vehicles is proposed in this thesis, with the intention of increasing the security of the most vulnerable road users. By means of the use of two calibrated cameras it is possible to obtain 3D information of the environment without the constraints related to monocular systems. Non dense 3D maps are generated by using the epipolar geometry and a robust correlation process applied over the Canny points, which are computed using adaptive thresholds based on the gradient magnitude. Depth accuracy of the 3D maps depends on the images resolution and the distance between the cameras.

Both, intrinsic and extrinsic camera parameters along with the camera height and the pitch angle, are computed with the help of a supervised calibration process, by using classic chessboard patterns of different sizes. Pitch angle is dynamically estimated at every frame, thanks to the geometric information of the environment. A good and adaptive estimation of the pitch is mandatory for a correct road/objects points separation. Generic obstacles are selected by using a *Subtractive Clustering* attention mechanism which has been adapted to the nature of the 3D data provided by the stereo reconstruction procedure. Accordingly, the amount of candidates per frame, in average, and their variability are strongly decreased, reducing the complexity of the later learning tasks.

Several databases containing thousands of pedestrian and non-pedestrian samples extracted from real traffic images have been created for learning purposes. A SVM-based pedestrian classifier has been designed according to different parameters which have been comprehensively analysed by using ROC curves: optimal kernel selection, holistic/by components performance comparison, second stage integration, performance analysis of different feature extraction methods, the suitability of using separated models depending on the illumination conditions and the candidate size, combination of optimal features for the different components, and finally, study of the effect of bounding box accuracy. Once the best single frame classification strategy has been fixed according to the last statements, the tracking, data association and multiframe validation stages are defined based on a linear Kalman filter, Mahalanobis distance along with 2D correlation, and probabilistic data association theory, respectively.

Finally, the global system has been implemented on different hardware platforms and different commercial vehicles, and tested on private circuits where several experiments have been carried out for collision avoidance and collision mitigation applications. In addition, the system has been exhibited for live demonstrations purposes in several international events, yielding promising results.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The analysis of the statistics of road traffic accidents becomes almost mandatory when designing ITS applications. Each year, thousands of pedestrians and cyclists are struck by motor vehicles. Only in the European Union about 8.000 pedestrians and cyclists are killed and about 300.000 injured. In North America approximately 5.000 pedestrians are killed and 85.000 injured. In Japan approximately 3.300 pedestrians and cyclists are killed and 27.000 injured [UNECE]. Most of these accidents take place in urban areas where serious or fatal injuries can be sustained at relatively low speed.

To reduce these figures, the European Commission is forcing the automotive industry to introduce safety measures to drastically cut the number of fatalities by 50% by 2010 compared to the figures of 2001. While in the first phase (up to 2005) passive measures were introduced to achieve the requirements, by modifying the frontal structures of vehicles, the way to reach the final requirements in the second phase (from 2005 to 2010) seems to demand the introduction of active or preventive safety measures.

The range of active safety measures is quite wide [Meinecke 05] including ideas like active hood systems, outside airbags, active bumpers or automatic deceleration [Meinecke 03]. Since these actuators have to be activated just before the crash occurs, sensors such as radar and cameras have compulsorily to be used in order to provide a measure of the time-to-collision well in advance. The study of the cumulative frequency of crashes between vehicles and pedestrians [UNECE] shows that a crash speed of up to 40 km/h can cover more than 75% of total pedestrian injuries. Thus, if a speed of up to 40 km/h is considered, the levels of injury suffered by pedestrians involved in frontal impacts with motor vehicles will be significantly reduced. Furthermore, some accidents are likely to be avoided, for velocities well bellow 40 km/h, if pedestrians are detected by the sensors onboard the car with enough an-

ticipation. In such cases, the deployment of deceleration strategies makes sense not only for collision mitigation but also for collision avoidance.

One of the most popular measures for collision mitigation is the use of the so-called active hood system [Siemens 07] [Autoliv 07]. These types of systems raise the hood of the vehicle in case of an unavoidable crash. This way a more elastic deformation of the hood can be achieved to get a reduced force over the pedestrian, especially over the head. Most of the pedestrians involved in a car-to-pedestrian accident have the first contact with the car's frontal region. This usually means that the legs make contact with the front bumper and after 50 to 150 ms the body, and especially the head, hit the bonnet or the windscreen of the car as stated in [Fuerstenberg 05]. For adult leg injuries, the major source is the front bumper of vehicles. When an adult pedestrian is struck by a vehicle, the first impact is generally between the pedestrian knee region and the vehicle's front bumper. Because this initial contact is below the pedestrian's centre of gravity, the upper body begins to rotate toward the vehicle. The pedestrians body accelerates linearly relative to the ground because the pedestrian is being carried along by the vehicle. The second contact is between the upper part of the grille or front edge of the bonnet and the pedestrian's pelvic area. The final phase of the collision involves the head and thorax striking the vehicle with a linear velocity approaching that of the initial striking velocity of the vehicle. Research has shown that the linear head impact velocity is about 90 percent of the initial contact velocity [UNECE]. Child and adult heads and adults legs are the body regions to be most affected by contact with the front end of vehicles. On vehicles, the bonnet top and the windscreen are the vehicle regions mostly identified with a high potential for contact in a car-to-pedestrian accident. These areas can cover more than 65 per cent of the fatal and serious injuries. Others measures are taken by using the so-called pedestrian protection airbags placed just on these areas. Sensor systems onboard the car are mandatorily required for predicting the car-to-pedestrian distance and the time-to-collision, both for collision avoidance and for collision mitigation.

1.2 State of the art

The most successful human detection systems from a moving vehicle are being accomplished through computer vision as main sensor. Using the same sensor humans use for driving is not a triviality. It provides the main clues for pedestrian detection although other sensors, such as laser-scanners, radar, etc., have also been tested [Carrea 00], [Gavrila 01], [Ewald 00], [Fuerstenberg 02], [Premebida 06]. So far, several approaches have been proposed, but only a few have achieved the challenge of a "global design": candidate selection, classification and tracking. These ones are the main pieces of the vision-based pedestrian detection systems installed onboard of intelligent vehicles.

Vision-based pedestrian detection is a challenging problem in real traffic scenarios

since pedestrian detection must perform robustly under variable illumination conditions, variable rotated positions and pose, and even if some of the pedestrian parts or limbs are partially occluded. An additional difficulty is given by the fact that the camera is installed on a fast-moving vehicle. As a consequence of this, the background is no longer static, and pedestrians significantly vary in scale. This makes the problem of pedestrian detection for ITS quite different from that of detecting and tracking people in the context of surveillance applications, where the cameras are fixed and the background is stationary.

To ease the pedestrian recognition task in vision-based systems, a candidate selection mechanism is normally applied. The selection of candidates can be implemented by performing an object segmentation in either a 3-D scene or a 2-D image plane. Not many authors have tackled the problem of monocular pedestrian recognition [Shashua 04], [Gavrila 99] .The advantages of the monocular solution are well known. It constitutes a cheap solution that makes mass production a viable option for car manufacturers. Monocular systems are less demanding from the computational point of view and ease the calibration maintenance process. On the contrary, the main problem with candidate selection mechanisms in monocular systems is that, on average, they are bound to yield a large amount of candidates per frame in order to ensure a low false negative ratio (i.e., the number of pedestrians that are not selected by the attention mechanism). Another problem in monocular systems is the fact that depth cues are lost unless some constraints are applied, such as the flat and static terrain assumption, which is not always applicable. These problems can be overcome by using stereo vision systems, although other problems arise such as the need to maintain calibration and the high computational cost required to implement dense algorithms.

Tha flat and stationary terrain assumption is not applicable specially in urban areas where vehicles are exposed to changes in their pitch angle due to braking, accelarating, bumped pedestrian and pelican crossings, etc. Thus pitch estimation becomes compulsory for a robust object detection algorithm. Non-flat and specifically, non static terrain assumption was introduced by non-flat road approximations by series of planar surface sections over the v-disparity map [Grubb 04], [Labayrade 02]. In [Nedevschi 04] road vertical profile is modelled with a clothoid curve fitting directly on the detected 3D road surface points.

Among the frameworks that use stereo vision for candidate selection we emphasize the next ones. In [Zhao 00] a stereo vision system to generate 3D representation of the scene with disparity maps, is propounded. The candidates are classified as pedestrian or non-pedestrian using a trained neural network. Since this is the first stereo approach in the literature, the segmentation algorithms are very basic. In [Gavrila 04] an obstacle detection procedure is done by using a multiplexed depth map, and selecting regions of interest whose number of depth features exceeds a percentage of the window area. Then they extract edge images and match them to a set of learned examples using chamfer distance [Gavrila 99]. In order to extract information from 3D scene in [Grubb 04], [Bertozzi 05a] and [Bertozzi 07] a segmen-

tation based on v-disparity and u-disparity maps is performed. The information for performing generic obstacles detection is defined with vertical lines. This implies managing very little information to detect obstacles, which may work well for big objects detection, such as vehicles [Labayrade 02], but might not be enough for small, thin objects detection, such as pedestrian, especially in city traffic due to the heavy disparity clutter.

Several systems have been presented for pedestrian detection using infrared images [Nanda 02], [Bertozzi 03], [Bertozzi 04a], [Xu 05] and infrared stereo [Tsuji 02], [Liu 04], [Bertozzi 05a], [Bertozzi 07]. Nighttime detection is usually carried out using infrared cameras as long as they provide better visibility at night and under adverse weather conditions. However, the use of infrared cameras is quite an expensive option that makes mass production an untraceable problem nowadays, especially for the case of stereo vision systems where two cameras are needed. They provide images that strongly depend on both weather conditions and the season of the year. Additionally, infrared cameras (considered as a monocular system) do not provide depth information and need periodic recalibration (normally once a year). In principle, the algorithm described in this thesis has been tested using cameras in the visible spectrum. Nonetheless, as soon as the technology for night-vision camera production becomes cheaper, the results could easily be extended to a stereo nightvision system. In Table 1.1 a summary of the main pedestrian detection systems using infrared images is presented.

Table 1.1: Outline of the main IR vision-based pedestrian detection systems.

Ref.	Sensor	Candidates	Classification	Features	Train./Test	Tracking	DR y FPR [1]
[Tsuji 02]	stereo, FIR	adaptive thresholding	–	–	–	–	–
[Nanda 02]	monocular, FIR	thresholding, image scan	Bayessian	gray levels	Tr: 1.000 Ts: 60sec	–	75-90% dr 1 fp/frame
[Bertozzi 03] [Broggi 04] [Bertozzi 04a] [Bertozzi 05a] [Bertozzi 05b] [Bertozzi 07]	monocular, stereo, FIR	vertical edges symmetry, v-disparity map, snakes	morphological models	binary image	–	–	70-85% dr 7-10% fpr
[Liu 04] [Xu 05]	stereo, monocular, FIR	thresholding, equalization, average disparity, blob matching	Holistic, SVM	binary image, gray levels	–	Kalman	93% dr 5% fpr
[Mahlisch 05]	monocular, FIR	HPN networks	templates matching, cascade classfiers	–	–	Particle filter	97% dr 0.1 fp/frame
[Suard 06]	monocular, FIR	thresholding	Holistic, SVM	HOG	Tr: 1.000 Ts: 2.200 p 2.200 n	–	90% dr 0.03 fp/frame

[1] DR: detection rate. FPR: false positive rate)

Concerning the various approaches proposed in the literature, most of them are based on shape analysis. Some authors use feature-based techniques, such as recognition by vertical linear features, symmetry, and human templates [Broggi 00], [Bertozzi 03], Haar wavelet representation [Oren 97], [Papageorgiou 00], [Mohan 01], hierarchical shape templates on Chamfer distance [Gavrila 99], [Gavrila 04], correlation with probabilistic human templates [Nanda 02], sparse Gabor filters and support vector machines (SVMs) [Cheng 05], graph kernels [Suard 05], motion analysis [Franke 02],

[Curio 00], and principal component analysis [Franke 98]. Neuralnetwork- based classifiers [Zhao 00] and convolutional neural networks [Szarvas 05] are also considered by some authors. In [Xu 05], an interesting discussion is presented about the use of binary or gray-level images as well as the use of the so-called hotspots in infrared images versus the use of the whole candidate region containing both the human body and the road. Using single or multiple classifiers is another topic of study. As experimentally demonstrated in this thesis and supported by other authors [Shashua 04], [Xu 05], [Grubb 04], the option of multiple classifiers is definitely needed. Another crucial factor, which is not well documented in the literature, is the effect of pedestrian bounding box accuracy. Candidate selection mechanisms tend to produce pedestrian candidates that are not exactly similar to the pedestrian examples that were used for training in the sense that online candidates extracted by the attention mechanism may contain some part of the ground or may cut the pedestriansŠąÇ feet, arms, or heads. This results in significant differences between candidates and examples. As a consequence, a decrease in Detection Rate (DR) takes place. The use of multiple classifiers can also provide a means to cope with day and nighttime scenes, variable pose, and non entire pedestrians (when they are very close to the cameras). In sum, a single classifier cannot be expected to robustly deal with the whole classification problem.

In the last years, SVMs have been widely used by many researchers [Shashua 04], [Papageorgiou 00], [Mohan 01], [Grubb 04], [Suard 06] as they provide a supervised learning approach for object recognition as well as a separation between two classes of objects. This is particularly useful for the case of pedestrian recognition. Combinations of shape and motion are used as an alternative to improve the classifier robustness [Shashua 04], [Viola 03]. Some authors have demonstrated that the recognition of pedestrians by components is more effective than the recognition of the entire body [Shashua 04], [Mohan 01]. In Table 1.2 an outline of the main vision-based pedestrian detection systems in the visible spectrum is depicted.

Finally it is interesting to stress those works that are combining different sensors (multisensors) and different spectrums (multimodal). By using a laserscan and a camera in [Premebida 07] a multi-sensor pedestrian and vehicle detection system is presented. Other works [Krotosky 06], [Krotosky 07] are motivating a new and very interesting discussion of feature fusion techniques, including a cross-spectral stereo solution to the pedestrian detection problem.

In our approach, the basic components of pedestrians are first located in the image and then combined with an SVM-based classifier. The pedestrian searching space is reduced in an intelligent manner to increase the performance of the detection module. Accordingly, road lane markings are detected and used as the main guidelines that drive the pedestrian searching process. The area contained by the limits of the lanes determines the zone of the real 3-D scene from which pedestrians are searched. In the case where no lane markings are detected, a basic area of interest is used instead of covering the front part ahead of the ego-vehicle. A description of the lane marking detection system is provided in [Hernández 05]. The authors have also

Table 1.2: Outline of the main vision-based pedestrian detection systems in the visible spectrum.

Ref.	Sensor	Candidates	Classification	Features	Train./Test	Tracking	DR y FPR [1]
[Broggi 00] [Bertozzi 02] [Bertozzi 04b] [Broggi 03]	mono, stereo, visible	vertical edges symmetry, v-disparity maps	ACO, vertical edges entropy	–	–	Kalman	
[Zhao 00]	mono, stereo, visible	disparity image segmentation	neural networks	gradient magnitude	Tr: 1.012 p 4.306 n Ts: 8.400	–	85.2% dr 3.1% fpr
[Gavrila 99] [Franke 00] [Gavrila 00] [Gavrila 04] [Gavrila 07]	stereo visible	XOZ map features segmentation	neural networks, templates matching	gray level image	Ts: 694 p 17.067 n	α-β *tracker*	78-100% dr 0.3-3.5 fp/min
[Grubb 04]	stereo visible	v-disparity maps	SVM, holistic, multiple.	Sobel	Tr: 1.500 p 20.000 n Ts: 150 p 2.000 n	Kalman, Bayessian probability.	83.5% dr 0.4% fpr
[Oren 97] [Papageorgiou 00] [Mohan 01]	mono visible	image scanning	SVM, holistic, by components	*Haar wavelets*	Tr: 889 p 3.106 n Ts: 123 p 796.904 n	–	97% dr $<$ 0.1% fpr
[Shashua 04]	mono visible	textures, perspective	SVM, by components, multiple	SIFT	Tr: 54.282 (50% p/n) Ts: 150 p 2.000 n	periodicity gait, ego motion, paralax, classifier consistency	(i) 96% dr 1 fp/hour (ii) 93% dr 3 fp/hour (iii) 85% dr 102 fp/hour
[Dalal 05]	mono visible	–	SVM, holistic,	HOG, *Haar wavelets*, PCA	Tr: 1.208 p 1.218 n Ts: 566 p 453 n	–	90% dr 10^{-4} fp
[Munder 06]	mono visible	–	SVM, neural networks, k-NN	PCA, *Haar wavelets*, LFR	Tr: 4.800 p 5.000 n Ts: 4.800 p 5.000 n	–	90% dr 5% fpr

[1] DR: detection rate. FPR: false positive rate

developed lane tracking systems for unmarked roads [Sotelo 04a], [Sotelo 04b] in the past. Nonetheless, a key problem is to find out the most discriminating features in order to significantly represent pedestrians. For this purpose, several feature extraction methods have been implemented, compared, and combined. While a large amount of effort in the literature is dedicated to developing more powerful learning machines, the choice of the most appropriate features for pedestrian characterization remains a challenging problem nowadays to such an extent that it is still uncertain how the human brain performs pedestrian recognition using visual information. An extensive study of feature extraction methods is therefore a worthwhile topic for a more comprehensive approach to image understanding.

This paper is organized as follows: Section II provides a description of the candidate selection mechanism along with the pitch angle compensation procedure. Section III describes the component-based approach, the optimal combination of feature extraction methods, the SVM-based pedestrian classification system and the multiframe validation and tracking processes. The implementation and comparative results achieved to date are presented and discussed in Section IV. Finally, Section V summarizes the conclusions and future works.

Chapter 2

Stereo vision-based candidate selection mechanism

2.1 Camera pitch and height calibration

The proposed system for calibrating the pitch angle and the height of the camera with regard to the ground plane is based on the use of a classical chessboard pattern like the one depicted in Figure 2.1. A set of images are taken with the pattern laying on the ground in different poses (see Figure 2.2). By using the Camera Calibration Toolbox for Matlab [Matlab. 07] extrinsic parameters for each pattern pose are obtained. In Figures 2.3(a) and 2.3(b) two views of the different extrinsic parameters are depicted.

Figure 2.1: Chessboard pattern of 4×5 squares used for calibrating camera pitch and height with respect to the ground. Each square has a size of $400mm$.

Let T_i and R_i represent the translation vector and the rotation matrix of the pattern with respect to the left camera on image i. Then, as stated in [Matlab. 07], the third column of the rotation matrix r_i^3 expresses the normal vector of the pattern surface which is denoted as $\vec{n}_i$. As it is shown in Figures 2.4(a) and 2.4(b) the pitch angle can be computing as from the relation between the normalized vector $\vec{o} = (0, -1, 0)$ (note that the Y axis used by the Toolbox of Matlab points to the bottom) and the

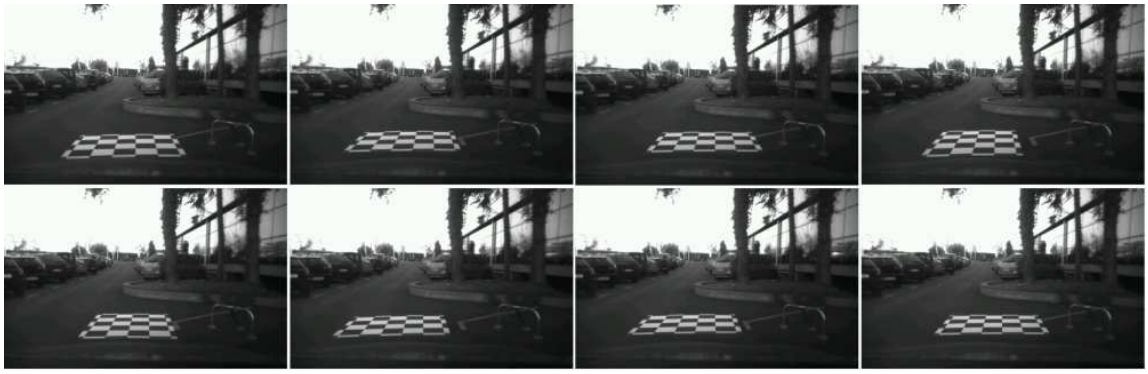
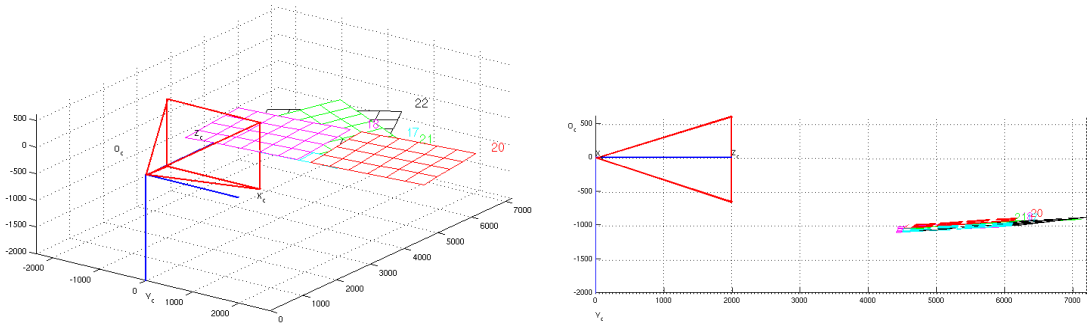


Figure 2.2: Set of images used for pitch and height of the camera calibration process.



(a) (b)

Figure 2.3: 3D pose of the pattern laying on the ground in different images with regard to the left camera point of view. (a) Transverse view; (b) Lateral vista.

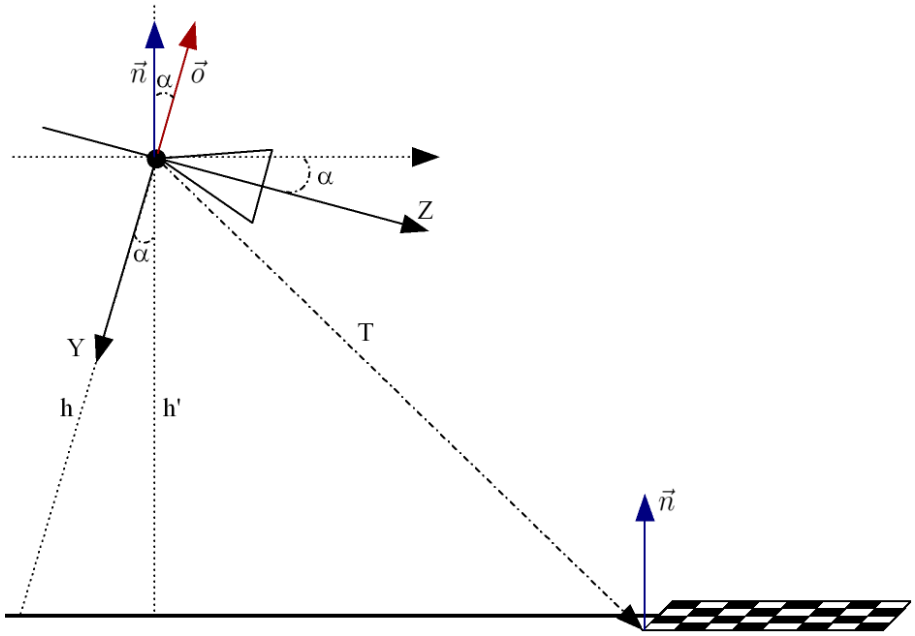
normal vector of the pattern surface $\vec{n}_i$. Taking into account that both vectors are normalized, the dot product is used for computing the pitch angle as follows:

$$\vec{o}.\vec{n_i} = |\vec{o}||\vec{n}_i| \cos\alpha_i = \cos\alpha_i \Rightarrow \alpha_i = cos^{-1}(\vec{o}.\vec{n_i}) \tag{2.1}$$

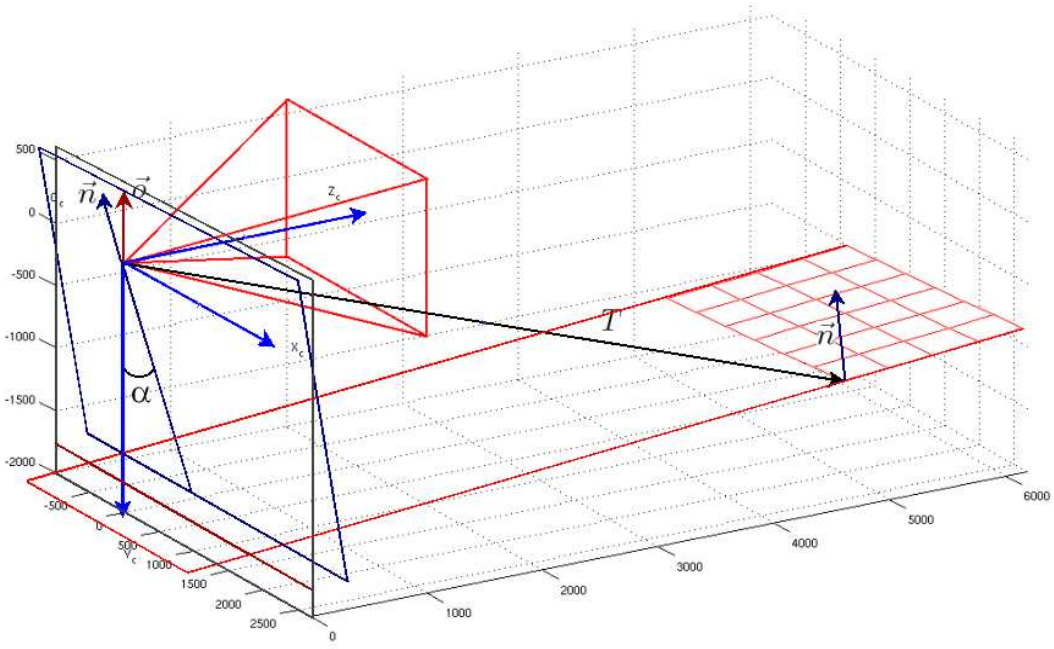
As a set of N images (like the ones depicted in Figure 2.2) is available, it is possible to use the average for each image to obtain a steadier solution:

$$\alpha = \frac{1}{N}\sum_{i=1}^{N}\alpha_i = \frac{1}{N}\sum_{i=1}^{N}\left(cos^{-1}(\vec{o}.\vec{n_i})\right) \tag{2.2}$$

The proposed method for computing the camera height with regard to the road is defined as follows. Let $T_i = (t_{xi}, t_{yi}, t_{zi})$ represents the translation vector and $\vec{n}_i = (a_i, b_i, c_i)$ represents the normal vector of the chessboard pattern surface. The normal plane that has the normal vector $\vec{n}_i$ and goes through the 3D point T_i is defined by applying the following expression:



(a)



(b)

Figure 2.4: Global views of the proposed method for calibrating the pitch angle and the height of the camera. (a) Lateral view with the ground plane in horizontal position (b) 3D view with the camera pitch angle in horizontal position.

$$a_i(x - t_{xi}) + b_i(y - t_{yi}) + c_i(z - t_{zi}) = 0 \Rightarrow a_i x + b_i y + c_i z = a_i t_{xi} + b_i t_{yi} + c_i t_{zi} \quad (2.3)$$

Image plane is defined by the expression $z = 0$. The line of intersection of the plane $z = 0$ and the plane defined in (2.3) is defined by:

$$a_i x + b_i y = a_i t_{xi} + b_i t_{yi} + c_i t_{zi} \Rightarrow a_i x + b_i y = d_i \tag{2.4}$$

where $d_i = a_i t_{xi} + b_i t_{yi} + c_i t_{zi}$. The distance between the line of intersection of both planes and the optical center $p = (0, 0, 0)$ is given by:

$$h_i = \frac{d_i}{\sqrt{a_i^2 + b_i^2}} = \frac{a_i t_{xi} + b_i t_{yi} + c_i t_{zi}}{\sqrt{a_i^2 + b_i^2}} \tag{2.5}$$

The actual height of the camera h' with respect to the ground plane is given as from pitch estimation, as can be observed in Figure 2.4(a):

$$h'_i = h_i \cos \alpha_i \tag{2.6}$$

The final value of the height of the camera is obtained by computing the average value of h'_i for each one of the N images as follows:

$$h' = \frac{1}{N} \sum_{i=1}^{N} h'_i = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{d_i}{\sqrt{a_i^2 + b_i^2}} \cos \alpha_i \right) \tag{2.7}$$

The initial estimation of the camera pitch and the camera height, obtained after applying the proposed method, will be very useful in next stages, not only as initial values but as default ones, when there will not be enough information to perform pitch estimation robustly.

2.2 Non dense reconstruction

2.2.1 Non-dense features selection

3D reconstruction can be achieved in a dense way, by using all the image points [Zhao 00], [Grubb 04], or in a non-dense way by taking into account only specific image points [Labayrade 02], [Franke 00]. Even the fact that there are several frameworks that implement dense reconstruction in real time [Scharstein 02], [Brown 03], [van der Mark 06] our approach does not need dense information in order to select candidates. Thus the computational cost is drastically reduced.

In our case it is necessary to achieve a trade-off between the number of features and their quality. Features as Harris corners [Harris 88] or KLT [Lucas 81], [Shi 94]

provide high quality points but these ones are not enough for the proposed method. The output of a standard Harris corners detector in a pedestrian sample image is depicted in Figure 2.5(b). The maximum number of points detected with an acceptable noise level is clearly insufficient. According to this the use of the well known Canny edge detector [Canny 86] is proposed since it maximizes the signal-to-noise ratio. As can be observed in Figure 2.5(c) Canny image provides a good representation of the discriminant features for pedestrians. Features as head, arms and legs are distinguishable when visible, and are not heavily affected by different colors or clothes.

(a) (b) (c)

Figure 2.5: (a) Pedestrian sample image; (b) Harris features; (c) Canny image.

The Canny hysteresis thresholds define the final behaviour of the filter. Any pixel in the image that has a value greater than th_L is presumed to be an edge pixel, and is marked as such immediately. Then, any pixels that are connected to this edge pixel and that have a value greater than th_H are also selected as edge pixels. The contrast in daytime images is always greater than the contrast in nightime ones. The mean and the variance of the gradient image histogram is clearly different between daytime and nightime images [1]. Accordingly the use of adaptive hysteresis thresholds is proposed in this work. These thresholds are defined by means of the analysis of the histogram of the magnitude of the gradient image. Let $\bar{x}$ be the mean value and σ the variance of the histogram of the gradient magnitude, i.e.:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i w_i}{\sum_{i=1}^{N} w_i} \tag{2.8}$$

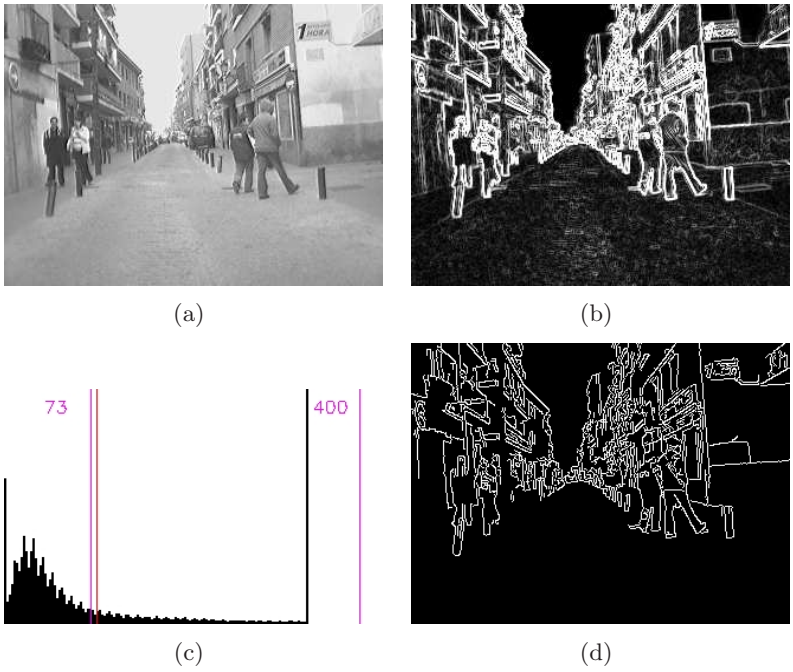
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2 w_i}{\sum_{i=1}^{N} w_i}} \tag{2.9}$$

where w_i is the number of pixels whose gradient magnitude is equal to x_i, and N is the maximum gradient magnitude value. After extensive trials in different lighting conditions (cloudless, cloudy, foggy, rainy, etc.), next adaptive threshold values are defined:

[1] By nightime we mean well illuminated urban environments

$$th_L = \bar{x} - \frac{\sigma}{16} \quad ; \quad th_H = \bar{x} + 4\sigma \tag{2.10}$$

In Figures 2.6 and 2.7 two examples in daytime and nightime conditions respectively are depicted. The adaptive value of the hysteresis thresholds can be observed in Figures 2.6(c) and 2.7(c).



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

Figure 2.6: Daytime scenario. (a) Original image ;(b) Gradient image; (c) Histogram of the gradient magnitude and adaptive thresholds ; (d) Final result after applying Canny edge detector.

2.2.2 Robust correspondence search

Thanks to the calibration process it is possible to use the epipolar geometry since fundamental matrix is known. Thus, the matching searching area is greatly decreased by using the parameters of the fundamental matrix F[Trucco 98], [Forsyth 03], [Hartley 03]. For each point on the left image $m_l = (u_l, v_l)$ the corresponding right epipolar line $a'_r u_r + b'_r v_r + c'_r = 0$ can be computed as follows:

$$\begin{pmatrix} a'_r \\ b'_r \\ c'_r \end{pmatrix} = F \begin{pmatrix} u_l \\ v_l \\ 1 \end{pmatrix} \tag{2.11}$$

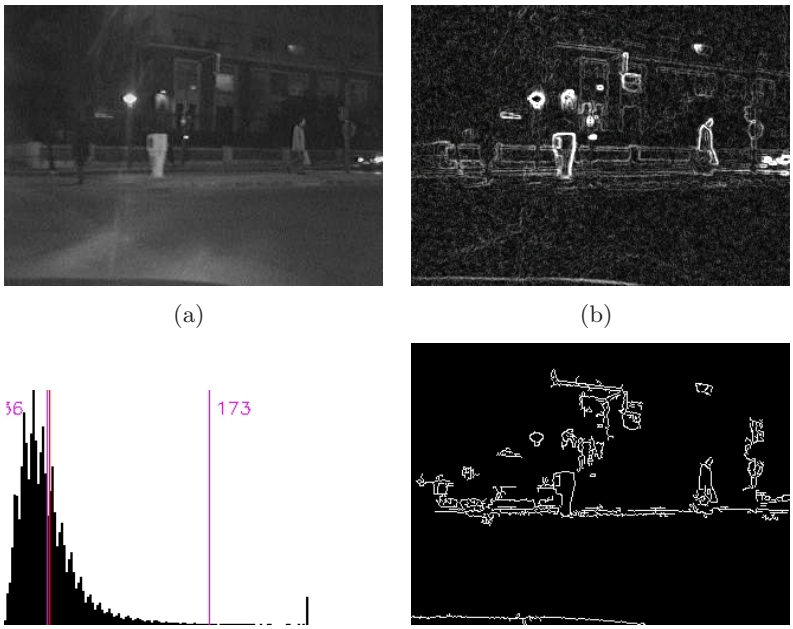


Figure 2.7: Nightime scenario. (a) Original image ;(b) Gradient image; (c) Histogram of the gradient magnitude and adaptive thresholds ; (d) Final result after applying Canny edge detector.

The use of parallel epipolar lines has been widely applied in obstacle detection applications [Se 98], [Yu 03], [Labayrade 02], [Zhao 00] and [Franke 00]. Nevertheless in practice epipolar lines are not parallel, even if the stereo structure is very precise. It is possible to manage parallel epipolar lines after applying images rectification [Fusiello 97], [Forsyth 03], [Trucco 98]. Several frameworks propose to rectify the stereo pair in order to simplify the correspondence search [Grubb 04], [van der Mark 06] y [Miled 07]. While image rectification provides a simple search area for correspondents and straightforward 3D reconstruction, the general geometry mode, without rectification, provides a better resolution since no image re-sampling is done [Nedevschi 04], [Nedevschi 06]. Moreover, epipolar geometry is precomputed when the application starts, and stored in a look-up table, so that, epipolar lines computation is avoided at run-time an thus the computational cost is further reduced.

Disparity search space is analysed by means of a specific software designed with that purpose. For each left image point $m_l = (u_l, v_l)$ the x disparity search space associated with the right images goes between $dx_{min} = u_l - 20$ to $dx_{max} = u_l + 25$, whose associate depths are $Z_{min} = 2m$ and $Z_{max} = 30m$ respectively. In consequence the x disparity search space has a total amount of 46 pixels which should be analyzed.

Most of the ITS frameworks use simple matching techniques such as SAD (sum of absolute differences) or SSD (sum of squared differences) [Grubb 04], [Nedevschi 04],

[van der Mark 06]. The main problem of that kind of matching techniques is that they need to apply a preprocessing step in order to compensate for intensity differences between the stereo pair of images, such as LoG [Bunschoten 00]. For avoiding the use of intensity differences compensation, and among the wide spectrum of matching techniques that can be used to solve the correspondence problem, we implemented the Zero mean Normalized Cross Correlation (ZNCC) because of its robustness against illumination level changes [Faugeras 93], [Boufama 94], [Krotkov 95], [Sun 02]. The ZNCC between points $p = (u, v)$ and $p' = (u', v')$ from two different images is defined by next expression:

$$ZNCC(p, p') = \frac{\sum_{i=-n}^{n} \sum_{j=-n}^{n} A \cdot B}{\sqrt{\sum_{i=-n}^{n} \sum_{j=-n}^{n} A^2 \sum_{i=-n}^{n} \sum_{j=-n}^{n} B^2}} \tag{2.12}$$

where A and B are defined by:

$$A = (I(u + i, v + j) - \overline{I(u, v)}) \tag{2.13}$$

$$B = (I(u' + i, v' + j) - \overline{I(u', v')}) \tag{2.14}$$

where $I(u, v)$ is the intensity level of pixel with coordinates (u, v), and $\overline{I(u, v)}$ is the average intensity of a $(2n + 1) \times (2n + 1)$ window centered around that point. As the window size decreases, the discriminatory power of the area-based criterion is decreased, and some local maximum appear in the searching regions (epipolar lines). On the contrary, an increase in the window size causes the performance to degrade due to occlusion regions and smoothing of disparity values across boundaries [van der Mark 06]. Accordingly to previous statements and to a computational cost criterion, a practical 7×7 correlation window size is selected.

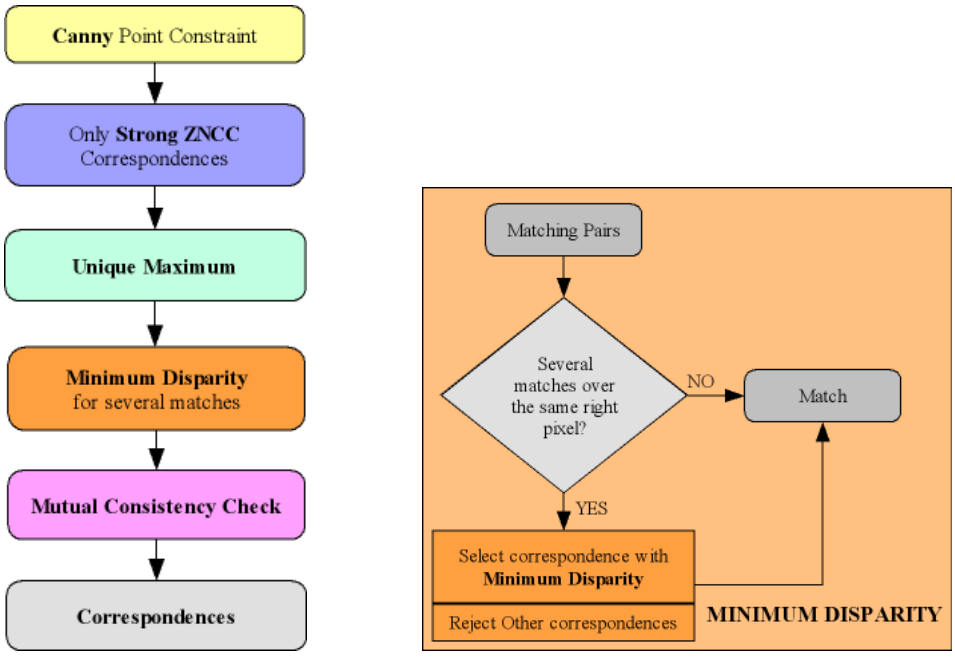
A post-processing is applied in the correlation step in order to increase robustness and reduce noise:

- Only strong responses of the correlation function along the epipolar line are considered as correspondents, so that, only responses grater than $ZNCC_{min}$ are taken into account, where $ZNCC_{min} = 0.9$.
- If the global maximum of the function is not strong enough relative to other local maximum, then the current left image point is rejected (*unique maximum*). In [Fusiello 00] the variance of the correlation values is used as a reliability

measure. In our case we propose to use the two global maximum values of the correlation function as in [Hirschmuller 02], [Muhlmann 02]. Let $C_1(u,v)$ represents the global maximum and $C_2(u,v)$ represents the next global maximum, the reliability coefficient is defined as:

$$W(u,v) = 1 - \frac{C_2(u,v)}{C_1(u,v)} \tag{2.15}$$

- Right image correlated points are also correlated over the left image (*mutual consistency check strategy*). If the new left matched points are not exactly the same than the original ones, these correspondences are considered as noise (*multi-correlation*).
- In case different left image points were correlated over the same right image point, two strategies could be taken: maximum correlation criterion [Stefano 02] or minimum disparity criterion. The second one is used since the noise due to structured backgrounds, which usually produces close 3D points, is avoided (*minimum disparity*).



(a) (b)

Figure 2.8: (a) Global scheme of the robust correspondence search method ;(b) Minimum disparity criterion scheme.

In Figure 2.8(a) the global outline of the proposed robust correlation method is depicted. Minimum disparity process when different left images points are correlated over the same right image point is depicted in Figure 2.8(b). The unique maximum

criterion is summarized in Figure 2.9. Finally the proposed method for mutual consistency check can be observed in Figure 2.10.

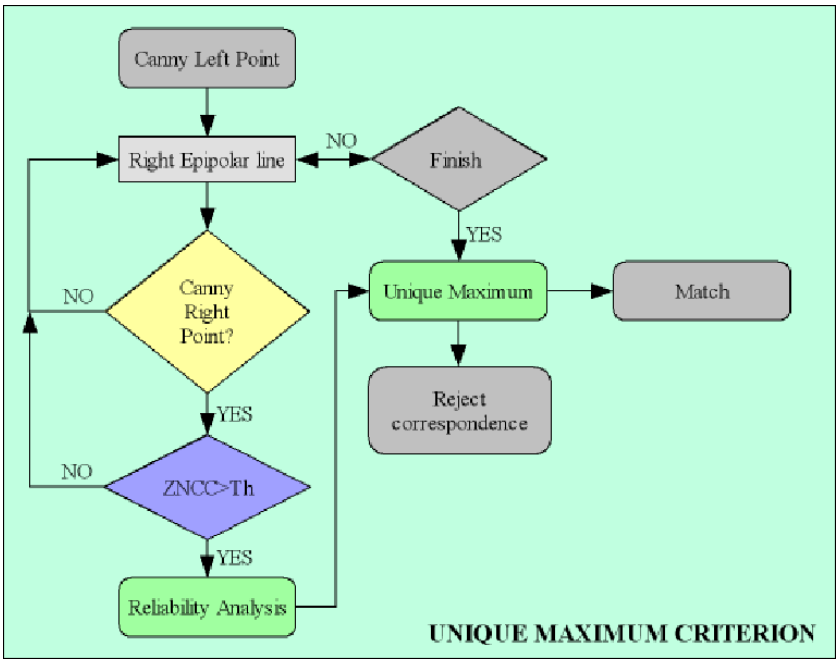


Figure 2.9: Outline of the unique maximum criterion.

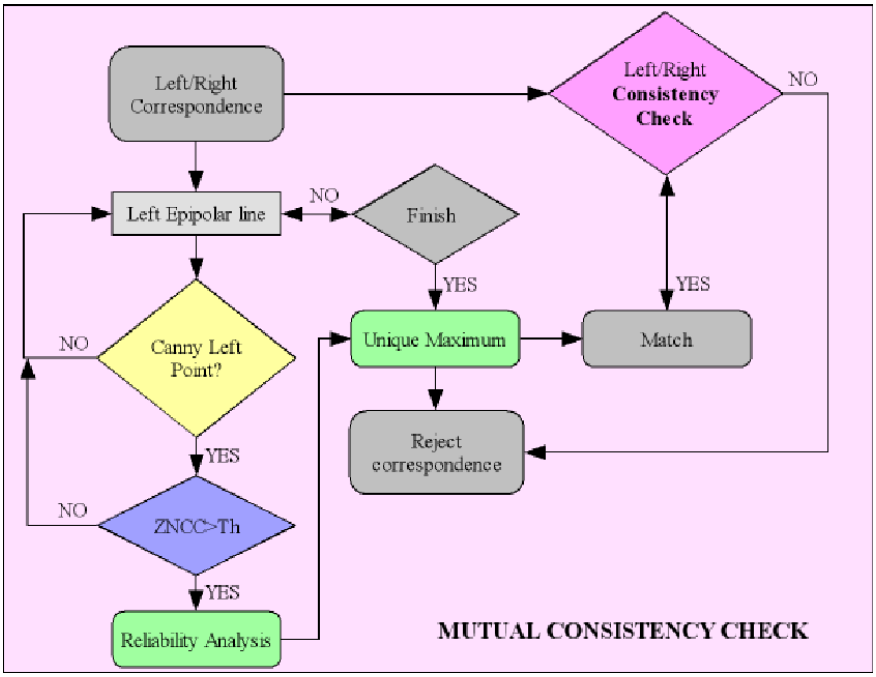


Figure 2.10: Global scheme of the mutual consistency check.

After applying the previous steps, the resulting correlation maps look much more

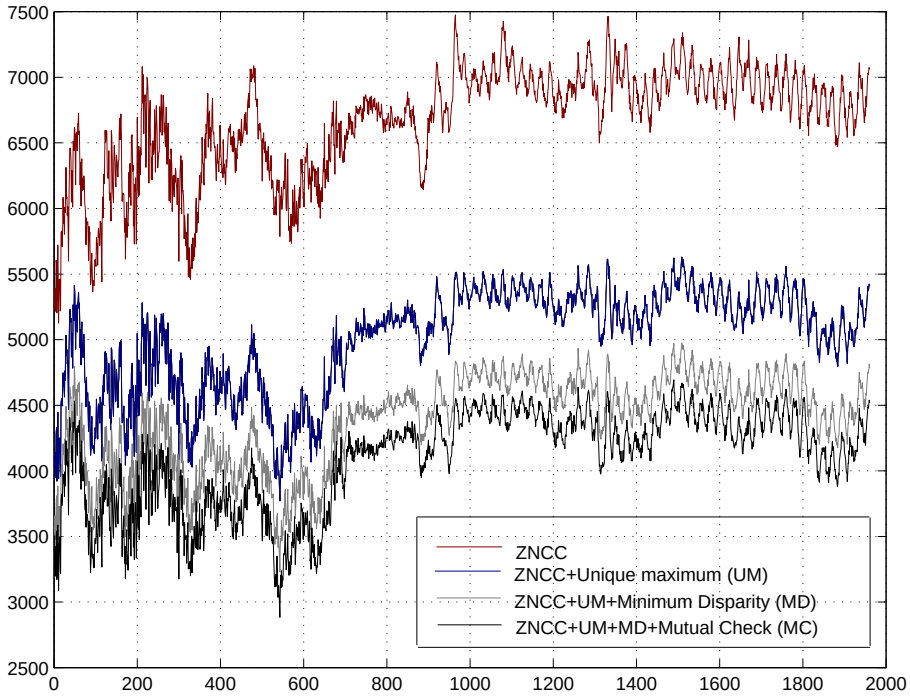


Figure 2.11: Number of correlated points as from the different steps used in the correlation process, in a sequence of 2000 frames.

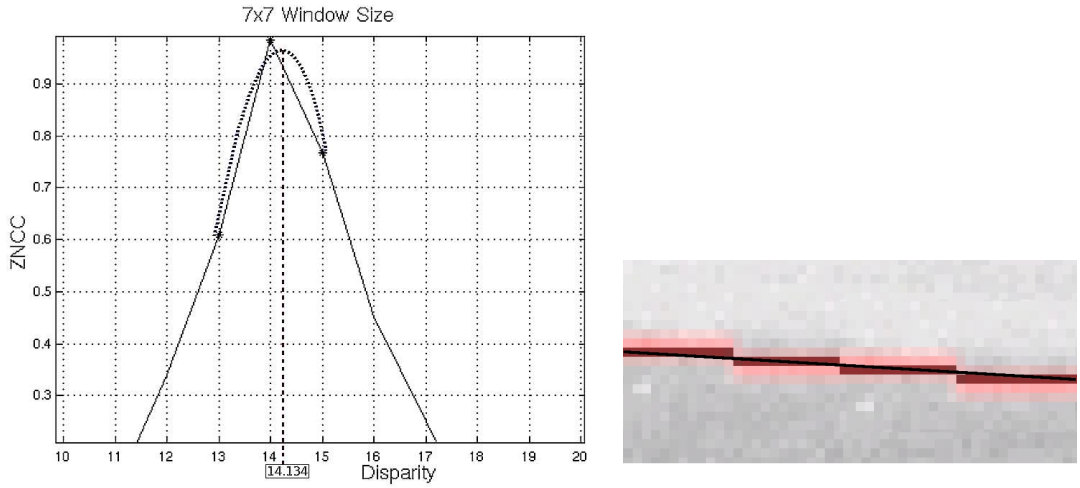
noise-free. As can be observed in Figure 2.11, the number of correlated points gets decreased by an average of 24.87 after using unique maximum criterion. By using both unique maximum and minimum disparity strategy an average of 34.14% of points are selected as noise. Finally, by adding the mutual consistency check an average of 38.54% of points are rejected. In practice most of the errors are du to structural backgrounds which very usual in urban environments. Occlusions also affect but to a lesser extent.

Besides using integer values for disparity correspondence, subpixel accuracy can be obtained by using floating point values. The subpixel value of the u coordinate of the right image can be approximated by a second degree polynomial. That kind of approach can be also used with rectified images and with different matching measures [Muhlmann 02], [Stefano 02], [Nedevschi 04] and [van der Mark 06]. Let $f(u)$ represents the ZNCC correlation value of the global maximum at pixel (u, v). Then the subpixel value in the u-axis will be computed as follows:

$$u_s = u + \frac{f(u-1) - f(u+1)}{2(f(u-1) - 2f(u) + f(u+1))} \tag{2.16}$$

where $f(u-1)$ and $f(u+1)$ are the ZNCC correlation values of the adjacent pixels of (u, v). Once u_s is computed, the subpixel value in the v-axis is then obtained by using the epipolar line equation (see Figure 2.12(b)):

$$a'_r u_s + b'_r v_s + c'_r = 0 \Rightarrow v_s = -\frac{c'_r + a'_r u_s}{b'_r} \tag{2.17}$$



(a) (b)

Figure 2.12: (a) Subpixel accuracy in the u axis by a second degree polynomial approaching. ;(b) Subpixel accuracy in the v axis as from the epipolar line equation.

2.2.3 3D Reconstruction

After solving the correspondence problem it is still necessary to solve the reconstruction problem, i.e., given a number of corresponding parts of the left and right image, and possibly information on the geometry of the stereo system, what can we say about the 3D location and structure of the observed objects? [Trucco 98]. In general these algorithms are so-called triangulations methods. The most used triangulation method is the one in which the intersection between the rays that join both optical centres with the correspondences, is computed. However, both rays will never, in practice, actually intersect, due to calibration and feature localization errors. In this context, various reasonable approaches to the reconstruction problem can be adopted, such as the method that finds the midpoint of the common perpendicular to the two rays in space [Xu 96], [Trucco 98], [Forsyth 03]. Nevertheless, this method is not suitable for projective reconstruction, since concepts such as distance and perpendicularity are not valid in the context of projective geometry. In fact, in projective reconstruction, this method will give different results depending on which particular projective reconstruction is considered, i.e., the method is not projective-invariant [Hartley 03].

In this work we propose the use of a triangulation method that is projective-invariant. As we know both intrinsic and extrinsic parameters of the stereo system, a scene point

can be reconstructed by using a purely algebraic approach. The equations that relate the projection of a 3D points onto the image plane are given by:

$$\begin{cases} u_l &= \frac{m_{11}X_l+m_{12}Y_l+m_{13}Z_l+m_{14}}{m_{31}X_l+m_{32}Y_l+m_{33}Z_l+m_{34}} \\ v_l &= \frac{m_{21}X_l+m_{22}Y_l+m_{23}Z_l+m_{24}}{m_{31}X_l+m_{32}Y_l+m_{33}Z_l+m_{34}} \end{cases} \tag{2.18}$$

$$\begin{cases} u_r &= \frac{m'_{11}X_r+m'_{12}Y_r+m'_{13}Z_r+m'_{14}}{m'_{31}X_r+m'_{32}Y_r+m'_{33}Z_r+m'_{34}} \\ v_r &= \frac{m'_{21}X_r+m'_{22}Y_r+m'_{23}Z_r+m'_{24}}{m'_{31}X_r+m'_{32}Y_r+m'_{33}Z_r+m'_{34}} \end{cases} \tag{2.19}$$

where (u_l, v_l) and (u_r, v_r) are the correspondent points in pixel coordinates from left and right cameras respectively, and $P_l = (X_l, Y_l, Z_l)$ and $P_r = (X_r, Y_r, Z_r)$ represent the 3D points with regard to the left and right camera respectively. The point P_r can be expressed as a function of P_l by means of the extrinsic relationship between both cameras by using the expression $P_r = RP_l + T$, i.e.:

$$\begin{cases} X_r &= r_{11}X_l + r_{12}Y_l + r_{13}Z_l + t_x \\ Y_r &= r_{21}X_l + r_{22}Y_l + r_{23}Z_l + t_y \\ Z_r &= r_{31}X_l + r_{32}Y_l + r_{33}Z_l + t_z \end{cases} \tag{2.20}$$

If radial and tangential distortions are compensated for, we can define a system of 4 equations in 3 unknowns (3D point coordinates) as follows:

$$\left.\begin{array}{rcl} u_l - u_{0l} &=& f_{xl}\frac{X_l}{Z_l} \\ v_l - v_{0l} &=& f_{yl}\frac{Y_l}{Z_l} \\ u_r - u_{0r} &=& f_{xr}\frac{r_{11}X_l+r_{12}Y_l+r_{13}Z_l+t_x}{r_{31}X_l+r_{32}Y_l+r_{33}Z_l+t_z} \\ v_r - v_{0r} &=& f_{yr}\frac{r_{21}X_l+r_{22}Y_l+r_{23}Z_l+t_y}{r_{31}X_l+r_{32}Y_l+r_{33}Z_l+t_z} \end{array}\right\} \Rightarrow \tag{2.21}$$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{pmatrix} \begin{pmatrix} X_l \\ Y_l \\ Z_l \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} \Rightarrow AP_l = b \tag{2.22}$$

where,

$$\begin{cases} a_{11} = f_{xl} \\ a_{12} = 0 \\ a_{13} = -(u_l - u_{0l}) \\ b_1 = 0 \\ a_{21} = 0 \\ a_{22} = f_{yl} \\ a_{23} = -(v_l - v_{0l}) \\ b_2 = 0 \\ a_{31} = r_{31}(u_r - u_{0r}) - f_{xr}r_{11} \\ a_{32} = r_{32}(u_r - u_{0r}) - f_{xr}r_{12} \\ a_{33} = r_{33}(u_r - u_{0r}) - f_{xr}r_{13} \\ b_3 = f_{xr}t_x - t_z(u_r - u_{0r}) \\ a_{41} = r_{31}(v_r - v_{0r}) - f_{yr}r_{21} \\ a_{42} = r_{32}(v_r - v_{0r}) - f_{yr}r_{22} \\ a_{43} = r_{33}(v_r - v_{0r}) - f_{yr}r_{23} \\ b_4 = f_{yr}t_y - t_z(v_r - v_{0r}) \end{cases} \tag{2.23}$$

The maximum likelihood estimate is given by the point which minimizes the reprojection error, i.e., the summed squared distances between the projections of P_l and the measured image points. Such a triangulation method is projective-invariant because only image distances are minimized, and the projections of P_l do not depend on the projective frame in which P_l is defined, i.e., a different projective reconstruction will project to the same points.

As we have an overconstrained system of four independent linear equations ($AP_l = b$) in the homogeneous coordinates of P_l, that is easily solved using the linear least-squares technique which implies to compute the pseudoinverse as follows:

$$AP_l = b \Rightarrow A^T AP_l = A^T b \Rightarrow P_l = (A^T A)^{-1} A^T b \tag{2.24}$$

where $A^\dagger = (A^T A)^{-1} A^T$ is the pseudoinverse of the matrix A which is computed by using singular value decomposition (SVD), since it is proved that the solution which minimizes the error $e = |AP_l - b|$ in the least squares sense, it is exactly the one that accomplishes with $V = 0$. After solving both correspondence and reconstruction problems, non-dense 3D maps are created like the ones depicted in Figure 2.13.

2.3 Pitch estimation

Detection range specification in vision based pedestrian detection applications is usually no longer than 30m due to several constraints like camera resolution, pedestrian size, etc. Thus, flat road geometry can be considered, i.e., road curvature can be

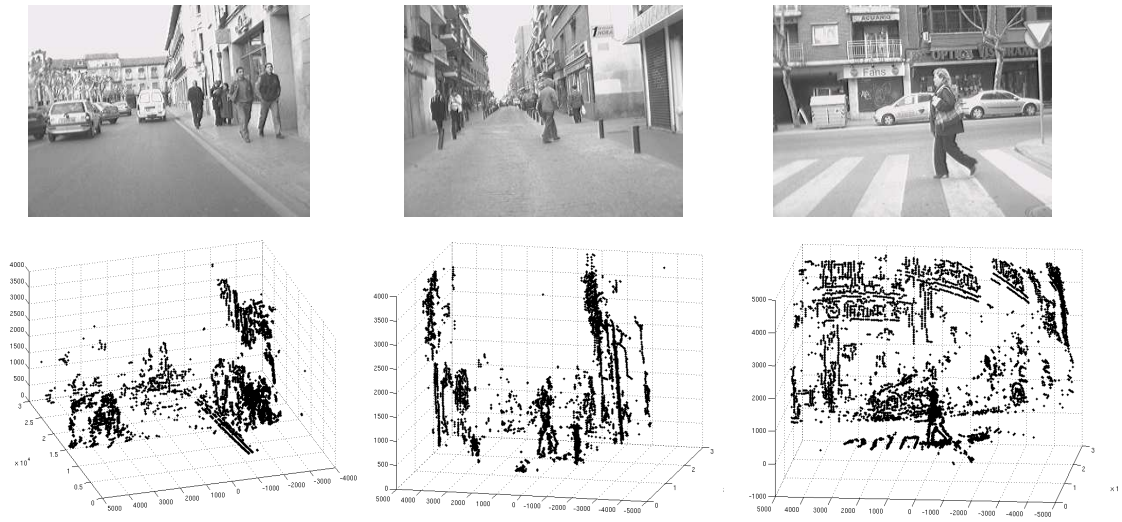


Figure 2.13: Upper row: original images. Lower row: non-dense 3D maps.

neglected in the near range. Thanks to the stereo approach the vertical road profile can be directly extracted. Two different methods have been tested in this work and explained in the following sections: pitch estimation based on projection map (YOZ), and pitch estimation using virtual disparity map.

2.3.1 Pitch estimation based on YOZ projection map

A robust correlation process reduces the number of 3D points located under the road (which is directly proportional to the amount of correlation errors). Taking into account a base plane without pitch change, the height of the camera relative to the base plane, and the camera vertical field of view, the origin of the world coordinate system is placed at the intersection point between the base plane and the lower boundary of the vertical field of view. Figure 2.14 depicts the lateral projection of 3D points on the YOZ plane.

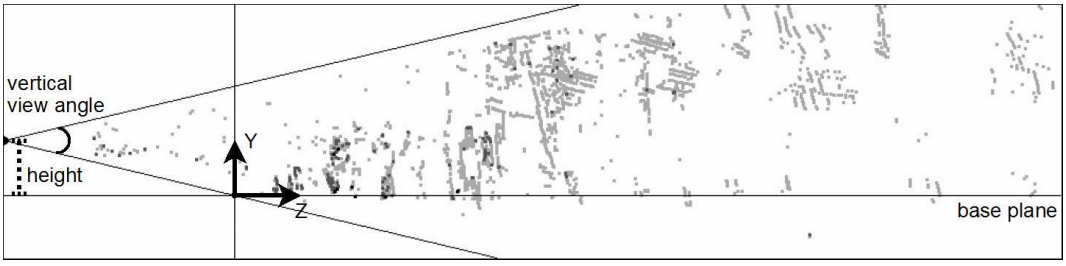


Figure 2.14: 3D projected points on the YOZ plane up to 30m.

The number of 3D projected points over the same 2D point in the lateral view are coded in a gray scale image. Thus the weight of matching errors is reduced.

As in [Nedevschi 04] we consider the vertical displacement due to roll negligible in comparison to the displacement due to pitch. From the point of view of the world coordinate system origin, and varying the slope to cover all possible pitch values, uniformly spaced rays are cast. Gray level values (number of points) along each ray i are counted in a histogram $H(i)$. The histogram is normalized and the mean value $\bar{h}$ is computed. A stable *jump* over $2/3\bar{h}$ in the histogram is looked for from the bottom of the road upwards. Being $i = 0$ the lowest ray and $i = N$ the highest one, pitch angle is selected as follows:

$$\begin{array}{l} \text{for } i = 0 \text{ to } N \\ \quad \text{if } (H(i) > \frac{2}{3}\bar{h} \text{ and } H(i+1) > \frac{2}{3}\bar{h} \text{ and } H(i+2) > \frac{2}{3}\bar{h} \\ \quad\quad \text{and } H(i), H(i+1), H(i+2) > H_{min}) \\ \quad\quad \text{then } \alpha = \alpha_i; \text{ break;} \\ \quad \text{else } \alpha = \alpha_{init}; \end{array} \tag{2.25}$$

The variable α_{init} is the pitch angle estimated after the calibration process. The parameter H_{min} is used to avoid pitch estimation errors when there are not enough road points detected. Figure 2.15 depicts three examples for positive, negative and zero pitch angle values. The darker the ray the higher the number of accumulated points. The estimated pitch angle is drawn in bold.

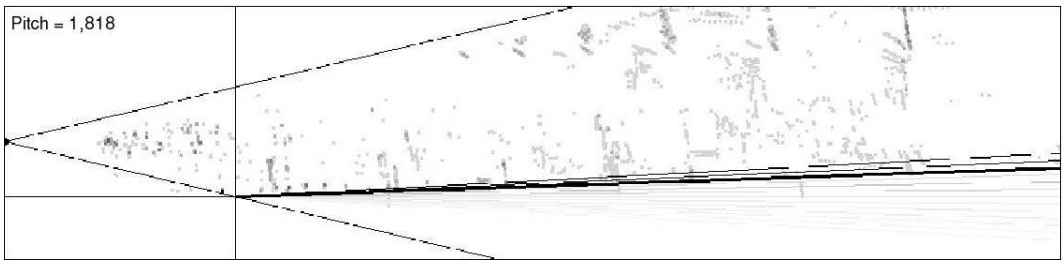
2.3.2 Pitch estimation using virtual disparity map

An improvement in pitch estimation accuracy is proposed based on the idea suggested in [Suganuma 07]. A rigid transformation of 3D points is carried out from the left camera with regard to the ground plane. For this purpose, a rotation around X axis and a translation along Y axis are performed as depicted in Figure 2.16). Parameters h (camera height) and α (camera pitch angle) obtained in the calibration stage are used in the transformation process.

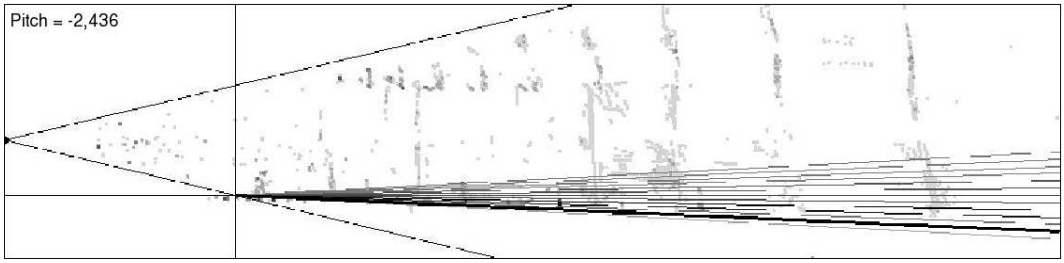
Let $P_l = (X_l, Y_l, Z_l, 1)^T$ represents the homogeneous coordinates of a 3D point with regard to the left camera. The coordinates of point $P_n = (X_n, Y_n, Z_n, 1)^T$ with regard to the ground plane can be obtained after applying the following transformation:

$$\begin{pmatrix} X_n \\ Y_n \\ Z_n \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -h \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha & 0 \\ 0 & \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X_l \\ Y_l \\ Z_l \\ 1 \end{pmatrix} \tag{2.26}$$

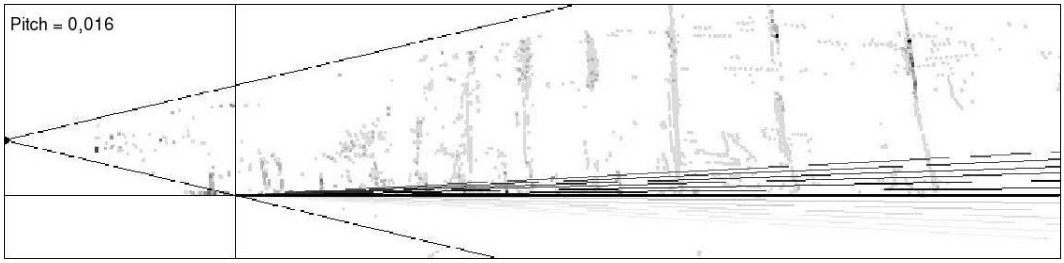
After transforming the reference frame, 3D points are backprojected on the image plane using the camera intrinsic parameters as follows:



(a)



(b)



(c)

Figure 2.15: Pitch angle estimation. (a) Positive pitch angle, (b) negative pitch angle and (c) pitch angle about 0 degrees.

$$\begin{pmatrix} su_n \\ sv_n \\ s \end{pmatrix} = \begin{pmatrix} f_{xl} & 0 & u_{0l} & 0 \\ 0 & f_{yl} & v_{0l} & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X_n \\ Y_n \\ Z_n \\ 1 \end{pmatrix} \tag{2.27}$$

Thus, the so-called *virtual disparity image* is obtained. The gray level of pixels in disparity images I_Δ is inversely proportional to their distance Z with regard to the camera, as depicted in Figure 2.17(b). As can be observed, the lesser the distance between the point and the camera, the higher the corresponding pixel gray level. After applying the transformation proposed in (2.26) and backprojecting pixels on the image plane using (2.27), the resulting virtual disparity image is shown in Figure 2.17(c).

The backprojection of all points along the ground plane in the virtual camera must

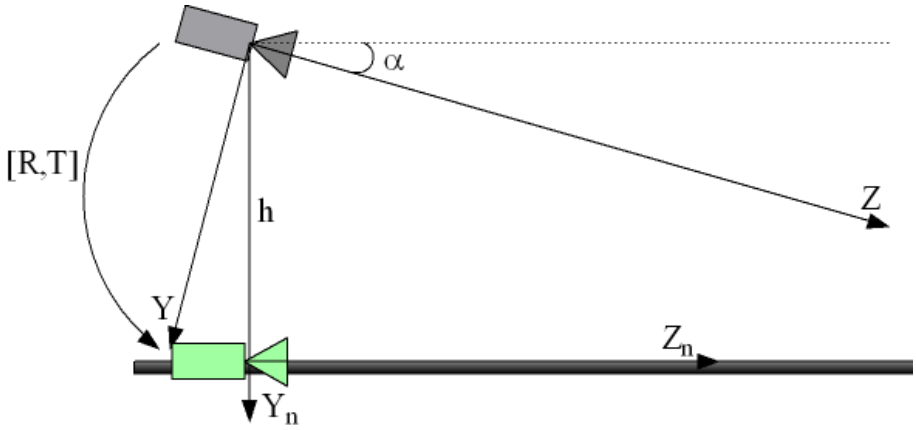


Figure 2.16: Rigid transformation - Rotation around X axis and translation along Y axis in order to virtually place the camera on the ground plane..

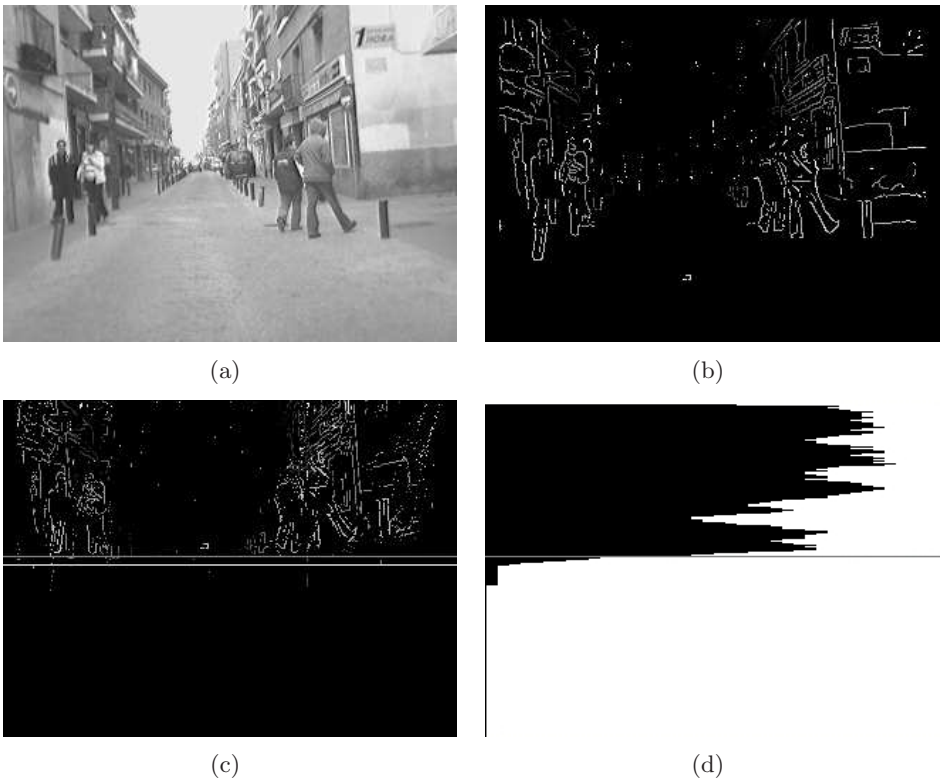


Figure 2.17: (a) Original left image; (b) Non-dense disparity image I_Δ ; (c) Virtual disparity image; (d) Vertical histogram of Virtual Disparity.

be located on the v coordinate corresponding to the camera central point v_{ol} if no variation with respect to the calibrated camera pitch angle takes place, as depicted in Figure 2.18(a). Points in the virtual camera image plane lying on coordinate v different from v_{ol} imply a variation in the pitch angle with regard to the calibrated

one. Figures 2.18(b) and 2.18(c) depict a change in the ground plane projection due to positive and negative variations of the pitch angle, respectively. In order to obtain the projection over the ground plane, a vertical histogram is computed accounting for the number of points lying on each raw in the virtual disparity image, as illustrated in Figure 2.17(d). The vertical histogram is smoothed using 3×1 window. After that, the average value $\bar{h}$ is recomputed as follows:

$$H_s(i) = \frac{1}{3}\sum_{j=-1}^{j=1} H(i-j) \quad ; \quad \bar{h} = \frac{1}{N}\sum_{i=0}^{i=N} H_s(i) \tag{2.28}$$

The proposed algorithm seeks a stable, significant jump in the vertical histogram starting from the bottom line $i =$ height up to the top line $i = 0$ implementing the following process:

$$\begin{array}{l} \text{for } i = \text{height to } 0 \\ \quad \text{if } (H_s(i) > \bar{h} \text{ and } H_s(i-1) > \bar{h} \text{ and } H_s(i-2) > \bar{h} \\ \quad\quad \text{and } H_s(i), H_s(i-1), H_s(i-2) > H_{min}) \\ \quad\quad \text{then } v_a = i\text{; break;} \\ \quad \text{else } v_a = v_{0l}\text{;} \end{array} \tag{2.29}$$

Parameter H_{min} is again used for setting a minimum number of points requested for achieving robust estimation. The computation of pitch angle with regard to the ground plane is carried out in the image plane using the camera intrinsic parameters f, f_{yl}, d_{yl} (mm/pixel in Y axis), and the camera optical centre (u_{0l}, v_{0l}), yielding:

$$v_a = \frac{y}{d_{yl}} + v_{0l} \Rightarrow \Delta v = v_a - v_{0l} = \frac{y}{d_{yl}} \Rightarrow y = \Delta v d_{yl} \tag{2.30}$$

$$\tan\alpha = \frac{y}{f} = \frac{\Delta v d_{yl}}{f_{yl} d_{yl}} \Rightarrow \alpha = tan^{-1}\left(\frac{\Delta v}{f_{yl}}\right) \tag{2.31}$$

Using this method, pitch estimation is computed more robustly than by using YOZ projection map. The use of the virtual disparity image allows also for providing robust estimation of camera height h and roll angle with respect Z axis, although at a higher computational cost.

2.3.3 Pitch correction

In order to have a steady estimation of the pitch angle, a linear Kalman filter [Kalman 60] is applied. The state vector contains the pitch angle and its velocity $x_k = \{\alpha_k, \dot{\alpha}_k\}$, while the measurement vector contains the pitch angle $z_k = \{\alpha_k\}$. The following equations show the proposed pitch angle estimation:

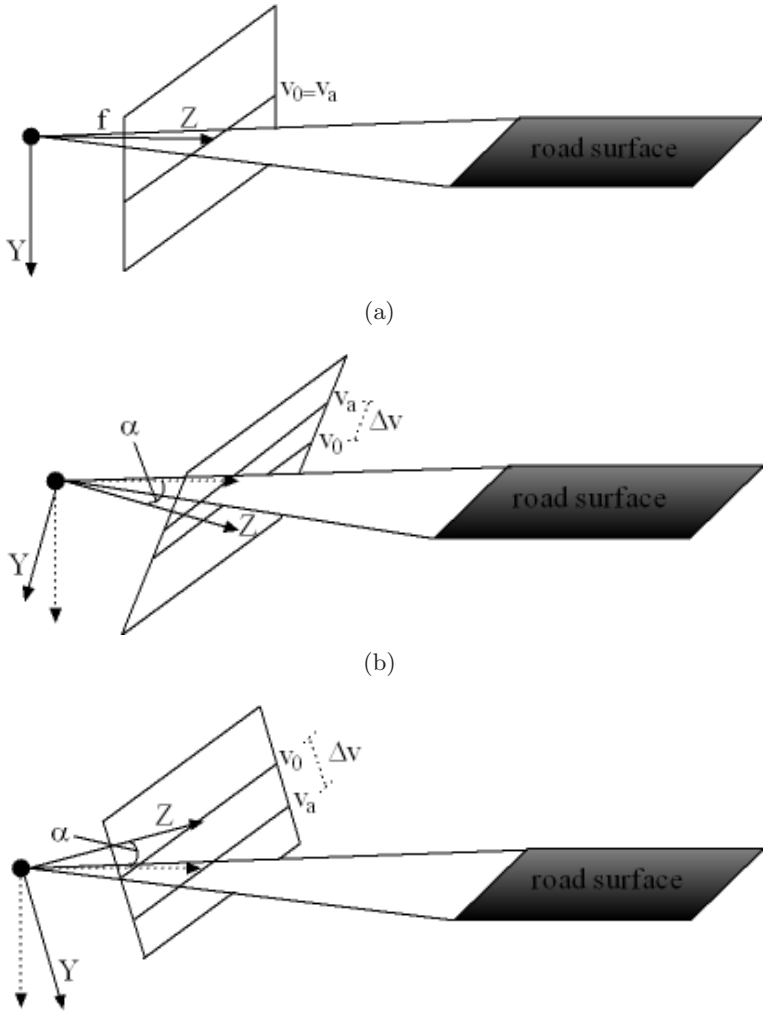


Figure 2.18: Ground plane projection on the virtual camera image plane (a) without pitch variation, (b) with positive pitch variation and (c) with negative pitch variation.

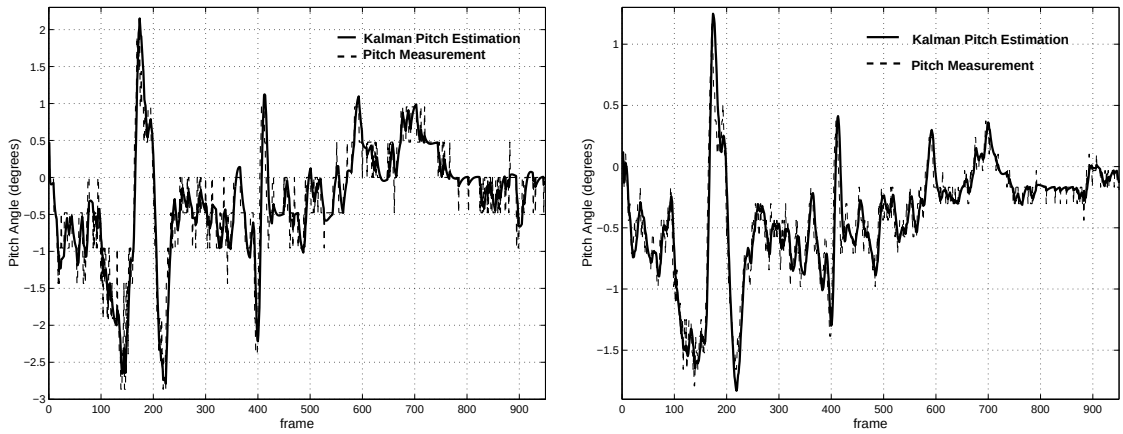
$$\vec{x_k} = \begin{pmatrix} \alpha_k \\ \dot{\alpha}_k \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_{k-1} \\ \dot{\alpha}_{k-1} \end{pmatrix} + \vec{r_k} \text{ state eq.} \tag{2.32}$$

$$z_k = \alpha_k + \vec{o_k} \text{ measurement eq.} \tag{2.33}$$

where $\vec{r_k}$ and $\vec{o_k}$ are the state vector noise and the measurement vector noise, respectively. Accordingly, a smoother pitch angle estimation is obtained. The transformation matrix that has to be applied in order to perform 3D points correction is:

$$R_\alpha = \begin{pmatrix} 1 & 0 & 0 \\ 0 & cos(\alpha) & -sin(\alpha) \\ 0 & sin(\alpha) & cos(\alpha) \end{pmatrix} \tag{2.34}$$

Once the longitudinal profile of the road has been extracted, and 3D points corrected, road surface points, which are not obstacle points, can be easily removed by analyzing their Y coordinate value. By doing so, these points do not perturb the clustering step. Figures 2.19(a) and 2.19(b) depict pitch angle estimation results in the same sequence using YOZ projection map and virtual disparity image, respectively. As can be observed, results obtained using virtual disparity image look more realistic and stable than those obtained using YOZ projection map.



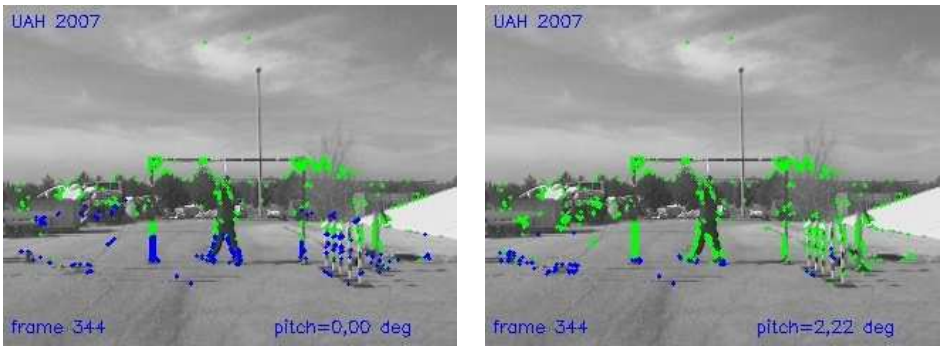
(a) (b)

Figure 2.19: Pitch angle measurement along with pitch angle estimation after Kalman filter: (a) using YOZ projection map and (b) using virtual disparity image.

Correct estimations of pitch angle have been used for compensating 3D measurements. As an example, Figures 2.20(a) and 2.20(b) show a sequence in which a car runs over a dummy. Points bellow $Y < 10cm$ are depicted in blue; points for which $Y \geq 10cm$ are depicted in green. When no pitch compensation is applied many points belonging to the dummy legs are considered as part of the ground plane.

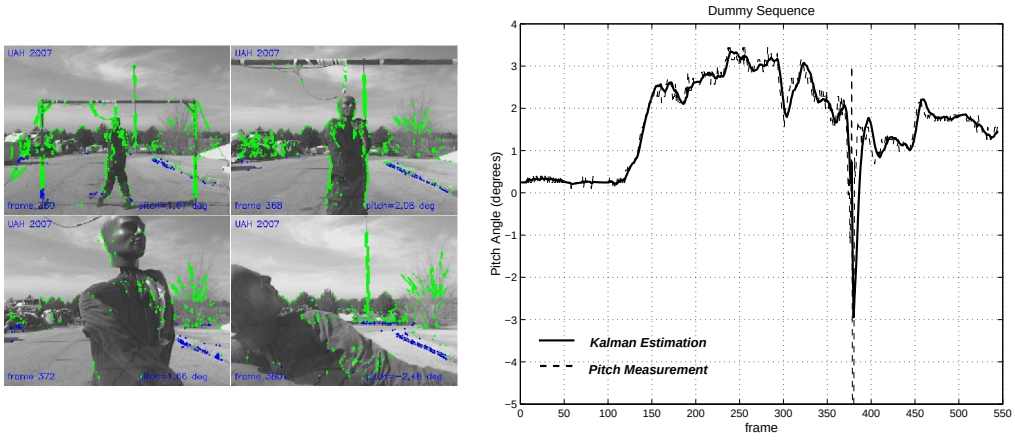
Pitch estimation is obtained using virtual disparity image, as depicted in Figure 2.21(b), where an abrupt change in the camera pitch angle can be appreciated when the car collides against the dummy (Figure 2.21(a)) due to the impact and the sudden braking after the collision.

Correct pitch estimation allows for correcting depth measurements. Thus, more accurate car-to-pedestrian distance and car-to-collision estimated times can be achieved. Figure 2.22(a) shows the absolute difference between depth measurements performed without pitch correction and depth measurements using pitch correction. The figure provides the absolute difference for angles between $1 - 10°$. For angles bellow $\pm 5°$ depth differences are not high (less than 10cm for depth up to 25m). However, for



(a) (b)

Figure 2.20: Separation between ground-plane(blue/objects(green); (a) without pitch estimation; and (b) with pitch angle correction.



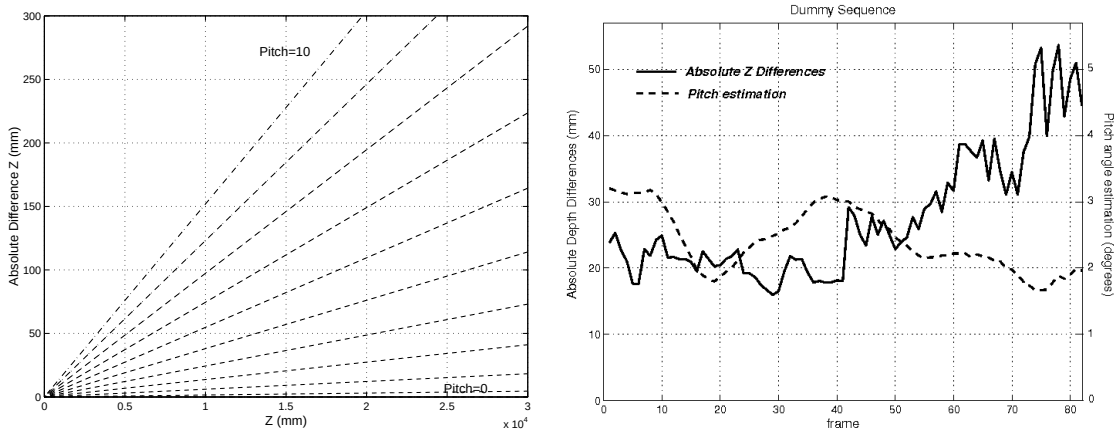
(a) (b)

Figure 2.21: (a) Separation between ground-plane/objects in a collision sequence; (b) Pitch estimation in a collision sequence.

large depth differences up to 10° errors in depth estimation can be between 15-40cm for depth ranging between 10-25m. In the example provided in Figure 2.22(b) the maximum depth correction is bellow 6cm.

Once the 3D points are compensated using the pitch angle estimation, separation between objects/ground-plane is carried out based on a simple height criteria. Given that $P'_l = (X'_l, Y'_l, Z'_l)^T = R_\alpha P_l$, the separation criteria is defined as follows:

$$\begin{cases} \text{if} \quad Y'_l \leq H_{min} \Rightarrow \text{ noise} \\ \text{if} \quad Y'_l > H_{min} \quad \text{AND} \quad Y'_l \leq H_{ground} \Rightarrow \text{ ground-plane point} \\ \text{if} \quad Y'_l > H_{ground} \quad \text{AND} \quad Y'_l \leq H_{max} \Rightarrow \text{ object point} \\ \text{if} \quad Y'_l > H_{max} \Rightarrow \text{ high object point} \end{cases} \tag{2.35}$$



(a) (b)

Figure 2.22: (a) Absolute depth difference with and without pitch correction; (b) Absolute depth difference with and without correction and pitch estimation in a sequence running over a dummy..

Parameter H_{min} determines the minimum distance under which detected points are considered as noise. H_{ground} determines the maximum height value for a 3D point to be considered as belonging to the ground plane. In practice, this value is set in the range between 10-20cm in order to reduce noise in the 3D reconstruction process. Parameter H_{max} establishes a maximum height value for a 3D point to be considered as part of a pedestrian. In practice this parameter has been set to 2.5m. In order to have a global idea about the object/ground-plane separation process, all the different stages of the algorithm are depicted in the scheme showed in Figure 2.23.

2.4 3D Clustering

Up to now all the frameworks that use stereo vision in order to deal with the candidates selection problem, operate, in different ways, over 2D images computed as from projections of the 3D points. For instance in [Zhao 00] a gray level segmentation over the disparity image is proposed. In [Franke 00] and [Gavrila 04] the obstacle detection procedure is done by using a multiplexed XOZ depth map, and selecting regions of interest whose number of depth features exceeds a percentage of the window are. Finally, in [Broggi 03] and [Grubb 04] the segmentation process is performed over v-disparity maps. The information for performing generic obstacles detection is defined with vertical lines.

In this section we describes the proposed segmentation method which directly operates over the 3D maps, exploiting the three dimensions in one step.

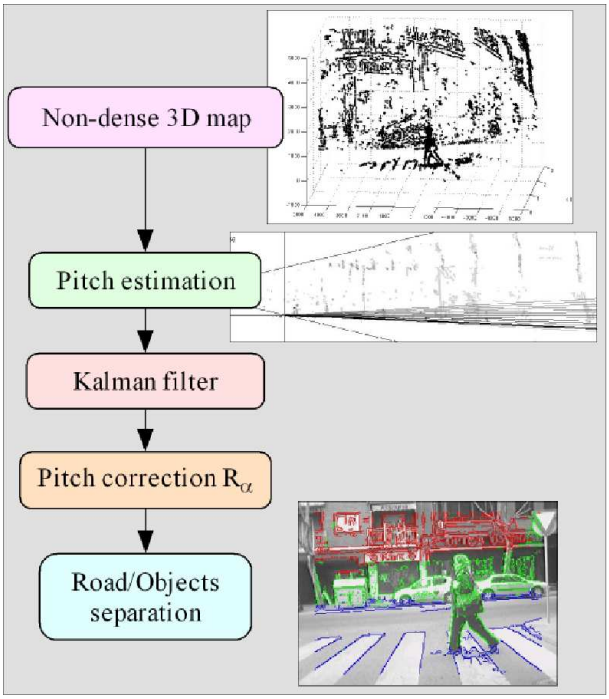


Figure 2.23: Global scheme of the ground-plane/object separation process.

2.4.1 Preprocessing

Spatial boundaries in the X axis are defined between $X_{min} = -5m$ and $X_{max} = 5m$ ($X_{lat} = \pm 5m$). The minimum depth distance is fixed to $Z_{min} = 2m$ and the maximum distance is pre-defined up to $Z_{max} = 30m$. Note that the cameras resolution (320×240) does not allow for a greater detection range. By taking into account the separation between objects/ground-plane carried out by means of the Equation 2.35, the segmentation process will only consider those points that satisfy next conditions:

$$\begin{cases} \text{if} & (Y'_l > H_{ground} \text{ and } Y'_l \leq H_{max}) \text{ AND} \\ & (X'_l > -X_{lat} \text{ and } X'_l \leq +X_{lat}) \text{ AND} \\ & (Z'_l > Z_{min} \text{ and } Z'_l \leq Z_{max}) \Rightarrow \text{accept point} \\ \text{else} & \Rightarrow \text{reject point} \end{cases} \tag{2.36}$$

Then, an outliers filtering process over the XOZ map (so-called *bird's eye map*) is carried out. The number of 3D projected points over the same 2D point in the XOZ map are taken into account. The XOZ map is filtered according to a neighbourhood criterion. For each I_{XOZ} image point, the sum of the points contained in a $(2m + 1) \times (2n + 1)$ window is computed as follows:

$$S(u,v) = \sum_{i=-m}^{i=m} \sum_{j=-n}^{j=n} I_{XOZ}(u+i, v+j) \tag{2.37}$$

If $S(u,v)$ is greater or lower than a pre-defined threshold, the point will be accepted or rejected, respectively. If a fixed threshold is used for all the XOZ points, there are problems due to the fact that close objects are defined by a large amount of points whereas far objects are defined with less number of points. In other words, if we use a high threshold value, some far objects may be eliminated since there were not enough number of projected points. On the other hand, if we use a low threshold value we will accept many errors that appear near the vehicle. Accordingly, the use of an adaptive threshold value Th_S is proposed as a linear function that varies in inverse proportion to the depth of the XOZ points. After several experiments this function has been defined as follows:

$$Th_S = Th_{max} - \frac{Z - Z_{min}}{Z_{max} - Z_{min}}(Th_{max} - Th_{min}) \tag{2.38}$$

where Th_{max} and Th_{min} are the required thresholds for points with minimum Z_{min} and maximum Z_{min} distances, respectively, which are manually tuned after extensive trials. The results after applying the proposed filtering process over the XOZ map are depicted in Figure 2.24.

2.4.2 3D Subtractive Clustering

Data clustering techniques are related to the partitioning of a data set into several groups in such way that the similarity within a group is larger than among groups [Kainulainen 02]. Normally the number of clusters is known beforehand. This is the case of *K-means* based algorithms. The needed clustering technique should be subject to some constraints:

- The number of clusters is considered unknown, since no a priory estimate of the number of pedestrians in scene can be reasonably made.
- The use of methods that measure the quality of the clustering result requires to launch the process N times, being N the maximum number of objects. This implies an intractable computational cost problem.
- The required clustering technique should not have convergence problems.
- The effects of outliers have to be reduced or completely removed in order to absorb correlation errors.

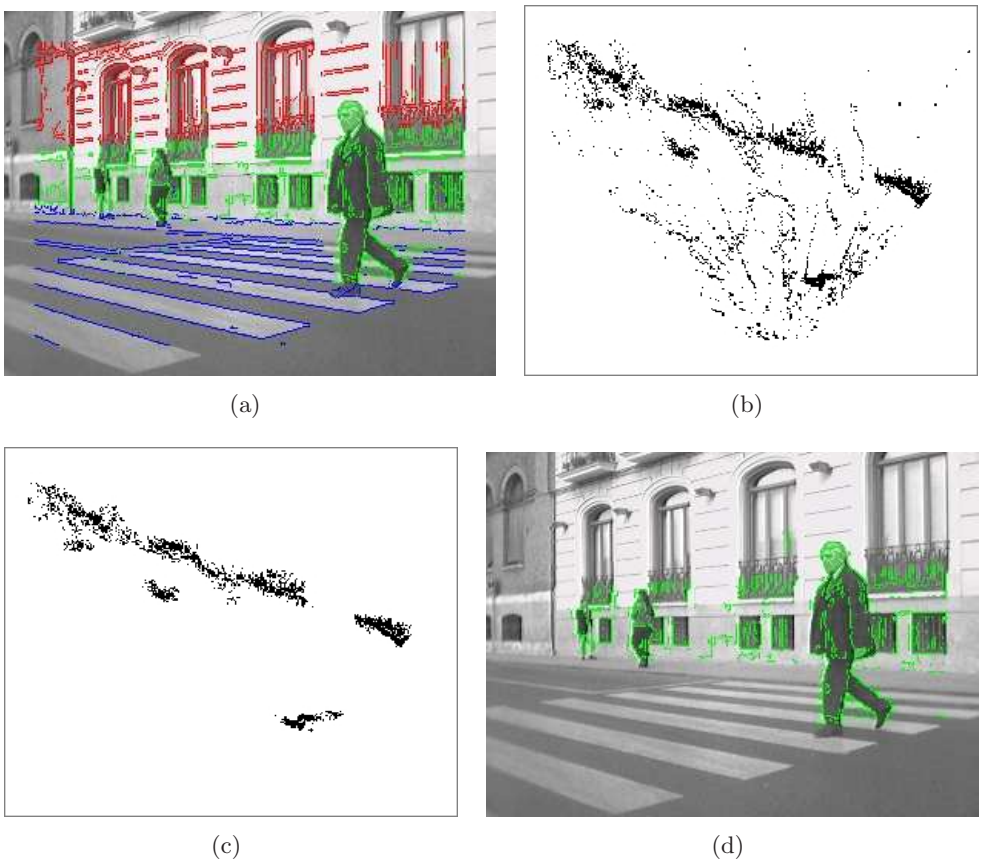


Figure 2.24: (a) 2D correlated points. Ground-plane points, object points and very high points are depicted with different colors; (b) Unfiltered XOZ map ; (c) Filtered XOZ map ; (d) Projection of the points that have not been rejected after the filtering process.

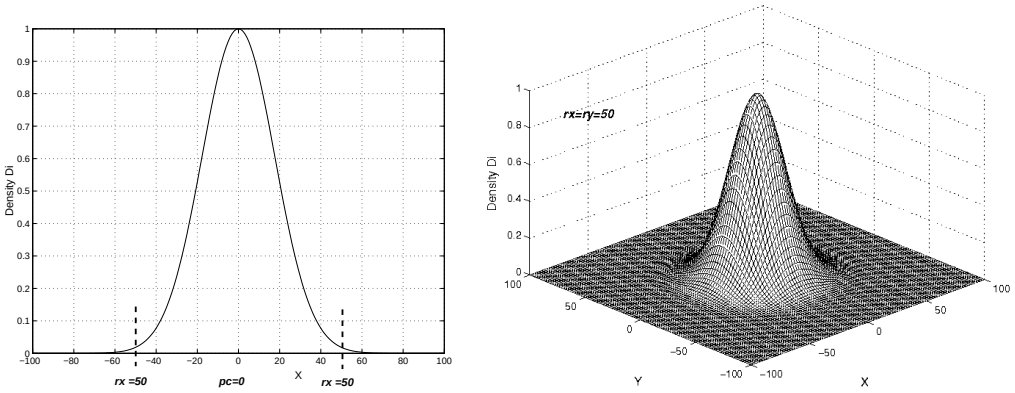
- It is necessary to define specific space characteristics in order to group different pedestrians in the scene.

For these reasons, a *Subtractive Clustering* method [Chiu 94a] is proposed, which is a well known approach in the field of *Fuzzy Model Identification Systems* [Chiu 94b]. The clustering is carried out in the 3D space, based on a density measure of data points. The idea is to find high density regions in 3D space. Objects in the 3D space are roughly modelled by means of Gaussian functions. It implies that, on principle, each Gaussian distribution represents a single object in 3D space. Nonetheless, objects that get too close from each other can be modelled by the system as a single one and, thus, represented by a single Gaussian distribution. The complete representation is the addition of all Gaussian distributions found in the 3D reconstructed scene. Accordingly, the parameters of the Gaussian functions are adapted in line with the depth by the clustering algorithm, so as to best represent the 3D coordinates of the detected pixels.

The 3D coordinates of all detected pixels are then considered as candidate clusters centers. Thus, each point p_i with coordinates (x_i, y_i, z_i) is potentially a cluster centre whose 3D spatial distribution D_i is given by the following equation:

$$D_i = \sum_{j=1}^{N} exp\left(- \frac{(x_i - x_j)^2}{\left(\frac{r_{ax}}{2}\right)^2} - \frac{(y_i - y_j)^2}{\left(\frac{r_{ay}}{2}\right)^2} - \frac{(z_i - z_j)^2}{\left(\frac{r_{az}}{2}\right)^2} \right) \tag{2.39}$$

where N represents the number of 3D points contained in a neighborhood defined by radii $r_a = (r_{ax}, r_{ay}, r_{az})$. Cluster shape can then be tuned by properly selecting the parameters r_{ax}, r_{ay}, r_{az}, which are related to 3D actual dimensions. As can be observed, candidates p_i surrounded by a large number of points within the defined neighborhood will exhibit a high value of D_i. Points located at a distance well above the radius defined by r_a will have almost no influence over the value of D_i (see Figure 2.25).



(a) (b)

Figure 2.25: Density function D_i. (a) One dimensional case with $r_x = 50$; (b) Two dimensional case with $r_x = r_y = 50$.

Equation 2.39 is computed for all 3D points measured after stereo reconstruction. Let $p_{cl} = (x_{cl}, y_{cl}, z_{cl})$ represent the point exhibiting the maximum density denoted by D_{cl}. This point is selected as the cluster centre at the current iteration of the algorithm. Densities of all points D_i are corrected based on p_{cl} and D_{cl}. For this purpose, the subtraction represented in equation 2.40 is computed for all points.

$$D_i = D_i - D_{cl} exp\left(- \frac{(x_i - x_{cl})^2}{\left(\frac{r_{bx}}{2}\right)^2} - \frac{(y_i - y_{cl})^2}{\left(\frac{r_{by}}{2}\right)^2} - \frac{(z_i - z_{cl})^2}{\left(\frac{r_{bz}}{2}\right)^2} \right) \tag{2.40}$$

where parameters $r_b = (r_{bx}, r_{by}, r_{bz})$ define the neighborhood where the correction of points densities will have the largest influence. The density of data point which

is close to the first cluster centre will be reduced, so that these data points can not become next cluster centre. Normally, parameters (r_{bx}, r_{by}, r_{bz}) are larger than (r_{ax}, r_{ay}, r_{az}) in order to prevent closely spaced cluster centres. Commonly let $r_{bx} = 1.5r_{ax}$, $r_{by} = 1.5r_{ay}$ and $r_{bz} = 1.5r_{az}$ [Chiu 94a].

After the subtraction process, density corresponding to the cluster center p_{cl} gets strongly decreased. Similarly, densities corresponding to points in the neighborhood of p_{cl} get also decreased by an amount that is a function of the distance to p_{cl}. All the points that satisfy the condition $D_i \leq Th_p$ are associated to the first cluster computed by the algorithm and will have no effect in the next step of the subtractive clustering. In fact these points are *subtracted* and restored as a 3D candidate. Th_p is experimentally set to a value of $Th_p = 0.1$.

After the correction of densities, a new cluster center $p_{cl,new}$ is selected, corresponding to the new density maximum $D_{cl,new}$ and the process is repeated whenever the condition expressed in equation 2.41 is met.

$$\text{if } U_{rel} > \frac{D_{cl}}{D_{cl,new}} \text{ and } D_{cl,new} > U_{min} \Rightarrow \text{new cluster} \tag{2.41}$$

where U_{rel} and U_{min} are experimentally tuned parameters ($U_{rel} = 0.3$ and $U_{min} = 0.35$) that permit to define a termination condition based on the relation between the previous cluster density and the new one, and a minimum value of the density function. The process is repeated, *subtracting* the points of each new cluster, until the termination condition given by equation 2.41 is not met.

2.4.3 Adaptive 3D Subtractive Clustering approach

There are two main reasons to propose the use of an adaptive approach to deal with the clustering process. Firstly, depth accuracy of the 3D maps depends on the images resolution and the distance between the cameras. And secondly, the number of points that defines an object will be decreased in proportion to the distance with regard to the vehicle.

Figure 2.26(a) shows the detected distance of the dummy, that appears in the dummy sequence used in Section 2.3.3, for both, pixel and subpixel accuracy, cases. As can be observed the accuracy is poorer when the dummy appears at a distances greater than 15m. In order to model this effect an adaptive value has been proposed to r_{az} taking into account the depth resolution of the stereo sensor. Even though stereo geometric relationship is known, for this case, we suppose the depth computation as follows:

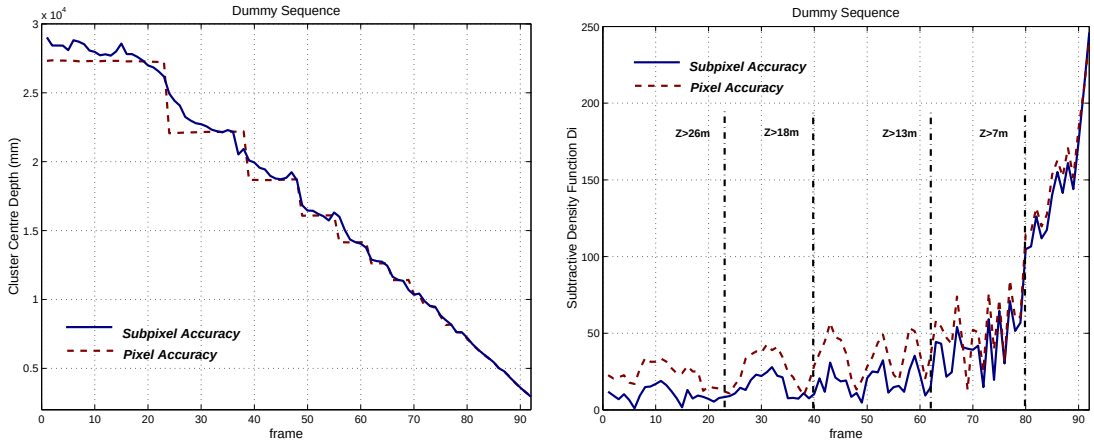
$$z = \frac{f_x B}{d_u} \tag{2.42}$$

where B is the stereo baseline length, f_x is the focal length expressed in units of horizontal pixels and $d_u = u_l - u_r$ is the horizontal disparity in pixels. Depth resolution is computed by using next equation:

$$\Delta z_i = f_x B \left(\frac{1}{d_{ui}} - \frac{1}{d_{ui} - 1} \right) = f_x B \left(\frac{1}{d_{ui}^2 - d_{ui}} \right) \tag{2.43}$$

According to equations (2.42) and (2.43) the adaptive value of r_{az} for a 3D point $p_i = (x_i, y_i, z_i)$ is given by:

$$r_{az} = 2\Delta z_i = 2 f_x B \frac{1}{d_{ui}^2 - d_{ui}} = 2 \frac{z_i^2}{f_x B + z_i} \tag{2.44}$$



(a) (b)

Figure 2.26: Dummy sequence. Pixel and subpixel accuracy analysis of the (a) distance of the dummy to the vehicle and (b) the subtractive clustering density function.

Figure 2.26(b) depicts the subtractive clustering density function for the cases with pixel and subpixel accuracy respectively, in the dummy sequence. The first conclusion we can obtain is that the differences in the density function are not relevant between pixel and subpixel approaches. However in both cases, the density function increases approximately exponentially with decreasing the distance between the dummy and the vehicle. When far and near objects would appear simultaneously, far objects could not be selected since U_{rel} and U_{min} are fixed. In order to avoid that problem we propose to use a correction factor f_c over the density function values D_i, which is defined as a function of the distance z_i. As can be observed in Figure 2.26(b) f_c has to be non linear. We propose to use next expression:

$$f_c(z_i) = 6 \exp^{-1} \left(\frac{z_{max} - z_i}{20} \right) - 0.7 \tag{2.45}$$

Note that f_c is only applied when the process is evaluating the termination condition expressed in equation (2.41) as from the Z coordinate of the clusters, i.e.:

$$\text{if } U_{rel} > \frac{f_c(z_{cl})D_{cl}}{f_c(z_{cl,new})D_{cl,new}} \text{ and } f_c(z_{cl,new})D_{cl,new} > U_{min} \Rightarrow \text{new cluster} \qquad (2.46)$$

Figure 2.27 is obtained by using exactly the same sequence depicted in Figure 2.26(b), and it depicts the ideal correction function $f_c(z_i)$ expressed in equation (2.45), the actual correction function for the distances measured in the dummy sequence, as well as the density function values obtained with and without correction. The use of f_c allows to artificially increase the density function values of the far objects, while the density function values of the near objects are almost not affected.

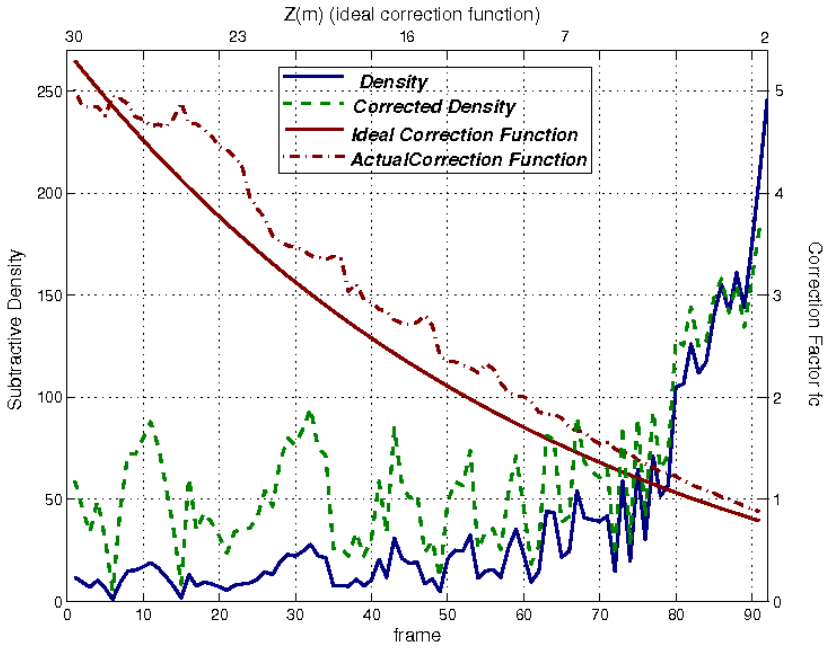


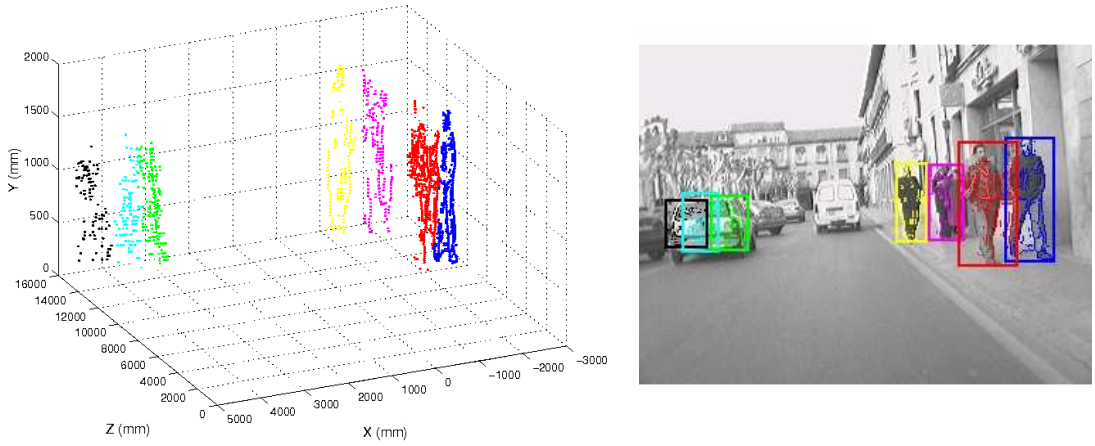
Figure 2.27: Dummy sequence. Subtractive clustering density function correction as from the correction factor which is defined as a function of the distance between the clusters and the vehicle.

.

2.4.4 Candidates analysis

After applying subtractive clustering to a set of input data, each cluster finally represents a candidate. Pedestrian candidates are then considered as the 2D region of interest (ROI) defined by the projection in the left image plane of the 3D candidate

regions. For instance, in Figures 2.28(a) and 2.28(b) the results after applying subtractive clustering and computing the 2D boundaries as from the projection of the 3D points for each cluster are depicted.



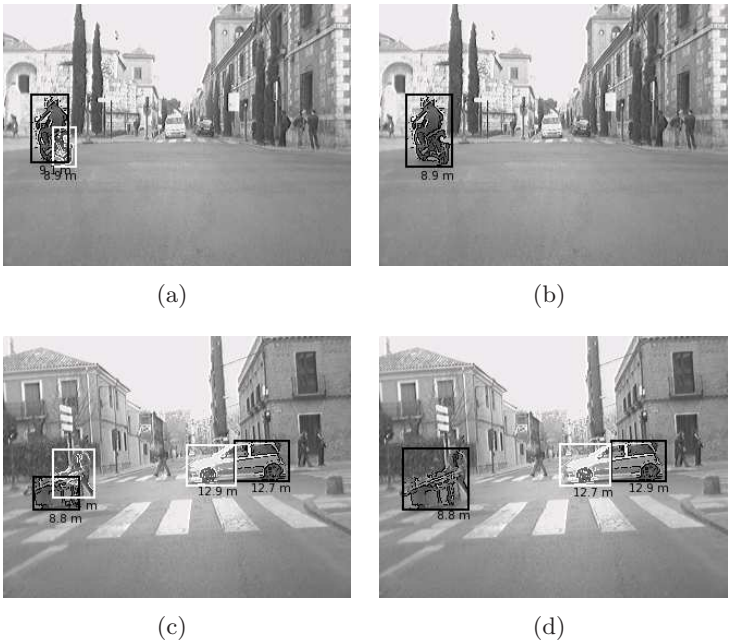
(a) (b)

Figure 2.28: (a) Subtractive clustering results in the 3D space. (b) Candidates selection by computing the 2D boundaries as from the projection of the 3D points in the left image plane for each cluster.

As it was stated in Section 2.4.3 the third parameter of the radii $r_a = (r_{ax}, r_{ay}, r_{az})$ is defined as a function of the depth accuracy. Thus, several cases are corrected such as the ones where far candidates are split in two parts due to the poor depth accuracy at those distances. This effect can be observed in Figure 2.29. On the other hand the second parameter r_{ay} is less problematic as long as road points and very high points had been properly rejected. If we define a too high value for r_{ay}, the density function values will be decreased, but all clusters would be affected in the same extent. Accordingly this parameter should be defined by taking into account the parameter $U_{min} = 0.35$ of the termination condition. For $U_{min} = 0.35$ the second parameter of the radii is fixed to $r_{ay} = 100cm$. If we use greater values for r_{ay} the subtractive clustering may yield very high candidates as the ones depicted in Figure 2.30.

The most critical parameter of the radii r_a is the first one r_{ax}. It is necessary to achieve a trade-off between a minimum value able to be discriminant for pedestrians volumes, and a maximum value that does not merge very close pedestrians. After comprehensive experiments this parameter is tuned to a value of $r_{ax} = 70cm$. Several examples of the clustering results depending on the value of r_{ax} are depicted in Figure 2.31.

The 2D projection of the 3D points in the left image plane only assures a good road contact (contact point between candidates and road) if a good estimation of the pitch has been previously carried out. In Figure 2.32 two examples with both accurate and inaccurate pitch estimation are shown. In cases where there were a steep change in

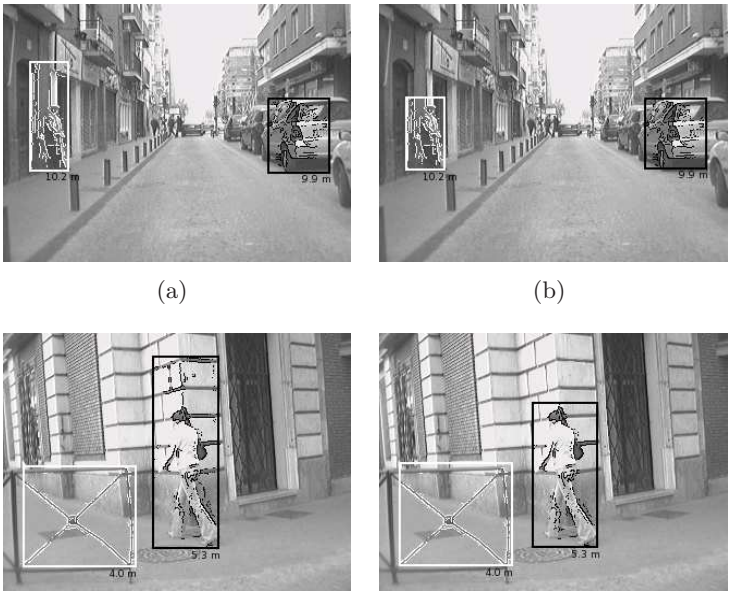


(a) (b)

(c) (d)

Figure 2.29: (a)(c) Examples with $r_{az} = 100cm$. Candidates are divided in two parts due to the depth accuracy ; (b)(d) Corrected examples with adaptive value $r_{az}(z_i) = 2z_i^2/(f_x B + z_i)$.

.

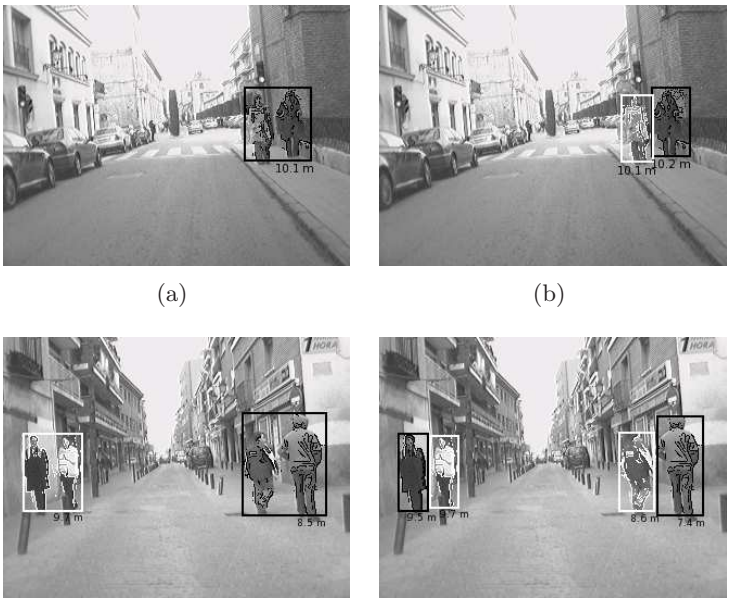


(a) (b)

(c) (d)

Figure 2.30: (a)(c) Examples with $r_{ay} = 150cm$. Very high candidates ; (b)(d) Corrected examples with $r_{ay} = 100cm$.

.

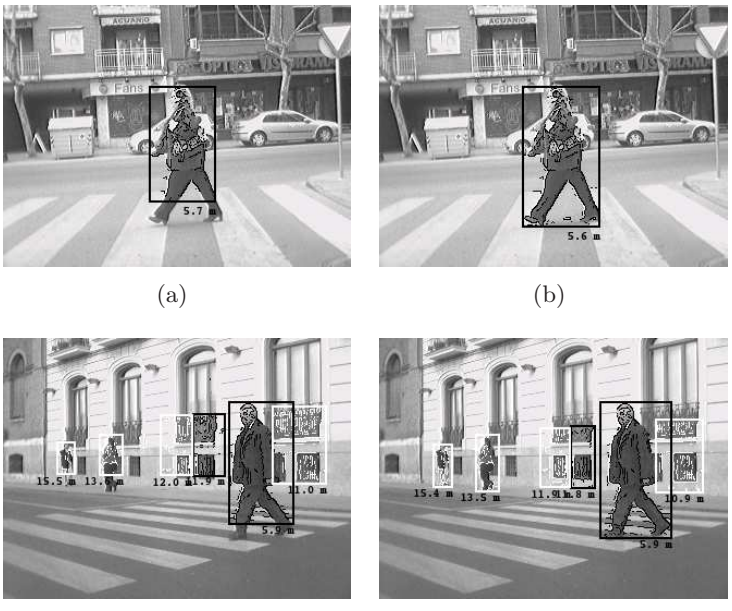


(a) (b)

(c) (d)

Figure 2.31: (a)(c) Examples with $r_{ax} = 110cm$. Two pedestrians are merged as only one candidate ; (b)(d) Corrected examples with $r_{ax} = 70cm$.

.



(a) (b)

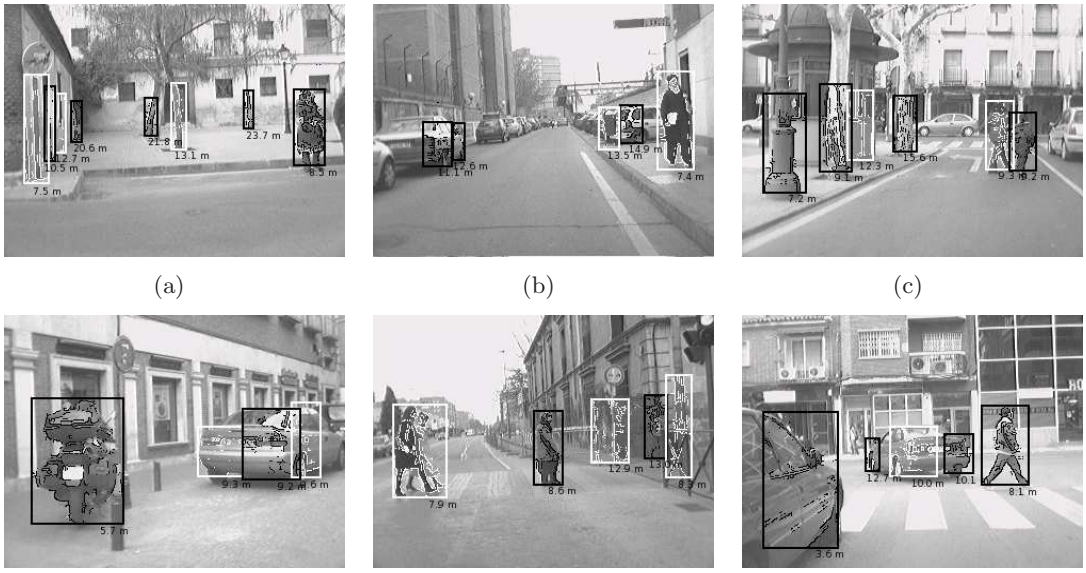
(c) (d)

Figure 2.32: Contact point between the road and the candidates (a)(c) Inaccurate results without pitch compensation. (b)(d) Accurate results with pitch compensation.

.

the pitch angle pitch estimation also allows for correcting depth measures, achieving more accurate car-to-pedestrian distance and car-to-collision estimated times.

Thanks to the use of the adaptive subtractive clustering technique the amount of candidates per frame, in average, and their variability are strongly decreased, reducing the complexity of the later learning tasks. As can be observed in Figure 2.33 the type of candidates yielded by the proposed method have a 3D volume similar to pedestrians, i.e., trees, lampposts, wastebaskets, traffic lights, fences, containers, the side of the cars, building fronts, etc.



(a) (b) (c)

(d) (e) (f)

Figure 2.33: Several examples of the types of candidates yielded by the adaptive subtractive clustering method in urban environments.

2.5 Conclusions

The previous discussion lets us state that the non-dense representation of the environment, as from Canny points obtained by using adaptive thresholds as a function of the gradient magnitude, proves successfully to provide a robust representation which allows for a correct object segmentation. After the supervised calibration process, based on the Camera Calibration Toolbox for Matlab, an automatic pitch and height calibration method has been proposed. Thus an initial estimation of both parameters is obtained, being available as default values in later stages.

A robust correspondences search algorithm has been proposed, based on ZNCC matching technique. Even the fact that ZNCC is heavy from a computational point of view, it allows to do not compensate images for illumination differences. By using unique maximum criteria, mutual consistency check and minimum disparity selection

when there were multiple correspondences, a considerable amount of stereo-matching errors are avoided. Subpixel accuracy improves the stereo depth accuracy. The proposed method for carrying out the triangulation process is projective-invariant.

Two different methods for pitch compensation have been proposed and analyzed. Both need to use 3D information from the environment. The first one is based on the use of YOZ projection of the 3D points whereas the second one resides on the use of the so-called virtual disparity image. Both methods are filtered by means of a linear Kalman filtering. The use of the virtual disparity image is less sensitive to reconstruction errors. A correct pitch estimation at each frame is mandatory for a correct separation between road/objects points.

An adaptive subtractive clustering technique has proved to be robust in order to detect generic obstacles with volumes similar to pedestrians. This method directly operates over the non-dense 3D maps. That is not the case of most of the frameworks which usually work with bi-dimensional information ([Zhao 00], [Franke 00], [Gavrila 04], [Broggi 03] and [Grubb 04]).

Thanks to the proposed stereo candidate selection method the amount of candidates per frame and the variability are strongly decreased. Thus, low false positive rates are assured as long as the candidates were visible and distinguishable and at a distances lower than 30m. In comparison with monocular approaches the stereo approach allows for lower false positive rates along with lower variability, reducing the complexity of the later learning tasks.

Chapter 3

Pedestrian detection using SVM

Pedestrian detection is done using Support Vector Machines (SVM) [Vapnik 95], [Cristianini 00]. The use of SVM in the field of pedestrian recognition has become a common approach for many researchers in the last years since it was first proposed in [Oren 97], [Papageorgiou 00] and [Mohan 01]. Traditional training techniques for classification tasks, such as multilayer perceptrons (MLP) use empirical risk minimization and only assure minimum error over the training data set. In contrast, the SVM machinery uses structural risk minimization which minimizes a bound on the generalization error and, therefore, should perform better on novel data. In addition, the SVM decision surface, i.e., the hyperplane that optimally separates two high-dimensional classes of objects, depends only on the inner product of the feature vectors. This leads to an important extension since we can replace the Euclidean inner product by a kernel. The use of a kernel is equivalent to mapping the feature vectors to a higher dimensional space and, thereby, significantly increasing the discriminative power of the classifier. For a more detailed description see [Christopher 98].

3.1 Training strategy

The first step in the design of the training strategy is to create representative databases for learning and testing. The training and test sets were manually constructed using the TSetBuilder tool [Nuevo 05], developed in our lab. The following considerations must be taken into account when creating the training and test sets.

- The ratio between positive and negative samples has to be set to an appropriate value. A very large number of positive samples in the training set may lead to a high percentage of false positive detections during on-line classification. On the contrary, a very large number of negative samples produces misslearning.

A trade-off of 1 positive sample for every 2 negative samples was initially chosen in our application and compared to the 1/1 option. Although the results attained after experiments do not exhibit a dramatic difference in performance, a slightly superior behavior is obtained by using training sets following the 1/2 positive/negative ratio. Accordingly, the rest of the experiments were carried out using that ratio between the number of positive and negative samples.

- The size of the database is a crucial factor to take care of. As long as the training data represent the problem well, the larger the size of the training set the better for generalization purposes. Nonetheless, the value of the regularization coefficient C [Christopher 98] is important since this parameter controls the degree of overlearning. Thus, a small value of C allows a large separation margin between classes, reducing overlearning and improving generalization. In this work, a value C=1.0 has been used after extensive trials. This value can be considered as a small one. The dimension of the database has to be accordingly designed in order to achieve real generalization.

- The quality of negative samples has a strong effect in the detection rate. Samples that are too easy to learn are not good for generalization. Quite the opposite, complicate samples are needed to achieve fine grain separation as indicated in [Shashua 04]. Negative samples belonging to the ground or the sky, for instance, do not greatly contribute to achieving a refined classifier, as long as the problem of class separation becomes an easy task under those circumstances. Negative samples have to be properly selected so as to account for ambiguous objects, such as poles, trees, advertisements, and the like. Only by following this strategy when creating the training sets a really powerful classifier can be achieved in practice.

- A sufficiently representative test set must be created for verification. The content of the test set has similar characteristics to those of the training sets in terms of variability, ratio of positive/negative samples, and quality of negative samples.

A detailed observation of the classifier operation in practice suggests the subdivision of the classification task into several, more tractable learning sets according to different practical considerations. A major issue is the effect of illumination conditions. It is clear that daytime and nighttime samples must be compulsorily separated in order to create multiple specialized classifiers. The nighttime classifier can be reasonably expected to operate correctly only in very short distances (bellow 6-8 m) for non-illuminated areas, where pedestrians can be appropriately illuminated by the car beams (infrared images would be needed in order to achieve long range detection). Nonetheless, nighttime pedestrian detection can be done up to 15-20 m in illuminated areas.

The separation between day and night specialized classifiers may not be enough to cover the most significant cases of pedestrians variability. In fact, as observed

in practice, the effect of depth is determinant. SVM classifiers tend to produce different outputs depending on the original candidate size in the image plane, in spite of the re-scaling stage that is executed prior to classification. The appearance of normalized candidates itself is quite different depending on the distance between the pedestrian and the cameras. Shapes and edges are not so neatly distinguished when pedestrians are beyond 12-15 m from the cameras. Accordingly, the effect of depth suggests the development of specialized SVM classifiers at daytime. Albeit several subdivisions could be done for very short, short, medium, long, and very long range, two specialized classifiers for short and long range detection have been considered to be enough in practice. The threshold between short and long range has been empirically set to 12 m.

The effect of pose must also be taken into account as a significant source of variability in pedestrians appearance. Most pedestrians appear standing in the surrounding of the vehicle, either walking or simply in a stationary position. The differences between walking and stationary pedestrians are clear. There are even some remarkable differences between pedestrians moving laterally, with regard to the vehicle trajectory, and those moving longitudinally. Pedestrians intersecting the vehicle trajectory from the sideways are usually easier to recognize since their legs are clearly visible and distinguishable. In fact, some authors have proposed two separate SVM classifiers according to this statement [Grubb 04]. A more complicated case occurs when a pedestrian crouches or bends down.

The appearance of pedestrians in those cases is dramatically different from that of standing pedestrians. Changes due to different clothing also contribute to further complexity in the variability problem. Thus, large skirts and coats make pedestrians look very different from those in trousers and suits. Likewise, pedestrians bringing trolleys or bags make the recognition problem even more difficult. Had it not been enough, pedestrians legs are not always visible in the image, specially when pedestrians are very close to the vehicle. This is a critical case of great importance for pre-crash protection systems (active hood, pedestrian protection airbag, etc.).

In order to handle all these variability cases, we have created separate training sets intended to perform pedestrians learning in short and long range, respectively, at daytime and nighttime. Four training sets were built for this purpose containing a number of negative samples that doubles the number of positive ones: a training set of 9.000 daytime long-range samples, denoted by DL, a training set with 15.000 daytime short-range samples, denoted by DS, a third training set containing 6.000 nighttime samples, denoted by N, and a global training set containing the concatenation of all samples in DL and DS (24.000 samples), denoted by G. Similarly, four test sets were created and denoted by TDS (test set for daytime short-range, 5.505 samples), TDL (test set for daytime long-range, 4.320 samples), TN (test set for nighttime, 3.225 samples), and TG (global test set composed by the concatenation of TDS and TDL, 9.825 samples), respectively. Variability due to pose, clothing, and other artifacts is handled by creating adequate training databases containing as many representative cases as possible. In this stage, pedestrians in different pose (standing, walking,

ducked) and clothing (coats, skirts, etc) are included in the database, as well as pedestrians with handbags and other artifacts. In total, the training sets contain up to 30.000 samples, while the test sets amount up to 13.050 samples.In Table 3.1 a detailed description of the different training and test data sets is depicted.

Train DB	**# Samples**	**Illumination**	**Range**	**Pos/Neg Ratio**
DS	15.000	daytime	short	1/2
DL	9.000	daytime	long	1/2
G	24.000	daytime	short & long	1/2
N	6.000	nighttime	short & long	1/2
F	3.000	daytime	short & long	1/2
B	2.000	daytime	short & long	1/2
Test DB	**# Samples**	**Illumination**	**Range**	**Pos/Neg Ratio**
TDS	5.505	daytime	short	1/2
TDL	4.320	daytime	long	1/2
TG	9.825	daytime	short & long	1/2
TN	3.225	nighttime	short & long	1/2
TB	1.000	daytime	short & long	1/2

Table 3.1: Number of samples, nomenclature, positive/negative ratio, illumination conditions and range for all the training and test data sets used throughout this Section.

With the purpose of showing the influence of the bounding box accuracy in the global performance of the classifier two training data set was specifically devised: a subdivision of G of 3.000 samples, denoted by F, and a new training data set of 2000 badly bounded candidates, denoted by B. In addition a test set containing 1000 badly bounded samples, denoted by TB, was created to study the bounding box effect.

The quality of the classification system is mainly measured by means of the *detection rate* (DR) and *false positive rate* (FPR) ratio (see contingency table in Figure 3.1). These two indicators are graphically bounded together in a *Receiver Operating Characteristic (ROC)* and can be computed as follows:

$$DR = \frac{TP}{P} = \frac{TP}{TP + FN} \quad ; \quad FPR = \frac{FP}{N} = \frac{FP}{FP + TN} \tag{3.1}$$

The selection of the FPR value has been made to show performance in representative points where differences between curves can be optimally appreciated. FPR must be a value for which DR exhibits an acceptable value. This leads to selecting 5% in some cases or 10% in others. For cases in which 10% has been chosen, a value of 5% would not make sense since DR would be a really poor value in those conditions. In addition, FPR has been chosen as a value from which practically no cross points occur among the ROC curves of the different features. This means that a curve that is better than another at a given FPR_i remains better for almost all FPR values greater than the given FPR_i , as can be observed in the Figures provided in this

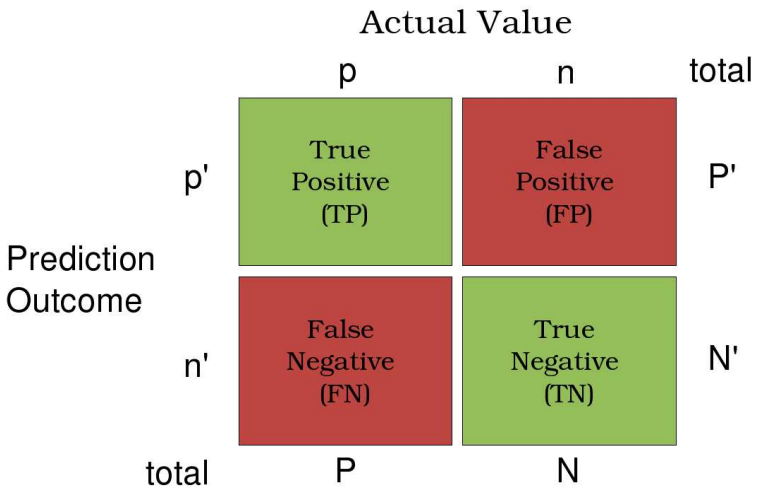


Figure 3.1: Contingency table for binary classification.

section. Accordingly, different FPR values have been selected for different types of tests in order to provide meaningful comparisons.

3.2 Classifier structure

A two-stage classifier is proposed in order to cope with the components-based approach, as depicted in Figure 3.2. In the first stage of the classifier, features computed over each individual fixed sub-region are fed to the input of individual SVM classifiers. Thus, there are 6 individual SVM classifiers corresponding to the 6 candidate sub-regions. These individual classifiers are specialized in recognizing separate body parts corresponding to the pre-specified candidate sub-regions. It must be clearly stated that no matching of parts is carried out. Instead, each individual SVM is fed with features computed over its corresponding candidate sub-region, and provides an output that indicates whether the analysed sub-region corresponds to a pedestrian part (+1, in theory) or not (-1, in theory). The training set is supposed to contain enough number of pedestrian samples in different poses so as to provide the variability needed by the components-based approach to operate properly. For example, let us consider the case of a pedestrian crossing the street laterally. The fixed sub-regions assigned to the candidate arms will most likely not contain any pedestrian's arm in some frames, since arms are occluded to the camera or are located in the same sub-region where the torso is, when a pedestrian is walking laterally. However, classification is still possible based on the pedestrian's head and legs. This statements constitute the main rationale for supporting the use of a components-based classifier.

In the second stage of the classifier, outputs provided by the 6 individual SVMs are combined (see Figure 3.2). Two different methods have been tested to carry out this

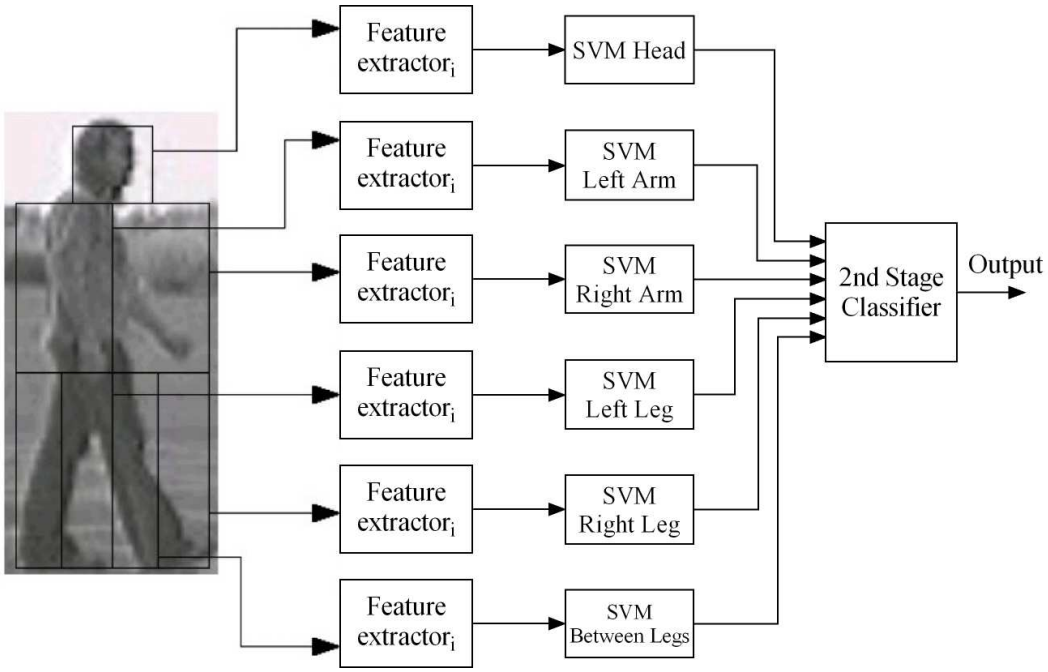


Figure 3.2: Outline of the two stage classifier.

operation. The first method implements what we denote as simple-distance criterion. A simple addition is computed, as indicated in next equation:

$$S_{distance-based} = \sum_{i=1}^{6} S_i \tag{3.2}$$

where S_i represents the output of the SVM corresponding to sub-region i. In theory, sub-regions corresponding to non-pedestrians or missing parts should contribute with negative values to $S_{distance-based}$. Likewise, sub-regions corresponding to pedestrians parts should contribute with positive values to the final sum. A threshold value T_{SVM} is then established in order to perform candidates classification. This threshold is parametrized for producing the *Receiver Operating Characteristic (ROC)*. The difference between pedestrians and non-pedestrians is set depending on the distance between T_{SVM} and $S_{distance-based}$. Thus, if $S_{distance_based}$ is greater than T_{SVM} the candidate is considered as pedestrian. Otherwise, it is regarded as non-pedestrian. This simple mechanism is what we denote as distance-based criterion which can be summarized as follows:

$$\begin{cases} \text{if} & S_{distance-based} \geq T_{SVM} \Rightarrow \text{ pedestrian} \\ \text{else if} & S_{distance-based} < T_{SVM} \Rightarrow \text{ non-pedestrian} \end{cases} \tag{3.3}$$

The second method that has been tested to implement the second stage of the clas-

sifier relies on the use of another SVM classifier. A second-stage SVM merges the outputs of the 6 individual first-stage SVM classifiers and provides a single output representing the candidate classification result. The resulting global structure is denoted as 2-stage SVM classifier. Obviously, the second-stage SVM classifier has to be trained with supervised data. The training set for the second-stage SVM classifier has been built as follows. First, the 6 individual first-stage SVM classifiers are properly trained using training set DS (containing 15.000 samples), in which the desired outputs (pedestrian or non-pedestrian) are set in a supervised way. Then, a new training set is created by taking as inputs the outputs produced by the 6 already trained first-stage SVM classifiers (in theory, between -1 and +1) after applying the 15.000 samples contained in DS, and taking as outputs the supervised outputs of DS. The test set for the second-stage SVM classifier is created in a similar way, using test set TDS (containing 5.505 samples).

3.3 Features extraction methods

The choice of the most appropriate features for pedestrian characterization remains a challenging problem nowadays, since recognition performance depends crucially on the features that are used to represent pedestrians. In a first intuitive approach, some features seem to be more suitable than others for representing certain parts of human body. Thus, legs and arms are long elements that tend to produce straight lines in the image, while the torso and head are completely different parts, not so easy to recognize. This statement, although based on intuition, suggests the combination of several feature extraction methods for the different sub-regions into which a candidate is divided. Accordingly, we have tested a set of 8 different feature extraction methods. The selection of features was made based on intuition, previous work carried out by other authors, and our own previous work on other applications. The proposed features are briefly described in the following lines.

- Canny image: the Canny edge detector [Canny 86] computes image gradient, highlighting regions with high spacial derivatives. It is known to many as the optimal edge detector. The computations of edges significantly reduces the amount of data that needs to be managed and filters out useless information while preserving shape properties in the image. The result obtained after applying a Canny filter to the region of interest is directly applied to the input of the classifier. The Canny-based features vector is the same size as the candidate image, i.e., $w_{ROI} \times h_{ROI}$.
- Haar Wavelets: originally proposed for pedestrian recognition in [Oren 97] and [Papageorgiou 00]. There are four components with a size of $(w_{ROI} \times h_{ROI})/4$ each one. This yields a features vector of $w_{ROI} \times h_{ROI}$.
- Gradient magnitude and orientation: the magnitude of the spatial derivatives, g_x and g_y are computed for all pixels in the image plane after using Sobel

operator [Sobel 68]. After that, orientation is calculated as $\theta = \arctan(g_x, g_y)$. The resulting features vector has twice the size of the candidate image, i.e., the vector has $2 \times w_{ROI} \times h_{ROI}$ elements.

- Coocurrence matrix: the coocurrence is specified as a matrix of relative frequencies $P_{i,j}$ with which two neighbouring pixels, separated by distance d at orientation θ, co-occur in the image, one with gray level i and the other with gray level j [Haralick 79]. The coocurrence matrix is computed over the Canny image. Resulting matrices are symmetric and can be normalized by dividing each entry in a matrix by the number of neighboring pixels used in the matrix computation. In our approach we propose a distance of 1 pixel and 4 different coocurrence matrices for the following orientations (bins): $(0°, 45°, 90°, 135°)$. The resulting size of the feature vector is $4 \times 2 \times 2 = 16$, which implies to manage a feature extraction method that does not depend on the image size.

- Histogram of intensity differences: the relative frequencies of intensity differences are computed between neighbouring pixels along 4 orientations over a normalized image of 128 gray levels. This generates a features vector of $4 \times 128 = 512$ which is also independent on the image size.

- Texture unit number (NTU): the local texture information for a pixel can be extracted from a neighbourhood of 3×3 pixels, that represents the smallest complete unit of texture. The corresponding texture unit is computed by comparing the pixel under study with his 8 neighboring pixels [Wang 90]. The NTU process generates a features vector with the same size as the candidate image, i.e., a features vector of $w_{ROI} \times h_{ROI}$ elements.

- Histograms of oriented gradients locally normalized (HOG): this method has been successfully applied for pedestrian detection in [Dalal 05]. The aim of this method is to describe an image by a set of local histograms which count occurrences of gradient orientation in a local part of the image. The image is split into cells. For each cell we compute histogram of gradients by accumulating votes into bins for each orientation. The normalization is done among a group of cells, which is referred to as a block. The final descriptor is built by concatenating all histograms into a single vector. The vector dimension computation is more complex: let (w_{ROI}, h_{ROI}), (w_{CELL}, h_{CELL}) y (w_{BLOCK}, h_{BLOCK}) represent the size of the image, cells and blocks, respectively. Then:

$$\begin{cases} CI_r = w_{ROI}/w_{CELL} & \Rightarrow & \#\text{ cells per row} \\ CI_c = h_{ROI}/h_{CELL} & \Rightarrow & \#\text{ cells per column} \\ CB_r = w_{ROI}/w_{CELL} & \Rightarrow & \#\text{ blocks per row} \\ CB_c = h_{ROI}/h_{CELL} & \Rightarrow & \#\text{ blocks per column} \\ DB_r = CI_r - CB_r + 1 & \Rightarrow & \#\text{ overlapped blocks per row} \\ DB_c = CI_c - CB_c + 1 & \Rightarrow & \#\text{ overlapped blocks per column} \end{cases} \tag{3.4}$$

The final feature vector dimension is computed by $DB_r \times DB_c \times CB_r \times CBc \times b$, where b is the number of bins per histogram.

- Histogram of gradient orientations (HON): the idea of HON descriptor is based on the HOG features. The gradient image is considered. Orientation is discretized to 20 bins (corresponding to an accuracy of 18°. Only pixels in the gradient image exhibiting a magnitude greater than some threshold (10) are considered. For those pixels, the values of gradient are accumulated in a 20-bins histogram. There is not cells distribution neither blocks normalization. Thus, the resulting features vector has 20 elements.

For better invariance to illumination and shadowing, all feature vectors are normalized. The main advantage of scaling is to avoid attributes in greater numeric ranges dominate those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation since large attribute values might cause numerical problems. As we use the library LIBSVM [Chang 01] for implementing SVM, we use the recommended range $[-1, +1]$ for linearly scaling each attribute.

3.4 Optimal kernel selection

An optimal kernel selection for SVM-based pedestrian classification has been carried out between polynomial, radial basis function (RBF) and sigmoid kernels. Linear kernel does not achieve a solution on the problem since it is a non linear problem. A training data set of 10.000 samples has been created by using a subdivision from DS. The test data set, with a total amount of 3670 samples, has been also performed by using a subdivision of TDS.

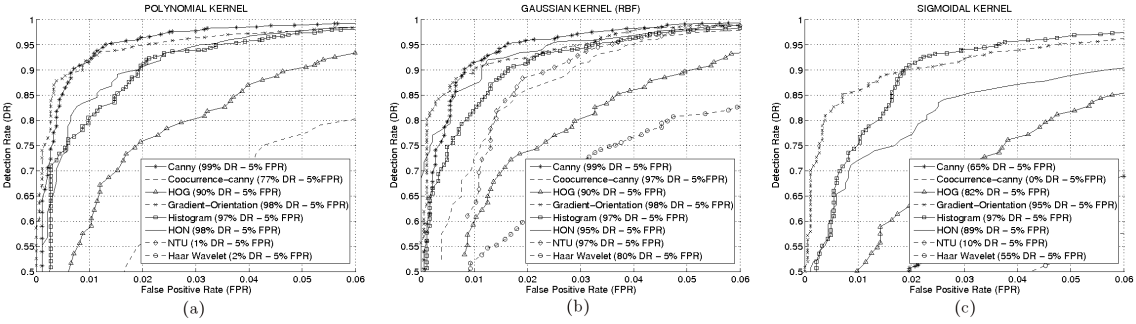


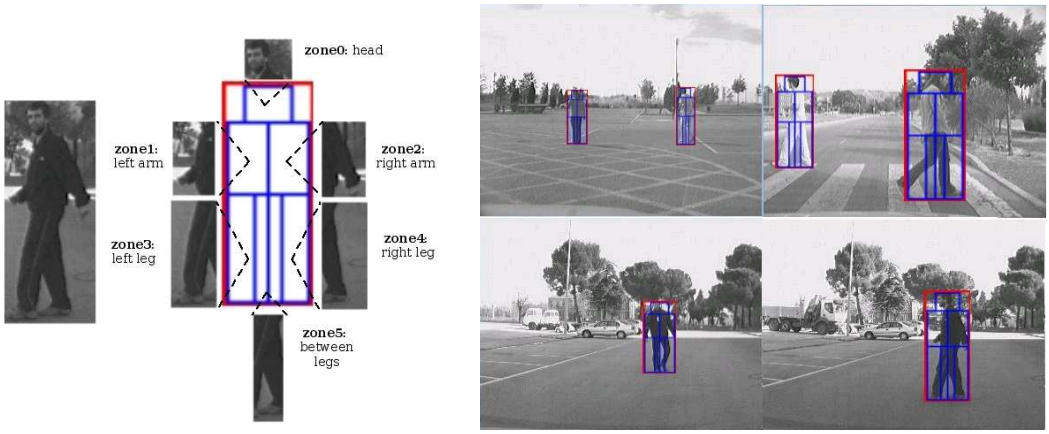
Figure 3.3: ROC curves for (a) polynomial, (b) radial basis function (RBF) and (c) sigmoid kernels.

As can be observed in Figure 3.3 only RBF kernel guarantees a reasonable solution for all the feature extraction methods. Polynomial kernel is not able to satisfactorily model the learning problem for Haar Wavelet and NTU descriptors, which also happens with the Canny features with sigmoid kernel. Even the fact that some descriptors perform slightly better with other kernels (for instance, the combination of HON descriptors with the polynomial kernel yields 98% of DR versus 95% of DR in case of the RBF kernel) we conclude that the optimal performance for almost all

the feature extraction methods is achieved by using the radial basis function RBF kernel.

3.5 Holistic vs components-based

There are some important aspects that need to be addressed when constructing a classifier, such as the global classification structure and the use of single or multiple cascaded classifiers. These issued are strongly connected to the way features are extracted. The first decision to make implies the development of a holistic classifier against a components-based approach. In the first option, features are extracted from the complete candidate described by a bounding box in the image plane. The components-based approach suggests the division of the candidate body into several parts over which features are computed. Each pedestrian body part is then independently learnt by a specialized classifier in a first learning stage. The outputs provided by individual classifiers, corresponding to individual body parts, can be integrated in a second stage that provides the final classification output. In section 3.2, two possible methods for developing a second stage classifier are described. As long as a sufficient number of body parts or limbs are visible in the image, the components-based approach can still manage to provide correct classification results. This allows for detecting partially occluded pedestrians whenever the contributions of the pedestrian visible parts are reliable enough to compensate for the missing ones.



(a) (b)

Figure 3.4: (a) Decomposition of a candidate region of interest into 6 sub-regions (b) Sub-regions examples in a set of images.

By using independent classifiers for each body part the learning process is simplified, due to the fact that a single classifier has only to learn individual features of local regions in certain conditions. Otherwise, it would be unrealistic to expect an acceptable learning result using a holistic approach, for the appearance of pedestrians in the scene presents a high intraclass variability (due to lateral and longitudinal

movements, different shapes, pose, clothing, etc). After extensive trials, we propose a total of 6 different sub-regions for each candidate region of interest, which has been re-scaled to a size of 24×72 pixels. This solution constitutes a trade-off between exhaustive sub-region decomposition and the holistic approach. The optimal location of the six sub-regions, empirically achieved after hundred of trials, has been chosen in an attempt to detect coherent pedestrian features, as depicted in Figure 3.4. Thus, the first sub-region is located in the zone where the head would be. The arms and legs are covered by the second, third, fourth, and fifth regions, respectively. An additional region is defined between the legs, covering an area that provides relevant information about the pedestrian pose. This sub-region is particularly useful to recognize stationary pedestrians.

A first comparison is made in order to state the best performing approach among the holistic and components-based options. For this purpose, both the holistic and components-based classifiers were trained and tested using the same set. In particular, the training and test sets were designed to contain 10.000 and 3.670 samples, respectively. These sets were created as subsets of DS and TDS (the same data sets used in Section 3.4). All samples were acquired in daytime conditions.

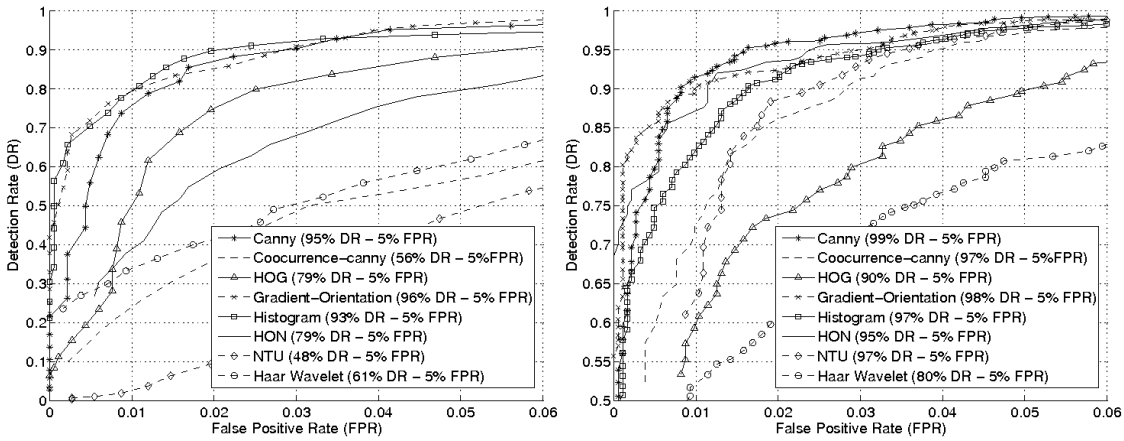


Figure 3.5: ROC curves. (a) Holistic approach. (b) Components-based approach.

As depicted in Figure 3.5, the performance of the holistic approach for all feature extraction methods is largely improved in the components-based approach. In the components-based approach, the outputs of the 6 SVMs corresponding to the 6 candidate sub-regions are combined in a simple-distance classifier, as explained in Section 3.2. Almost every feature extraction method produces an acceptable result in the components-based approach, where the Detection Rate (DR) is between 80% (Haar Wavelet) and 99% (Canny image) at a False Positive Rate (FPR) of 5% for all feature extraction methods. The DR ranges are lesser in the holistic classifier: 48% (NTU), 56% (coocurrence over Canny), 61% (Haar Wavelet), up to a 95% (Canny image), at a FPR of 5%. This shows that breaking the pedestrian into smaller pieces and specifically training the SVM for these pieces reduces the variability and lets

the SVM generalize the models much better. It can then be stated, as previously agreed by other researchers [Papageorgiou 00], [Mohan 01], that the components-based approach clearly outperforms the global classifier.

3.6 Combination of optimal features

In theory, the best performing feature extractor method should be selected in order to implement the final detection system. However, a detailed observation of partial results reveals that some feature extraction methods prove to be more discriminant than others for certain sub-regions, as depicted in Figure 3.6, where the same training and test data sets applied in Sections 3.4 and 3.5 have been also used in this Section. Thus, NTU and Histogram of intensity differences perform the best for head and arms, while HON, Canny, and Histogram of intensity differences seem to perform the best for legs. Similarly, the area between-the-legs is best recognized by NTU. There seems then to be an optimal feature extraction method for each candidate sub-region.

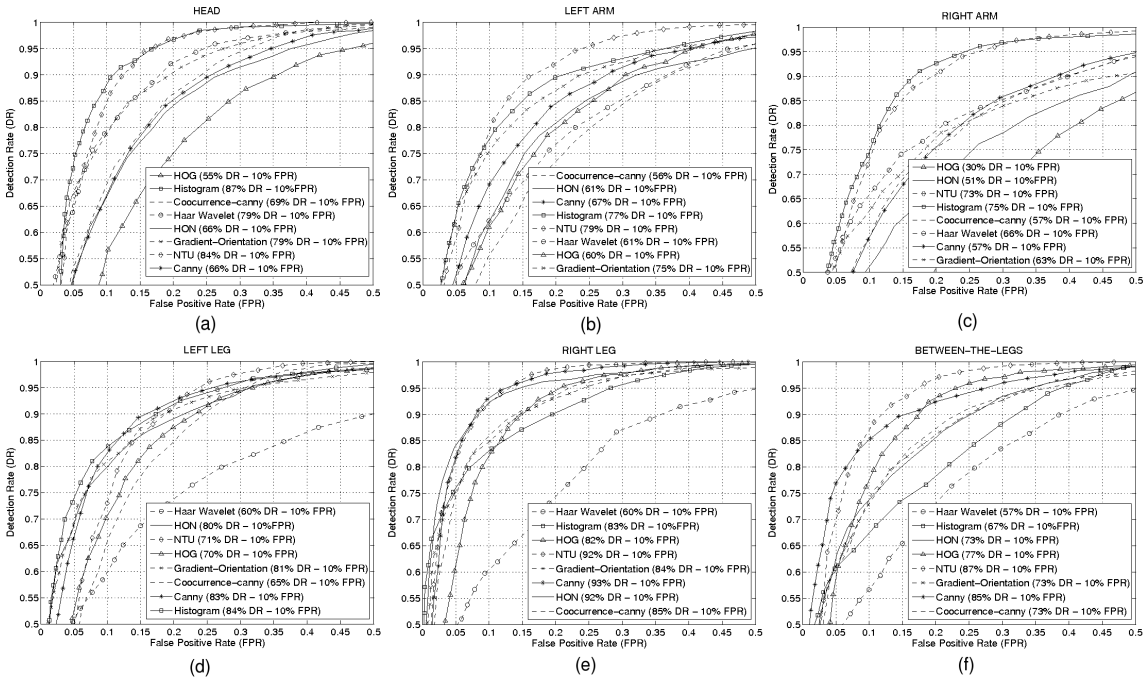


Figure 3.6: ROC curves. (a) Head. (b) Left arm. (c) Right arm. (d) Left leg. (e) Right leg. (f) Between-the-legs.

Accordingly, each candidate sub-region will be learnt separately by an independent classifier. The input to the classifier associated to a given sub-region will be the features vector corresponding to the best performing method for such sub-region. The fine-grain selection of optimal feature extraction methods has been carried out as described next. First, the 3 best performing methods have been selected for each

subregion. Then, the performance difference among the 3 selected feature extraction methods has been evaluated. If there is a method that clearly outperforms the rest of methods it is selected as the optimal method for the sub-region under consideration. Otherwise, a decision is made considering other aspects such as the features vector size. In such a case, the 2 feature extraction methods yielding a smaller vector size are chosen among the 3 best performing ones. According to these parameters, a pre-selection of features is made, with the following result: head - NTU; arms - NTU and Histogram of intensity differences ; legs - HON and Canny; between-the-legs - NTU. An iterative process is started to test the 4 possible combinations using the previously mentioned pre-selected feature extraction methods. The comparison among the results achieved in the 4 experiments yields the final combination of features used in this work: head-NTU, arms-Histogram of intensity differences, legs-HON, between-the-legs-NTU.

The combined used of optimal features leads to a clear increase in the overall classifier performance with regard to individual feature extractors, as depicted in Figure 3.7, where a DR of 99.1% is achieved for a FPR of 2%. These results improve the performance of Canny's detector, which is the best performing feature extractor (in the conditions of the experiment conducted and described in section 3.5), exhibiting a DR of 95% at a FPR of 2%. The optimal combination of feature extraction methods eases the learning stage, making the classifier less sensitive to pedestrian variability.

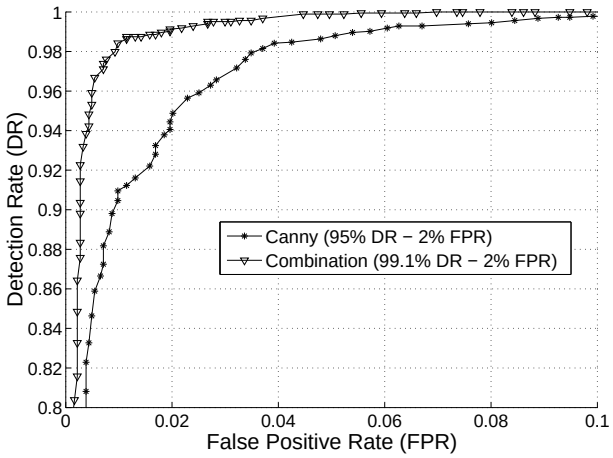


Figure 3.7: ROC curves. Comparison between features combination and Canny's extractor.

3.7 Analysis of the Second-stage Classifier

Another comparison has been studied in order to analyze the influence of the second-stage classifier that combines the information delivered by the six specifically trained SVM models. In a first approach, we have used a simple distance criterion (i.e. distance to the hyperplane separating pedestrians from non-pedestrians) that computes

the addition of the 6 first-stage SVM outputs and then decides the classification by setting a threshold. Another option has been tested by training a two-stage SVM (2-SVM). Once again, the same training and test sets as in Sections 3.4, 3.5 and 3.6 were used in this experiment. The results achieved up to date show that the simple-distance criterion clearly outperforms the 2-SVM classifier, as depicted in Figure 3.8 where a comparison between the both methods is shown when optimal feature extraction methods are applied (see Section 3.6). Thus, the simple-distance classifier exhibits a DR of 99.1% at $FPR = 2\%$, while the performance of 2-SVM classifier is $DR = 30\%$ for the same $FPR = 2\%$. As a consequence of this, the combined use of components-based optimal feature extracion methods in a second distance-based classifier is proposed as a reliable, sufficient solution for single frame pedestrian classification.

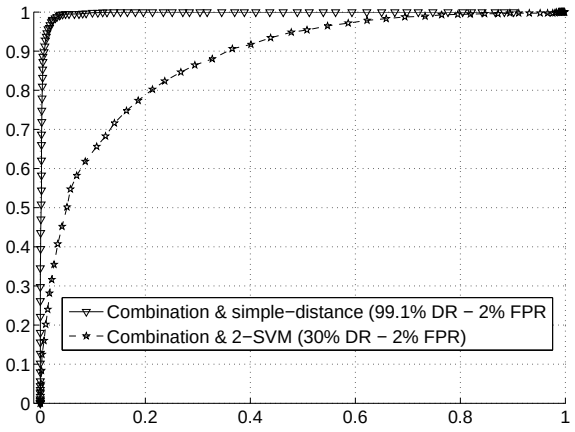


Figure 3.8: ROC curves. Comparison between simple-distance classifier and two-stage SVM.

3.8 Effect of illumination conditions

The need of separate training sets for day and night is analyzed in this section. Nighttime samples were acquired only in illuminated urban and non urban environments, where pedestrian detection remains feasible under distances between 15-20 m. Non-illuminated areas have not been considered in this analysis since pedestrian detection would not be possible beyond a few meters (6-8 m), and infrared cameras should be needed.

A SVM classifier was trained using set G (containing all daytime samples) and tested using set TN. Next, a different SVM classifier was trained using set N (nighttime samples) and tested using the same set TN. The purpose of this experiment is to analyse the performance of nighttime classification using a global daytime classifier.

The observation of Figure 3.9(a) reveals that nighttime pedestrian detection exhibits a high performance when training is carried out using a database containing nighttime samples. Thus, the DR is between 74% (Haar Wavelet) to 94% (Gradient magnitude and orientation) for all feature extraction methods at a FPR of 10%. Figure 3.9(b) shows that nighttime pedestrian detection is not accurate when training is carried out using daytime samples (DR is between 11% for the Histogram of intensity differences to 42% for Gradient magnitude and orientation at a FPR of 10%). In such a case, only two of the proposed feature extraction methods (NTU and HOG) exhibit acceptable operation, which can be explained by the strong normalization process carried out by both methods. Nevertheless, in a first approach, our conclusion is that separate training sets for daytime and nighttime is definitely advisable for optimal classification. Illumination conditions are too different between day and night, making difficult to maintain the same training set and the same classifier for all cases, since generalization becomes a really complex problem.

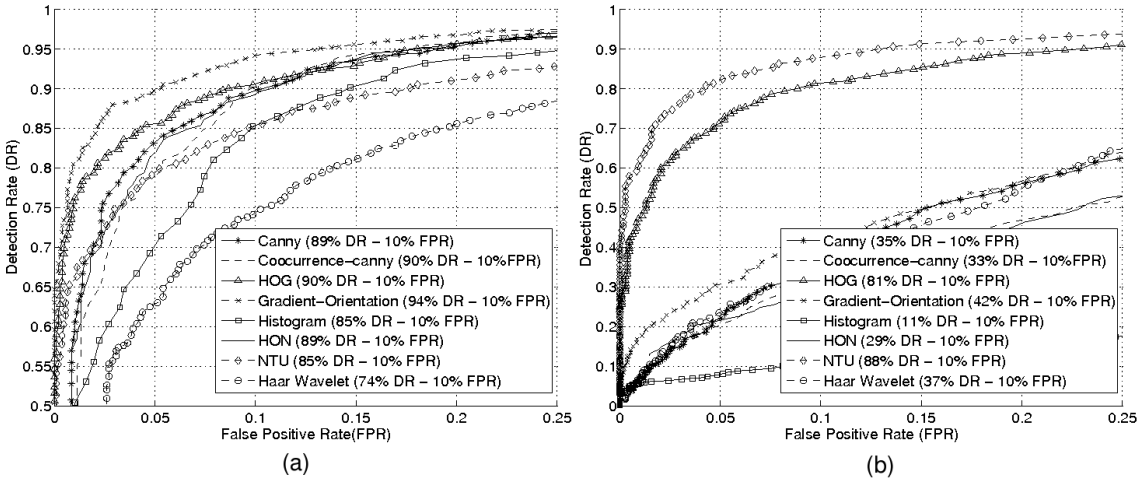


Figure 3.9: ROC curves for nighttime pedestrian detection. (a) Classification of nighttime test samples using training set N (nighttime samples). (b) Classification of nighttime test samples using training set G (daytime samples).

3.9 Effect of the distance and candidate size

The need of separate training sets depending on the different candidate size is analyzed in this section. The separation between short-distance and long-distance pedestrian detection has been empirically set to 12 m. Three different SVM classifiers were trained using sets DS, DL and G, respectively. The trained classifiers were tested against sets TDS and TDL. The purpose of this experiment is to test the necessity or convenience of training separate classifiers for short and long range at daytime.

In a first step all three classifiers were tested using TDS in order to demonstrate the influence of specialized classifiers in daytime short-range classification. The results

are illustrated in Figure 3.10(a). In a second step, the test is repeated using this time TDL so as to check how daytime long-range classification gets affected by learning specialization. Figure 3.10(b) shows the results of this experiment. In both cases, the tests are executed using the optimal combination of features described in Section 3.6.

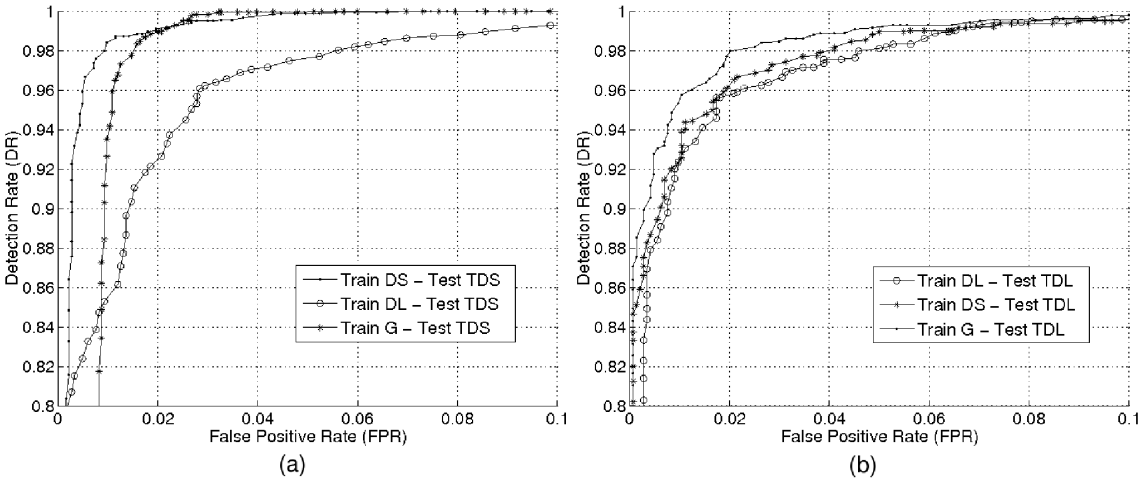


Figure 3.10: ROC curves for daytime pedestrian detection. (a) Pedestrian detection at short distance ($\leq 12m$). (b) Pedestrian detection at long distance ($\geq 12m$).

As can be observed in Figure 3.10(a), the classifier specialized in short-distance pedestrians exhibits only a bit better performance than the rest. Thus, the detection rate for DS-based classifier (SVM classifier trained using set DS) is higher than the detection rate for G-based classifier for a False Positive Rate (FPR) bellow 2%, while the G-based classifier performs better for FPR greater than 2%. Similarly, the results depicted in Figure 3.10(b) show that the G-based classifier clearly outperforms the rest of classifiers for long-distance pedestrian detection. Despite the fact that short-distance pedestrian detection is slightly improved by using separate training sets, our conclusion, contrary to the initial intuition, is that a single SVM classifier trained with a single database containing all types of pedestrians at short and long distance proves to be more effective than separate classifiers for short and long distance, respectively. Let us state clear that this statement remains applicable only for daytime pedestrian detection.

3.10 Effect of bounding box accuracy

The accuracy exhibited in bounding candidates is limited for all the possible candidate selection methods. Although this topic is usually not considered by most authors, we have state its importance. In [Shashua 04] a monocular attention mechanism generates up to 75 windows per frame which are fed to the classifier as potential pedestrians. Their training data set contains a large amount of samples, whose neg-

atives were automatically generated using their attention mechanism. It means, as they say, that the negatives are not randomly generated. We want to stress that not only the negatives samples should be generated by the attention method, but also the positive samples, in order to enhance the single frame performance.

As we saw in Chapter 2 the candidate selection method yields generic obstacles with a 3D shape similar to pedestrians. The 2D candidates are selected by projecting the 3D points over the left image and computing the box that bounds these points. Two bounding box boundaries are defined, one for maximum width and height, and one for minimum ones, taking into account people taller than 2 m and kids with a height lower than 1 m. The 3D candidate position is always known thanks to the stereo candidate selection approach (subtractive clustering) whose outputs are the 3D cluster center coordinates, but the 2D bounding box could not be perfectly fixed due to several effects: body parts which are partially occluded or camouflaged with the background, 3D close objects which have been subtracted along with a pedestrian (for example, pedestrians beside traffic signals, trees, cars, etc.), low contrast pedestrians defined with a few number of 3D points, etc. These badly bounded pedestrians will be classified as non pedestrians if the positives samples used to train the classifier were well-fitted. Figure 3.11 depicts some examples of bad-fitted candidates. Note that this problem also appears with 2D candidate selection mechanisms [Shashua 04], [Mohan 01] with the main drawback of losing the actual pedestrian depth due to monocular constraints.

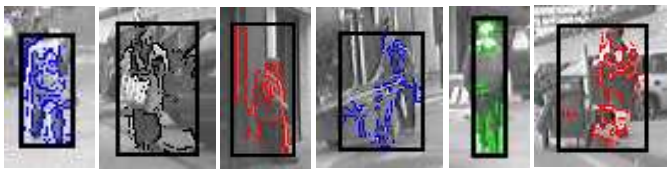


Figure 3.11: Some bad-fitted candidates.

Two strategies are proposed to solve the "badly bounded effect" and both of them need to know the 3D pedestrian depth without monocular lacks. Both strategies are described in the following sections.

3.10.1 Off-line study of the bounding box effect

This approach consists in training the classifier with also bad-fitted pedestrians in an attempt of absorbing the extra information due to larger bounding boxes and the loss of information due to smaller ones. In other words, training the classifier with the positive samples yielded by the candidate selection method. For that purpose, it is necessary to execute the candidate selection process with an off-line validation to distinguish pedestrians from non-pedestrians. In [Shashua 04] and [Mohan 01] the same procedure is only applied for non-pedestrian samples.

With the purpose of showing the influence of the bounding box accuracy in the global performance, we devised an experiment in which a SVM classifier was off-line trained using a training set F of 3000 well fitted or tightly bounded candidates (i.e., the bounding box of candidates fits the real position of the corresponding pedestrians in the image), while a different SVM classifier was off-line trained using a training set B containing 2000 badly bounded candidates. Next, the system is evaluated using a test set TB containing 1000 off-line badly bounded candidates. Figure 3.12(a) depicts the performance obtained after testing a set of badly bounded samples using a classifier trained on badly bounded samples. Figure 3.12(b) shows the performance obtained after testing a set of badly bounded samples using a classifier trained on well fitted ones.

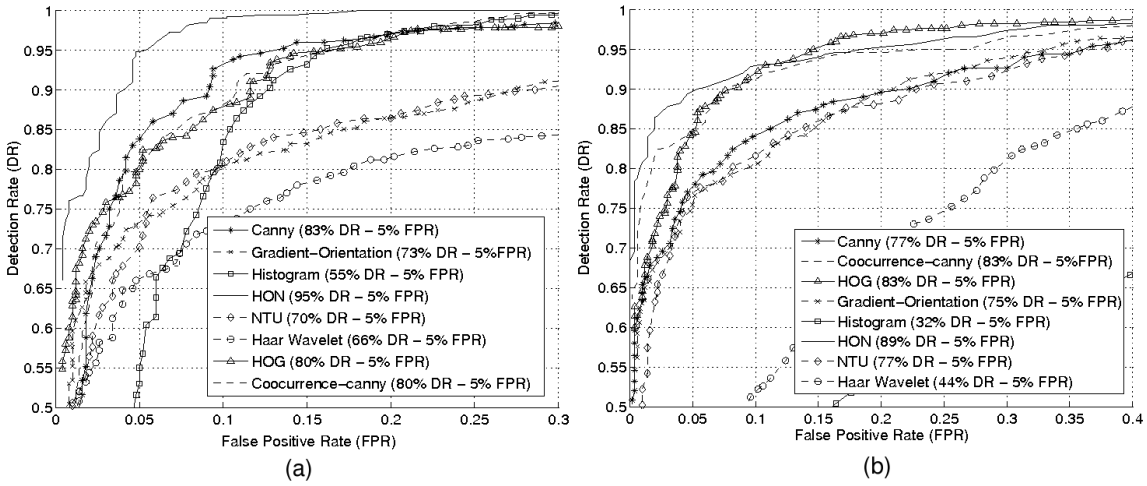


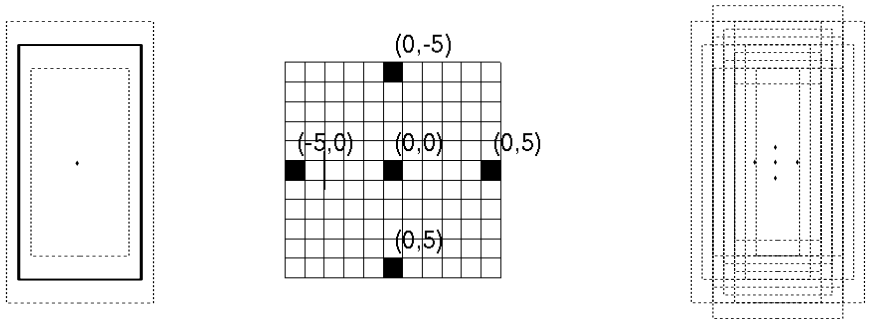
Figure 3.12: Off-line Receiver Operating Characteristic (ROC) for bounding box accuracy. Classification of badly bounded samples (TB) using (a) training set containing badly bounded samples (B) and (b) using training set containing only well-fitted samples (F).

Practical results show that the performance of every feature extraction method has a remarkable decrease when no badly bounded samples are used for the training process. In Figure 3.12(b) most of the methods exhibit much worse figures as long as none of the proposed extractors succeeds in providing a detection rate (DR) above 89% (for the case of HON, which is the best performing one) at a false positive rate (FPR) of 5%. Therefore it is clear that every feature extraction method reduce their performance when the selected candidates are not exactly bounded as the candidates used in the training step, which demonstrates that not only the negatives samples should be generated with the attention mechanism, but also the positive samples in order to enhance the single frame performance.

3.10.2 Multicandidate (MC) generation

By using a well-fitted trained classifier, we propose to perform a multi-candidate generation for every extracted candidate, trying to hit the target and add redundancy.

Three window sizes are defined: the one generated by the candidate selection method, a 20% oversized window and a −20% downsized one as can be observed in Figure 3.13(a). These three windows are moved 5 pixels for each direction: top, down, left and right (see Figure 3.13(b)). Thus a total of 15 candidates are generated as we can see in Figure 3.13(c).



(a) (b) (c)

Figure 3.13: Multicandidate (MC) generation approach: (a)Oversized and downsized windows ; (b) Spatial centers for each window ; (c) 15 candidates generated.

The multi-candidate strategy yields a pedestrian when more than a predefined amount of candidates have been classified as pedestrians. Let S_i represents the classifier error $S_{distance-based}$ for each one of the 15 candidates ($i = 1, \ldots, 15$). Then:

$$\begin{cases} \text{if} & S_i \geq T_{SVM} \Rightarrow D_i = 1 \\ \text{else if} & S_i < T_{SVM} \Rightarrow D_i = 0 \end{cases} \tag{3.5}$$

The final result will be defined by the addition of D_i, i.e.:

$$R_{OUT} = \sum_{i=1}^{15} D_i \tag{3.6}$$

Specifically, the multicandidate approach yields a pedestrian when more than 5 candidates have been classified as pedestrians. This number has been defined after comprehensive experiments. Then:

$$\begin{cases} \text{if} & R_{OUT} \geq 5 \Rightarrow \text{pedestrian} \\ \text{else if} & R_{OUT} < 5 \Rightarrow \text{non-pedestrian} \end{cases} \tag{3.7}$$

On average the candidate selection mechanism generates 6 windows per frame, which yields a total of 90 candidates per frame, after the multi-candidate process. In case the attention mechanism generated more generic candidates per frame this approach would become impractical.

3.11 Tracking and multiframe validation

Each candidate is tracked by using a linear Kalman filter. Thus a softer spatial estimation is achieved. It is necessary to define a method to associate the data at the current time t_i with the data at time t_{i-1}. Finally a multiframe validation process has to be defined in order to provide an estimate of pedestrian detection certainty over time. These three steps are summarizing in the following sections.

3.11.1 Tracking by means of a Kalman filter

Spatial estimates of the detected pedestrians are given by a linear Kalman filter. Tracking is done in 3D space. The state vector is composed of next elements:

$$\vec{s_i^k} = \{x_i^k, y_i^k, z_i^k, w_i^k, h_i^k, \dot{x}_i^k, \dot{y}_i^k, \dot{z}_i^k, \dot{w}_i^k\}^T \tag{3.8}$$

where i and k represents the frame and the number of candidate respectively. This vector contains the 3D pedestrian position (x_i^k, y_i^k, z_i^k) (indeed, the position of the mass center), the pedestrian width (w_i^k) and height (h_i^k), the 3D relative velocity between the car and the target pedestrian $(\dot{x}_i^k, \dot{y}_i^k, \dot{z}_i^k)$ and the 1D relative velocity of the width of the pedestrian $(\dot{w}_i^k)$. The model equations are defined as follows:

$$\begin{bmatrix} x_i^k \\ y_i^k \\ z_i^k \\ w_i^k \\ h_i^k \\ \dot{x}_i^k \\ \dot{y}_i^k \\ \dot{z}_i^k \\ \dot{w}_i^k \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \Delta_t & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & \Delta_t & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & \Delta_t & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \Delta_t \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} . \begin{bmatrix} x_{i-1}^k \\ y_{i-1}^k \\ z_{i-1}^k \\ w_{i-1}^k \\ h_{i-1}^k \\ \dot{x}_{i-1}^k \\ \dot{y}_{i-1}^k \\ \dot{z}_{i-1}^k \\ \dot{w}_{i-1}^k \end{bmatrix} + \vec{w}_i \tag{3.9}$$

where Δ_t is the sampling period, and $\vec{w}_i$ is the noise vector related to the system dynamics. The measurement vector is defined by the 3D pedestrian position, the pedestrian width and the pedestrian height, i.e.:

$$\vec{m_i^k} = \{x_i^k, y_i^k, z_i^k, w_i^k, h_i^k\}^T \tag{3.10}$$

The measurement equation that relates the measurement vector m_i^k as a function of the state of the system s_i^k is then given the following expression:

$$\begin{bmatrix} x_i^k \\ y_i^k \\ z_i^k \\ w_i^k \\ h_i^k \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} . \begin{bmatrix} x_{i-1}^k \\ y_{i-1}^k \\ z_{i-1}^k \\ w_{i-1}^k \\ h_{i-1}^k \\ \dot{x}_{i-1}^k \\ \dot{y}_{i-1}^k \\ \dot{z}_{i-1}^k \\ \dot{w}_{i-1}^k \end{bmatrix} + \vec{v}_k \tag{3.11}$$

where $\vec{v}_k$ represents the noise vector related to the accuracy of the measurements. The noise vector related to the system $\vec{w}_k$ and the noise vector related to the measurements $\vec{v}_k$ are mutually independent and they are defined as white noise with zero mean, and normal distributions ($\vec{w}_k \sim N(0, Q)$ and $\vec{v}_k \sim N(0, R)$).

3.11.2 Data association by means of Mahalanobis distance and ZNCC matching

The fundamental problem of target tracking is measurement association. Data association techniques assign a probability for each one of the N measurement selected at frame t_i and use these probabilities to weight these measurements for update the M candidates at frame t_{i-1}. We propose the use of a merge function based on two indicators. The firs one is founded on the Mahalanobis distance [Mahalanobis 36] between the prediction of the state $\bar{s}_i^k$ (only the first five elements) and the measurement vector m_i^l at a frame t_i, i.e.:

$$d_{M_{kl}} = exp\left(-\frac{1}{2}(\bar{s}_i^k - m_i^l)^T C^{-1}(\bar{s}_i^k - m_i^l)\right) \tag{3.12}$$

where C is the covariance matrix which is defined by next expression:

$$C = \begin{bmatrix} \sigma_x^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_y^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_z^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_w^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_h^2 \end{bmatrix} \tag{3.13}$$

Variances are defined according to the pedestrian and vehicle dynamics, as well as depth accuracy of the stereo system and pitch angle variations. Correspondingly we define $\sigma_x = \sigma_w = r_{ax}/2$ and $\sigma_h = r_{ay}/4$. The z component is defined as from the

same adaptive value used for subtractive clustering, i.e., $\sigma_z = r_{az} = 2z_i^2/(f_x B + z_i)$. Finally y variance is defined according to r_{ay} and pitch variations. Let $\Delta_\alpha = \alpha_i - \alpha_{i-1}$ represents the pitch variation between two consecutive frames. Then $\sigma_y = r_{ay}/8 + Z_i tan(\Delta_\alpha)$.

Along with the Mahalanobis distance we propose the use of box areal overlapping, that is, matching regions of interest of candidate k with the measurement l. In order to get the box that bounds the candidate k at frame t_i the state prediction $\bar{s}_i^k$ is used. Both regions of interest are resized to a fixed size of 24×72. Then the Zero mean Normalized Cross Correlation (ZNCC) is computed for the center point (12,36) as follows:

$$d_{ZNCC_{kl}} = \frac{\sum_{i=-12}^{11} \sum_{j=-36}^{35} A \cdot B}{\sqrt{\sum_{i=-12}^{11} \sum_{j=-36}^{35} A^2 \sum_{i=-12}^{11} \sum_{j=-36}^{35} B^2}} \tag{3.14}$$

where A and B are defined by the following equations:

$$A = (I_{\bar{s}_i^k}(12+i, 36+j) - \overline{I_{\bar{s}_i^k}}) \tag{3.15}$$

$$B = (I_{m_i^l}(12+i, 36+j) - \overline{I_{m_i^l}}) \tag{3.16}$$

where $I_{\bar{s}_i^k}(u,v)$, $I_{m_i^l}(u,v)$ are the gray level intensity of pixel (u,v) in both candidate region of interest and measurement region of interest, respectively, and $\overline{I_{\bar{s}_i^k}}$, $\overline{I_{m_i^l}}$ are the intensity average for those points in both ROIs.

In order to get the final probability of associating candidate k with measurement l both elements are linearly combined according to next equation:

$$d_{kl} = \eta d_{M_{kl}} + (1-\eta) d_{ZNCC_{kl}} \tag{3.17}$$

In practice we have noted that Mahalanobis distance is more reliable than the ZNCC matching because of the loss of information due to resizing. Then we have defined η =0.6. The distance d_{kl} is normalized by default, so that, it can be observed as a probability value. In order to associate a candidate with a measurement a minimum of 0.7 has to be meet. All the distances between candidates and measurements are computed and stored in a table like the one depicted in Table 3.2. The process can be summarized as follows: the maximum distances are looked for each row and column, only when $d_{kl} \geq 0.7$, each measurement is then assigned with each candidate, new

candidates are generated as from non-matching measurements, and old candidates with no matching measurement are erased. This process is carried out for the three possible cases: (i) $M = N$, (ii) $M > N$ and (iii) $M < N$.

t_{i-1} \ t_i	1	2	...	N
1	d_{11}	d_{12}	...	d_{1N}
2	d_{21}	d_{22}	...	d_{2N}
...	...	...	...	...
M	d_{M1}	d_{M2}	...	d_{MN}

Table 3.2: Global matching between measurements (N) at the current frame t_i and candidates (M)at the last frame t_{i-1}.

3.11.3 Probabilistic multiframe validation

Multi-frame validation is needed to endow the tracking system with robustness. The use of Bayesian probability is proposed to provide estimates of pedestrian detection certainty over time. In few words, a sort of low-pass filter has been designed based on Bayesian probability. The process is divided in two stages: pre-tracking and tracking. Newly detected pedestrians enter the pre-tracking stage. Only after consolidation in the pre-tracking stage they start to be tracked by the system. If the candidate is considered as a new pedestrian, it is annotated in the tracked pedestrians list as a new element denoted by k, and its probability of being a pedestrian $P(k_i)$ is initialized according to the classification value given by the SVM classifier at frame t_i ($S_{distance-based,k,i}$), that is:

$$P(k_i) = \begin{cases} 1.0, & \text{if } f(D_i^k) > 1.0 \\ 0.0, & \text{if } f(D_i^k) < 0.0 \\ f(D_i^k), & \text{otherwise} \end{cases} \tag{3.18}$$

where $f(D_i^k) = 0.5 + D_i^k$, being $D_i^k = S_{distance-based,k,i} - T_{SVM}$. In case we were using the multicandidate (MC) approach D_i^k is obtained as a function of the classifier result. If more than 5 candidates have been classified as a pedestrian, D_i^k is computed by using the average value for each one of the positive samples. In case MC would yield a non pedestrian, D_i^k is then computed by using the average for the 15 candidates.

A pedestrian entering the pre-tracking stage must be validated in several iterations before entering the tracking stage. The algorithm followed to implement pedestrian validation during pre-tracking is described in the following lines.

1. Let $\bar{s}_i^k$ represents the predicted position of the pre-validated pedestrian k at frame t_i, and m_i^k represent the associated measure at frame t_i after performing

Probabilistic Data Association. The probability of pre-candidate k to be considered as pedestrian at frame t_i, denoted by $P(k_i)$, is given by the following equation:

$$P(k_i) = P(k_i/k_{i-1})P(k_{i-1}) = K \cdot f(D_i^k)d_{kk}(\bar{s}_i^k, m_i^k)P(k_{i-1}) \tag{3.19}$$

where K is a normalizing factor.

2. The pre-candidate is validated as pedestrian when its probability is above 0.5 during 3 consecutive iterations. Once a pre-candidate is validated pre-tracking stops and tracking starts.

The same condition applies during tracking, i.e., tracking of a pedestrian stops if its probability is below 0.5 during 7 consecutive iterations. The implementation of the multi-frame validation and tracking algorithm described in this section permits to achieve a compromise between robustness in new pedestrians detections and accuracy in pedestrians tracking.

3.12 Conclusions

We propose the use of Support Vector Machine as a good solution for solving the pedestrian learning problem. The regularization coefficient is fixed to a value C=1.0 which can be considered as a small one in order to achieve real generalization.

Recognition performance depends crucially on the features that are used to represent pedestrians. It is not obvious to know what features are better to be learnt for pedestrian detection and not many authors perform a comparative study. In addition, using a by components approach a new problem appears: what features are better for each component?. In this work we have studied the performance for 8 different feature extraction methods.

We have proved that radial basis function (RBF) performs better than polynomial and sigmoid kernels. The component based approach has been demonstrated to outperform the global classifier in practice. In addition, the combination of different feature extraction methods for different subregions leads to an increase in classifier performance. Accordingly the so-called optimal features have been identified for each subregion and combined in a more discriminant components-based classifier.

Likewise, the effects of illumination conditions and candidate size have been studied. Several training and test sets have been created for empirically demonstrating the suitability of multiple classifiers for daytime and nighttime at short and long ranges, respectively. At nighttime, the use of the pedestrian detection system is limited to well-illuminated areas.

Another important factor, usually disregarded by most authors. is the effect of the candidate bounding box accuracy. We propose the development of a multi-candidate generation strategy, in order to assure that issuance of some well-fitted candidates matches the samples used for training.

The use of the Mahalanobis distance along with ZNCC matching is proposed as a solution for the data association problem. A linear Kalman filter has been used for candidates tracking and the multiframe validation process has been carried out by using a probabilistic approach.

Chapter 4

Implementation and Results

4.1 Global implementation of the system

The complete system has been implemented by using Fire-i and Fire-i 400 Unibrain cameras (see Figure 4.1) [Unibrain. 07] with a resolution of 320×240 pixels, 4.3mm of focal length and an horizontal field of view of 42.25°.

(a) (b) (c)

Figure 4.1: (a) Fire-i camera and serial interconnection. (b) Fire-i 400 camera. (c) Inner electronic [Unibrain. 07].

By using a Firewire two cameras are series-interconnected so that, only one camera is needed to connect with the PC. If we have a laptop with a four lines firewire port we need an external battery as depicted in Figure 4.2. In case we would have a laptop or PC with a six lines firewire port, power supply will be obtained from the port itself as can be observed in Figure 4.3.

A first stereo platform were designed with a sucking disc at the back as can be observed in Figure 4.4(a). This platform is installed onboard the vehicle as depicted in Figure 4.4(b) and it does not allow independent movements between the cameras, so that, extrinsic parameters can not change in time. The baseline is about 30cm

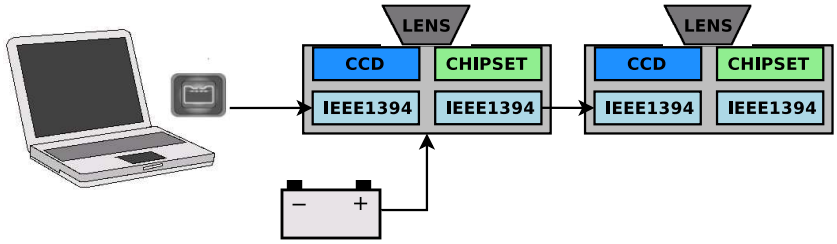


Figure 4.2: Stereo sensor and laptop interconnection with a 4 lines firewire port. An external battery is needed to get the power supply.

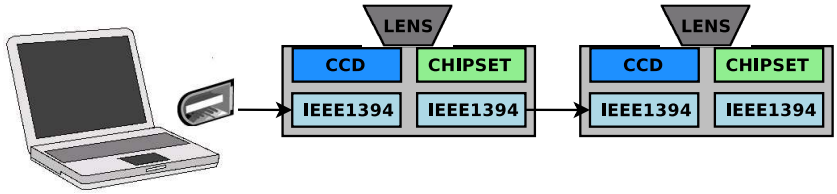


Figure 4.3: Stereo sensor and laptop interconnection with a 6 lines firewire port. Power supply is taken from the firewire port itself.

approximately as a tradeoff between depth accuracy and matching computational cost.

(a) (b)

Figure 4.4: (a) Stereo platform with a sucking disc (b) Stereo platform onboard the experimental vehicle at the University of Alcalá.

Another two different platforms have been used throughout this thesis. The first one was developed at the *Instituto de Automática Industrial* of CSIC (Arganda del Rey, Madrid) in the framework of the AUTOPIA project [AUTOPIA. 07]. This platform is based on Fire-i cameras as can be observed in Figure 4.5(b) and it is installed onboard a Citroen C3 Pluriel prototype as depicted in Figure 4.5(a).

(a) (b)

Figure 4.5: (a) Experimental vehicle Citroen C3 Pluriel. (b) Fire-i cameras based stereo platform used at the IAI of CSIC.

The second platform was developed at the *Centro de Homologación de Vehículos del Instituto Nacional de Técnica Aeroespacial* (INTA, Torrejón de Ardoz, Madrid). In this case we used two Fire-i 400 cameras fixed with an aluminium piece as higher as possible as can be observed in Figure 4.6(b). The experimental vehicle used in this case is a Seat Córdoba equipped with an active hood system and a pedestrian protection airbag (see Figure 4.6(a)).

(a) (b)

Figure 4.6: (a) Experimental vehicle Seat Córdoba used equipped with active hood and pedestrian protection airbag systems (b) Stereo platform with Fire-i 400 cameras used at INTA.

Up to know all the Hardware platforms used to run the system have been PC-based (including laptops). These kind of prototypes are computationally limited, so that, more powerful hardware implementations are needed (FPGAs, DSPs, etc.). Thus the costs of production will be also reduced.

4.2 Global results

4.2.1 Global performance analysis

The performance of the global system is evaluated in a set of sequences recorded in real traffic conditions. Some of the sequences were acquired in urban environments and others in non urban areas. The purpose of this evaluation is to assess the combined operation of the attention mechanism and the SVM-based classifier, including the MC generation strategy, and a multiframe validation stage using Kalman filtering.

The results obtained in the experiments are listed in Table 4.1 (urban environments) and in Table 4.2 (non urban areas). The non urban sequences were recorded in open roads as well as in the Campus of the University of Alcalá. The urban sequences were acquired in Alcalá de Henares and in Madrid. Only pedestrians below 25 m are considered. For each row in both tables the following information is provided: sequence duration, the used method (single or multicandidate), number of pedestrian detected, number of missed pedestrians, number of false alarms (F/A) issued by the system and number of candidates selected in average per frame. Let us remark that the generation of false alarms is also subject to multiframe validation in order to avoid glitches. Accordingly, a false alarm takes place only when a false positive persistently occurs in time. The global system was implemented according to the following features: subtractive clustering candidate selection; components-based SVM using the 6 sub-regions; combination of features ; multiple SVM for day and night-time classification; single and multi-candidate generation to compensate for the bounding box accuracy effect; multi-frame validation using kalman filtering.

Duration	Method	Detected	Missed	False alarms	Candidates per frame
20 min	Single	138	10	9	6
20 min	Multicandidate	143	5	8	6·15

Table 4.1: Global performance evaluated in a set of sequences recorded in urban environments.

Duration	Method	Detected	Missed	False alarms	Candidates per frame
72 min	Single	163	5	5	2
72 min	Multicandidate	166	2	5	2·15

Table 4.2: Global performance evaluated in a set of sequences recorded in non urban environments.

The analysis of results reveals that performance is quite different in urban and non urban environments. Thus, the pedestrian detection system exhibits a ratio of 9 false

alarms in 20 minutes of operation in urban scenarios. This yields a ratio of 27 false alarms per hour. The usual false positives yielded by the system are motorbikes, trees, lampposts, the sides of the cars and other urban furniture. In all false alarm cases, there was a missclassified real object causing the false alarm. Several false positives examples are depicted in Figure 4.7. Similarly, the Detection Rate is 93.24% in urban environments, where 10 pedestrians were missed by the system. Let us clarify the fact that all missed pedestrians were partially occluded or completely out of the vehicle path. Different examples of pedestrian detected by the system in urban environments, where the average number of candidates selected per frame is 6, are depicted in Figure 4.8.

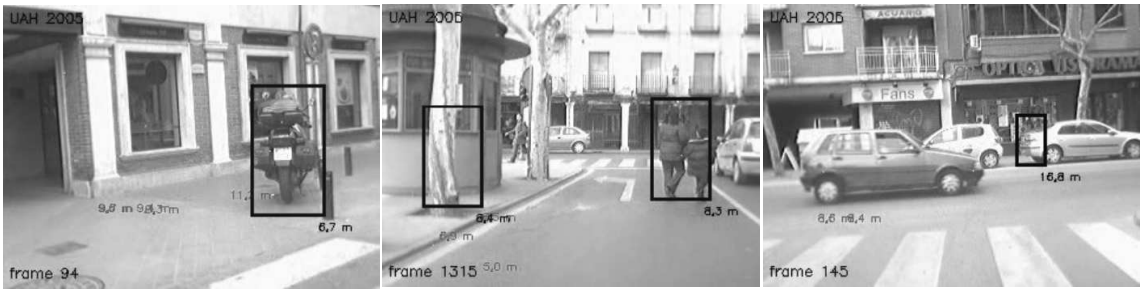


Figure 4.7: Examples of false positive detections in urban environments.

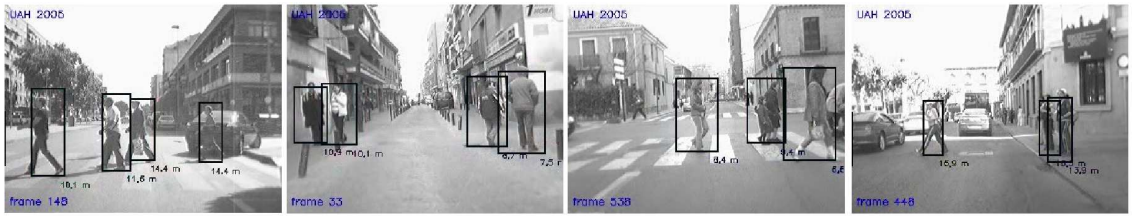


Figure 4.8: Examples of pedestrian detected in urban environments.

Concerning non urban environments, 5 pedestrians were missed by the system in 72 minutes of operation. In all cases, pedestrians were far from the car (20 m or beyond) and wore clothes that produced almost no contrast with the background. This yields a Detection Rate of 97.02% in non urban scenarios, where images are not so heavily corrupted with clutter. Similarly, 5 false alarms occured in the sequences, mainly due to lampposts and trees located by the edge of the road, yielding an average ratio of 4 false alarms per hour. As happens in urban environments, false alarms are caused by real objects. In Figure 4.9 some examples of pedestrians detected in non urban environments, where the average number of candidates selected per frame is 2, are depicted.

Multicandidate approach (MC) produces and increase in the performance for both cases urban and non urban areas. In urban areas 5 missed pedestrians are detected by using MC approach yielding an increase in the detection rate from 93.24% to 96.62%. Similarly in non urban environments the detection rate is improved from 97.02% to 98.81%. Several cases are corrected by this approach. For example, kids,

Figure 4.9: Examples of pedestrian detected in non urban environments.

which are usually selected as candidates with very few points, are better detected by using MC method. When several people are together in the same area the candidate selection method usually yields bounding boxes which fall between two people due to the 3D approach. Thanks to the MC method these pedestrians are well classified. Let us clarify the fact that the other missed pedestrians were partially occluded or completely out of the vehicle path. There is other important effect due to the use of MC classification: pedestrians are classified at larger distances, that is, before in time, with the improvement of increasing the time used in anticipating actions. Figure 4.10 depicts typical images from our test sequences. The number below each bounding box represent range. Right image shows a motorcyclist which is detected as a pedestrian (false positive). In the left image two kids are detected with a correct range.

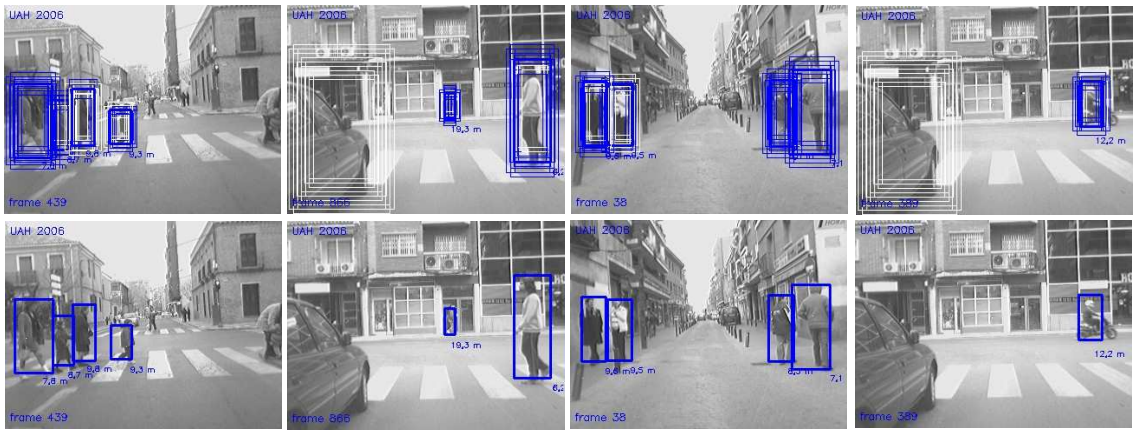


Figure 4.10: Upper row: multi-candidate generation. Lower row: results after classifying the 15 candidates.

Finally we would like to stress that former experiments were carried out in daytime conditions. For nighttime conditions more complex process are needed to evaluate the global performance, since, by one hand, it would be necessary to label the different sequences according to the degree of illumination and, by other hand, candidate selection method and SVM classifier are limited in nighttime conditions whenever visible spectrum cameras were used. In Figure 4.11 several examples in well illuminated areas in nighttime conditions are depicted. As can be observed pedestrian detection is greatly limited in range.

Figure 4.11: Examples of pedestrian detected in well illuminated urban areas in nighttime conditions.

4.2.2 Collision mitigation

Collision mitigation experiments were carried out using a light-weight dummy like the one depicted in Figure 4.12. The car ran over the dummy in several experiments at different velocities bellow 50 km/h. In all cases, the active hood had to be activated at a pre-programmed time before the collision took place, ranging between 200-350ms. The experiments were recorded using a high speed camera providing 1000 frames per second. Thus, the ground truth was obtained from recorded videos given that each frame corresponds to 1ms. With the use of special software applications digital videos recorded with high speed cameras can be played frame by frame in order to precisely determine the exact time between the activation of the active hood and the instant the car-to-dummy collision occurred. This measurement can be done with an accuracy of 1ms.

Figure 4.12: Arrangement used in collision mitigation experiments.

Normally, the specifications for hood activation are in the range 200-350ms. In previous versions of this work, the authors achieved an accuracy of 120ms in the time-to-collision estimation. This figure depends on the vehicle velocity and the number of correlated points for candidates selection. Let us take into account the fact that the exact desired activation time can occur while an image is being processed. Considering that the execution time of the algorithm is in the average of 50ms, and that during that time the system remains *blind*, issuing the activation signal at the end of the frame process would result in a poor resolution activation accuracy that depends on the processing time of the vision-based algorithm. To avoid this problem,

a timer is programmed 2 or 3 frames before the activation time is reached. After that, the process is resumed. The timer interrupts the main process when the programmed time expires. In that moment, the vehicle hood is activated and the vision process is overrun (see Figure 4.13).

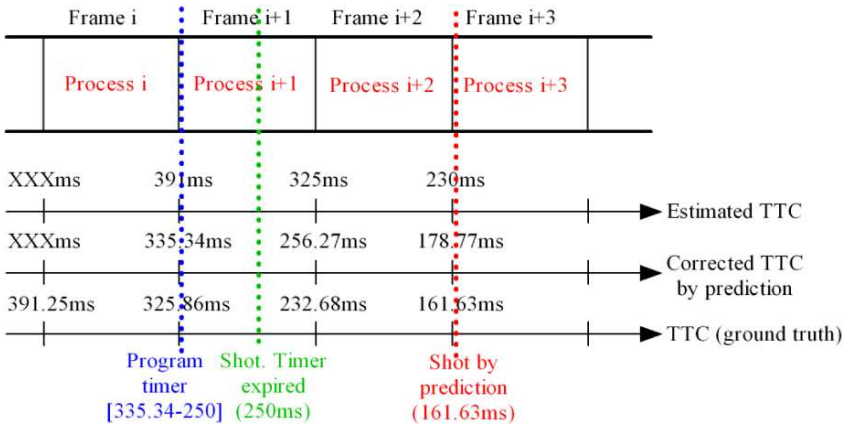


Figure 4.13: Activation time depending on the shot (by prediction or timer expired). The pre-programmed activation time was fixed to 250ms.

After applying the pitch correction method described in this paper, the accuracy in the issuing of the activation signal has been increased to some 50ms (worst case). Table 4.3 shows the results for three collision mitigation experiments at different velocities ranging between 20-50 km/h. The first column represents the vehicle velocity just before the collision. The second column shows the time between the activation of the timer and the real collision. The third column is the desired (pre-programmed) activation time, and the fourth column represents the ground truth provided by the high-speed camera recording the experiment.

Impact velocity	**Timer activation**	**Pre-programmed time**	**Ground truth**
25 km/h	373ms	350ms	350ms
34 km/h	308ms	250ms	235ms
40 km/h	290ms	250ms	250ms

Table 4.3: Time-to-Collision estimation in three different experiments.

Figure 4.14 depicts the result obtained by using the high speed camera and the results obtained by the stereo vision system. The *crash* signal indicates the exact moment when the active systems should be launched. That moment is also recorded by high speed cameras since an illumination signal is turned on in the car side marker light.

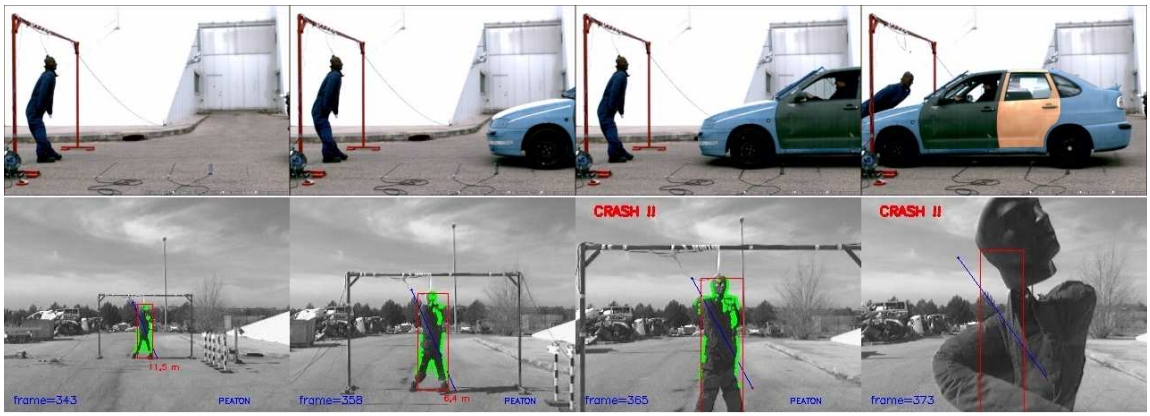
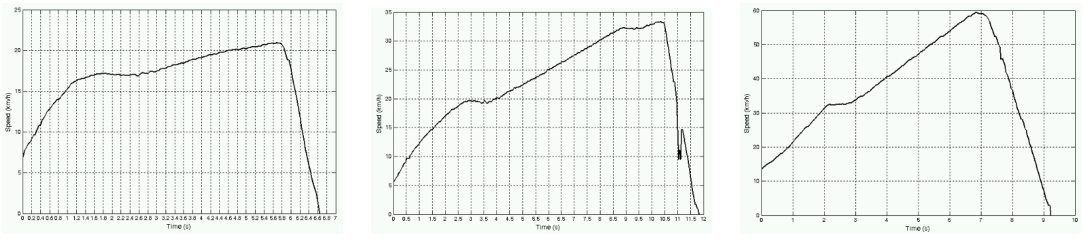


Figure 4.14: Dummy sequence. Upper row: high speed camera point of view. Lower row: inner results of the stereo vision system.

4.2.3 Collision avoidance

Collision avoidance experiments consists on emergency stop manoeuvres by brusquely decelerating the vehicle. Several experiments were conducted to achieve an estimation of the distance that is needed for preventing an accident as a function of the current vehicle speed. Although this computation can be easily done using theoretical equations, we decided to execute real emergency stop manoeuvres and derive a realistic estimation of the stopping time at different velocities. Figure 4.15 depicts the evolution of the vehicle velocity in several experiments where an emergency stop manoeuvre was executed. Based on the previous experiments, a suitable equation for computing the stopping time was estimated using linear regression techniques. As depicted in Figure 4.16 the estimated curve is a reasonable approximation of the real curve. The final expression of the estimated stopping time t_{stop}, including actuator latencies, is provided in next equation:



(a) (b) (c)

Figure 4.15: Emergency stops at different velocities. (a) 21 km/h; (b) 33km/h ; (c) 60 km/h

$$t_{stop} = 0.0003 \cdot x^2 + 0.006 \cdot x + 0.5757 \tag{4.1}$$

where x stands for vehicle velocity at the time of starting the emergency stop ma-

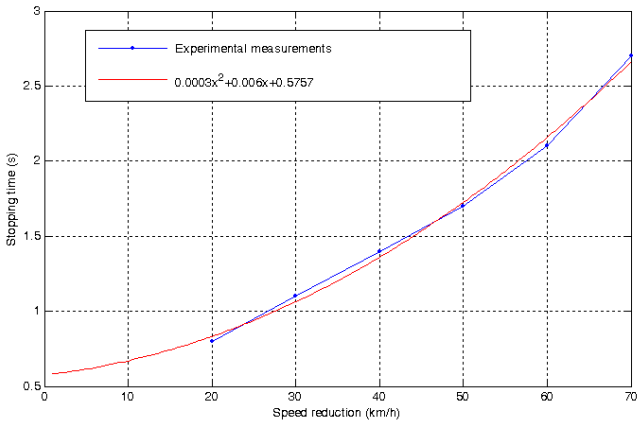


Figure 4.16: Regression curve.

noeuvre. Similarly, the estimated distance required by the vehicle to come to a full stop D_{stop} is provided in the following equation:

$$D_{stop} = 0.5 \cdot (0.0003 \cdot x^2 + 0.006 \cdot x + 0.5757) \tag{4.2}$$

The system was implemented on a G4 Power PC that was installed on a Citroen C3 Pluriel equipped with automatic steering wheel, brake and accelerator pedals. The Citroen C3 was tested on several private circuits for specific experiments, like the ones carried out at the IAI of CSIC (see Figure 4.18) and for live demonstration purposes, including the live demo carried out at the EUROCAST 2007 International Conference at the port of Las Palmas de Gran Canaria (Spain) in February 2007 (see Figure 4.17) and the live demos performed in the framework of the Cybercars project at INRIA [INRIA. 07] and in the framework of the PREVENT European Project [PReVENT. 07] at Versailles (France) in September 2007 (see Figures 4.19(a) and 4.19(a)).

Figure 4.17: Collision avoidance demonstration at the EUROCAST 2007 International Conference at the port of Las Palmas de Gran Canaria (Spain) in February 2007.

Video files showing the global system performance in its different steps, along with the system performance during the fore mentioned demos can be retrieved from ftp://www,depeca.uah.es/pub/vision/pedestrians.

Figure 4.18: Collision avoidance experiment with automatic with automatic steering wheel, brake and accelerator pedals, carried out at the *Instituto de Automática Industrial* of CSIC in Arganda del Rey, Madrid. Upper row: external camera point of view. Lower row: stereo vision system results

(a) (b)

Figure 4.19: (a) Collision avoidance demonstration carried out in the framework of the Cybercars project at INRIA. (b) Stereo vision system demonstration in the framework of the PREVENT European Project. Versailles (France) September 2007.

4.3 Conclusions

The system performs better in non urban areas where images are not so heavily corrupted with clutter. Single frame detection rate is improved by using the multi-candidate generation strategy which absorbs the false negatives that have been bad classified by the single strategy. Several cases are also corrected by this approach. For example, kids, groups of people, etc., are better detected by using MC method. There is other important effect due to the use of MC classification: pedestrians are classified at larger distances, that is, before in time. Obviously the number of classifications needed to accomplish the classification stage becomes 15 times more heavy, from a computational point of view. The stereo vision-based pedestrian detection system has been implemented on different hardware platforms, on different experimental vehicles, succeeding in different safety applications such as collision avoidance and collision mitigation applications.

Chapter 5

Conclusions and future work

To conclude, the next key points should be remarked. First of all, in this thesis we have presented a stereo vision-based system, in the visible spectrum, for pedestrian detection, designing and analyzing each one of the main parts of that kind of systems: candidate selection, classification and tracking. Pedestrian detection seems to be one of the main areas of interest for many automotive industries because of the new European Commission directive which is forcing to introduce safety measures to drastically reduce the number of fatalities of the most vulnerable road users.

The stereo approach allows us to know where the pedestrians are without scale constraints and allows us to reduce the average number of false positives per frame. Non dense 3D maps are created by using edge points. A robust correlation method reduce the amount of stereo-matching errors. Pitch angle was estimated as from both YOZ projection of 3D points and the so-called virtual disparity map. With the help of the pitch compensation process, which needs the 3D information provided by the stereo vision sensor, a correct obstacle/road separation is possible. In addition depth accuracy is increased and the number of points with which pedestrians are detected is improved.

An adaptive subtractive clustering technique has proved to be robust in order to detect generic obstacles with volumes similar to pedestrian. The use of subtractive clustering technique is well known in the area of fuzzy identification models, but its application to obstacles detection in the framework of ITS applications, is presented as a novel and powerful contribution to pedestrian detection.

Several training and test sets with pedestrian and non-pedestrian samples, have been created for empirically demonstrating several classification issues: the suitability of multiple classifiers for daytime and nighttime, the improvements on the classifier performance thanks to the component-based approach, the optimal features selection according to the different components, the fact that multiple models for short and long ranges are not needed, the suitability of radial basis function kernel for SVM classification, etc.

Another important factor, usually disregarded by most authors, is the effect of the candidate bounding box accuracy. We have demonstrated that the classifier performance is decreased when the on-line selected candidates are not bounded in the same way that the candidates used in the training step. Thus it is stated that not only the negatives samples used for training the classifier, should be generated with the attention mechanism, but also the positive samples, in order to enhance the single frame performance. In case we would have a model with well fitted samples we propose the use of the multicandidate generation strategy in order to assure that the issuance of some well-fitted candidates matches the samples used for training.

Although experimental results, which were carried out in different experimental vehicles for different safety applications, show that progress is being made in the right direction, further improvements need to be made before deploying a really robust vision-based pedestrian detection system for assisted driving in real traffic conditions. A FPGA-based hardware implementation of the system should be carried out in order to have a real time system as well as to reduce the costs of production. Infrared cameras have to be used and combined for not illuminated areas in nighttime scenarios. Multi-sensors solutions may be desirable in order to have depth measures faster and with higher accuracy, since active hood systems and pedestrian protection airgabs need as accurate information as possible. Monocular candidates selection methods can be also used (with redundancy) with the aim of improving a bit more the robustness of the stereo vision-based candidate selection mechanism. Adaboost techniques may lead to better classification performances. Finally, additional classifiers should be added to the system (motorbikes, cars, urban furniture and so on) in order to continue breaking the variability of the still huge pedestrian detection problem.

Chapter 6

Publications and projects

6.1 Publications arised from this thesis

- D. Fernández, I. Parra, M. A. Sotelo, L. M. Bergasa, P. Revenga, J. Nuevo, M. Ocaña. Pedestrian recognition for intelligent transportation systems. *In Proceedings of the International Conference on Informatics in Control, Automation and Robotics*. Barcelona, Spain. September 2005.

- I. Parra, D. Fernández, M. A. Sotelo, P. Revenga, L. M. Bergasa, M. Ocaña, J. Nuevo, R. Flores. Pedestrian recognition in road sequences. *In Proceedings of the WSEAS Int. Conf. on Signal Processing, Robotics and Automation*. Madrid, Spain. February 2006.

- I. Parra, D. Fernández, M. A. Sotelo, P. Revenga, L. M. Bergasa, M. Ocaña, J. Nuevo, R. Flores. Stereo Vision-based Pedestrian Recognition for ITS Applications. *In Transactions on Information Science and Applications WSEAS*. vol. 3, no. 3. pp. 554-561. March 2006.

- M. A. Sotelo, I. Parra, D. Fernández, E. Naranjo. Pedestrian detection using SVM and Multi-feature combination. *In Proceedings of the IEEE Intelligent Transportation Systems Conference*. Toronto, Canada. September 2006.

- D. Fernández, I. Parra, M. A. Sotelo, P. Revenga. Bounding Box Accuracy in Pedestrian Detection for Intelligent Transportation Systems. *In Proceedings of the Annual Conference of the IEEE Industrial Electronics Society IECON*. Paris, France. November 2006.

- M. A. Sotelo, D. Fernández, I. Parra, E. Naranjo. Improved Vision-based Pedestrian Detection System for Collision Avoidance and Mitigation. *In Proceedings of the IEEE International Conference on Robotics and Automation. Workshop on Planning, Perception and Navigation for Intelligent Vehicles*. Roma, Italy. April 2007.

- D. Fernández, I. Parra, M. A. Sotelo, P. Revenga, S. Álvarez. 3D Candidate Selection Method for Pedestrian Detection on Non-planar Roads. *In Proceedings of the IEEE Intelligent Vehicles Symposium*. Istanbul, Turkey. June 2007.

- I. Parra, D. Fernández, M. A. Sotelo, L. M. Bergasa, P. Revenga, M. Ocaña, J. Nuevo, M. A. García. Combination of Feature Extraction Methods for SVM Pedestrian Detection. *IEEE Transactions on Intelligent Transportation Systems*. vol. 8, no. 2, pp. 292-307. June 2007.

- D. Fernández, M. A. Sotelo, I. Parra, J. E. Naranjo, M. Gavilán, S. Álvarez. Pitch compensation in pedestrian protection systems for collision avoidance and mitigation. *IEEE Transactions on Intelligent Transportation Systems*. Under review.

6.2 Projects totally or partially arised from this thesis

- *Project Name*: Integración sensorial para asistencia activa a la conducción (ISAAC)
  Funding Entities: CICYT (Spanish Ministry of Science and Technology)
  Duration: from 1/12/2002 to 1/12/2005

- *Project Name*: Herramientas inteligentes de visión artificial para una conducción segura
  Funding Entities: Spanish Ministry of Development
  Duration: from 19/11/2002 to 19/5/2005

- *Project Name*: Red temática GUIA-EX
  Funding Entities: CICYT (Spanish Ministry of Science and Technology)
  Duration: from 1/9/2003 to 1/9/2004

- *Project Name*: Red temática GUIA-EX. Parte II.
  Funding Entities: CICYT (Spanish Ministry of Science and Technology)
  Duration: from 1/9/2005 to 1/9/2006

- *Project Name*: Desarrollo de un sistema binocular embarcado en vehículo experimental para detección anticipada de peatones y mitigación del efecto de los atropellos.
  Funding Entities: Instituto Nacional de Técnicas Aeroespaciales (INTA)
  Duration: from 1/10/2004 to 31/12/2004

- *Project Name*: Desarrollo de un sistema binocular embarcado en vehículo experimental para detección anticipada de peatones y mitigación del efecto de los atropellos. Parte II.
  Funding Entities: Instituto Nacional de Técnicas Aeroespaciales (INTA)
  Duration: from 1/1/2005 to 31/12/2005

- *Project Name*: Desarrollo de un coche de demostración Democar 1
  Funding Entities: Ficosa International S. A.
  Duration: from 1/1/2007 to 31/12/2007

Bibliography

[Autoliv 07] Autoliv, 2007. http://www.autoliv.com/.

[AUTOPIA. 07] AUTOPIA. *Instituto de Automática Industrial del CSIC*, 2007. http://www.iai.csic.es/autopia/.

[Bertozzi 02] M. Bertozzi, A. Broggi, A. Fascioli & P. Lombardi. *Vision-based Pedestrian Detection: will Ants Help?* In Proc. IEEE Intelligent Vehicles Symposium, 2002.

[Bertozzi 03] M. Bertozzi, A. Broggi, R. Chapuis, F. Chausse, A. Fascioli & A. Tibaldi. *Shape-based pedestrian detection and localization.* In Proc. IEEE Intelligent Transportation Systems Conference, 2003.

[Bertozzi 04a] M. Bertozzi, A. Broggi, A. Fascioli, T. Graf & M-M. Meinecke. *Pedestrian Detection for Driver Assistance Using Multiresolution Infrared Vision.* IEEE Transactions on Vehicular Technology, vol. 53, no. 6, pages 1666–1678, November 2004.

[Bertozzi 04b] M. Bertozzi, A. Broggi, A. Fascioli, A. Tibaldi, R. Chapuis & F. Chausse. *Pedestrian Localization and Tracking System with Kalman Filtering.* In Proc. IEEE Intelligent Vehicles Symposium. Parma, Italy, 2004.

[Bertozzi 05a] M. Bertozzi, E. Binelli, A. Broggi & M. D. Rose. *Stereo Vision-based approaches for Pedestrian Detection.* In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.

[Bertozzi 05b] M. Bertozzi, A. Broggi, A. Lasagni & M. D. Rose. *Infrared Stereo Vision-based Pedestrian Detection.* In Proc. IEEE Intelligent Vehicles Symposium, 2005.

[Bertozzi 07] M. Bertozzi, A. Broggi, C. Caraffi, M. Del Rose, M. Felisa & G. Vezzoni. *Pedestrian detection by means of far-infrared stereo vision.* Computer Vision and Image Understanding, vol. 106, no. 6, pages 194–204, 2007.

[Boufama 94] B. Boufama. Reconstruction tridimensionnelle en vision par ordinateur: Cas des cameras non etalonnees. Institut National Polytechnique de Grenoble, France, PhD Thesis, 1994.

[Broggi 00] A. Broggi, M. Bertozzi, A. Fascioli & M. Sechi. *Shape-based Pedestrian Detection.* In Proc. IEEE Intelligent Vehicles Symposium, 2000.

[Broggi 03] A. Broggi, A. Fascioli, I. Fedriga, A. Tibaldi & M. Del Rose. *Stereo-based Preprocessing for Human Shape Localization in Unstructured Environments.* In Proc. IEEE Intelligent Vehicles Symposium, 2003.

[Broggi 04] A. Broggi, A. Fascioli, M. Carletti, T. Graf & M. Meinecke. *A Multi-resolution Approach for Infrared Vision-based Pedestrian Detection.* In Proc. IEEE Intelligent Vehicles Symposium, 2004.

[Brown 03] M. Z. Brown, D. Burschka & G. D Hager. *Advances in computational stereo.* IEEE Transactions on Pattern Analisis and Machine Intelligence, vol. 25, no. 8, pages 993–1008, August 2003.

[Bunschoten 00] R. Bunschoten & B. Krose. *Range Estimation from a Pair of Omnidirectional Images.* In IEEE International Conference on Robotics and Automation, 2000.

[Canny 86] J. Canny. *A Computational approach to Edge-Detection.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 8, no. 6, pages 679–698, November 1986.

[Carrea 00] P. Carrea & G. Sala. *Short Range Area Monitoring for Pre-crash and Pedestrian Protection: the Chameleon and Protector Projects.* In Proc. of the 9th Aachen Colloquium on Automobile and Engine Technology, 2000.

[Chang 01] Chih-Chung Chang & Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/c̃jlin/libsvm.

[Cheng 05] H. Cheng, N. Zheng & J. Qin. *Pedestrian Detection using sparse Gabor filter and support vector machine.* In Proc. IEEE Intelligent Vehicle Symposium, 2005.

[Chiu 94a] S. L. Chiu. *A cluster extension method with extension to fuzzy model identification.* In Proc. of the Third IEEE Conference on Fuzzy Systems. IEEE World Congress on Computational Intelligence., 1994.

[Chiu 94b] S. L. Chiu. *Fuzzy Model Identification based on Cluster Estimation.* Journal of Intelligent and Fuzzy Systems, vol. 2, no. 3, pages 267–278, 1994.

[Christopher 98] J. C. Christopher. *A Tutorial on Support Vector Machines for Pattern Recognition.* In Data Mining and Knowledge Discovery. No. 2, pp. 121-167 Kluwer Academic Publishers.1, 1998.

[Cristianini 00] N. Cristianini & J. Shawe-Taylor. An introduction to support vector machines: and other kernel-based learning methods. Cambridge University Press, 2000.

[Curio 00] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas & W. V. Seelen. *Walking pedestrian recognition.* IEEE Transactions on Intelligent Transportation Systems, vol. 1, no. 3, pages 155–163, September 2000.

[Dalal 05] N. Dalal & B. Triggs. *Histograms of Oriented Gradients for Human Detection.* In International Conference on Computer Vision and Pattern Recognition, pages 886–893, 2005.

[Ewald 00] A. Ewald & V. Willhoeft. *Laser Scanners for Obstacle Detection in Automotive Applications.* In Proc. of the of the IEEE Intelligent Vehicle Symposium, 2000.

[Faugeras 93] O. Faugeras, B. Hotz, H. Mathieu, T. Vieville, Z. Zhang, F. Pascal, E. Theron, L. Moll, Berry G, Vuillemin J, P. Bertin & C. Proy. Real time correlation-based stereo: Algorithm, implementations and applications. INRIA Technical Report, France, 1993.

[Fernández 06] D. Fernández, I. Parra, M. A. Sotelo & P. Revenga. *Bounding Box Accuracy in Pedestrian Detection for Intelligent Transportation Systems.* In Proc. IEEE Industrial Electronics, IECON, 2006.

[Fernández 07] D. Fernández, I. Parra, M. A. Sotelo, P. Revenga, S. Alvarez & M. Gavilán. *3D Candidate Selection Method for Pedestrian Detection on Non-Planar Roads.* In Proc. IEEE Intelligent Vehicles Symposium, 2007.

[Forsyth 03] D.A Forsyth & J. Ponce. Computer vision: A modern approach. Prentice Hall PTR, 2003.

[Franke 98] Y. Franke, D. M. Gavrila, S. Gorzig, F. Lindner, F. Paetzold & C. Wohler. *Autonomous driving goes downtown.* In IEEE Intelligent Systems, vol. 13, no. 6, pages 40–48, December 1998.

[Franke 00] U. Franke & A. Joos. *Real-Time Stereo Vision for Urban Traffic Scene Understanding.* In Proc. IEEE Intelligent Vehicles Symposium, 2000.

[Franke 02] U. Franke & S. Heinrich. *Fast obstacle detection for urban traffic situations.* In Proc. IEEE Intelligent Transportation Systems Conference, 2002.

[Fuerstenberg 02] K. C. Fuerstenberg, K. J. Dietmayer & V. Willhoeft. *Pedestrian Recognition in Urban Traffic using a Vehicle based Multilayer Laserscanner.* In Proc. IEEE Intelligent Vehicles Symposium, 2002.

[Fuerstenberg 05] K. Ch. Fuerstenberg. *Pedestrian Protection using Laserscanners.* In Proc. IEEE Intelligent Transportation Systems Conference. Vienna, Austria, September, 2005.

[Fusiello 97] A. Fusiello, E. Trucco & A. Verri. *Rectification with unconstrained stereo geometry.* In Proc. British Machine Computer Vision, 1997.

[Fusiello 00] A. Fusiello, V. Roberto & E. Trucco. *Symmetric stereo with multiple windowing.* International Journal of Pattern Recognition and Artificial Intelligence, vol. 14, no. 8, pages 1053–1066, December 2000.

[Gavrila 99] D. M. Gavrila & V. Philomin. *Real-Time Object Detection for Smart Vehicles.* In Proc. of International Conference on Computer Vision, 1999.

[Gavrila 00] D. M. Gavrila. *Pedestrian Detection from a Moving Vehicle.* In Proc. of European Conference on Computer Vision, 2000.

[Gavrila 01] D. M. Gavrila, M. Kunert & U. Lages. *A multi-sensor approach for the protection of vulnerable traffic participants -the PROTECTOR project.* In Proc. IEEE Instrumentation and Measurement Technology Conference, 2001.

[Gavrila 04] D. M. Gavrila, J. Giebel & S. Munder. *Vision-based Pedestrian Detection: The PROTECTOR System.* In Proc. of IEEE Intelligent Vehicles Symposium, 2004.

[Gavrila 07] D. M. Gavrila & S. Munder. *Multi-Cue Pedestrian Detection and Tracking from a Moving Vehicle.* International Journal of Computer Vision, vol. 73, no. 1, pages 41–59, June 2007.

[Grubb 04] G. Grubb, A. Zelinsky, L. Nilsson & M. Rilbe. *3D Vision sensing for improved pedestrian safety.* In Proc. IEEE Intelligent Vehicles Symposium. pp. 19-24, 2004.

[Haralick 79] R. M. Haralick. *Statistical and Structural Approaches to Texture.* In Proc. IEEE. vol67, no. 5, pp. 786-804, May, 1979.

[Harris 88] C. Harris & M. Stephens. *A combined corner and edge detector.* In Proc. Fourth Alvey Vision Conference, 1988.

[Hartley 03] R. Hartley & A. Zisserman. Multiple view geometry in computer vision. Cambridge University Press, 2003.

[Hernández 05] C. Hernández. *Sistema de asistencia a la conducción de vehículos de carretera mediante la detección y aviso de salida de carril*. In M.S thesis. Univ. of Alcalá, 2005.

[Hilario 05] C. Hilario, J. M. Collado, J. Ma Armingol & A. de la Escalera. *Pedestrian Detection for Intelligent Vehicles Based on Active Contour Models and Stereo Vision*. Lecture Notes In Computer Science, vol. 3643, pages 537–542, October 2005.

[Hirschmuller 02] H. Hirschmuller, P. R Innocent & J. M. Garibaldi. *Real-time Correlation Based Stereo Vision with Reduced Borders Errors*. International Journal of Computer Vision, vol. 47, no. 1-2-3, pages 229–246, 2002.

[INRIA. 07] INRIA. *Institut National de Recherche en Informatique et en Automatique*, 2007. http://www.inria.fr/.

[Kainulainen 02] J. Kainulainen. *Clustering Algorithms: Basics and Visualization*. In Laboratory of Computer and Information Science. Special Assignement, Helsinki, 2002.

[Kalman 60] R.E. Kalman. *A New Approach to Linear Filtering and Prediction Problems*. In Transactions. ASME Journal of Basic Engineering. vol. 82, no. 1,pp. 35 45, 1960.

[Krotkov 95] E. Krotkov, M. Hebert & R. Simmons. *Stereo perception and dead reckoning for a prototype lunar rover*. Autonomous Robots, vol. 2, no. 4, pages 313–331, December 1995.

[Krotosky 06] S. Krotosky & M. Trivedi. *Multimodal Stereo Image Registration for Pedestrian Detection*. In Proc. of the IEEE Intelligent Transportation Systems Conference, 2006.

[Krotosky 07] S. Krotosky & M. Trivedi. *A Comparison of Color and Infrared Stereo Approaches to Pedestrian Detection*. In Proc. of the IEEE Intelligent Vehicle Symposium, 2007.

[Labayrade 02] R. Labayrade, D. Aubert & J. P. Tarel. *Real Time Obstacle Detection in Stereovision on Non Flat Road Geometry Through V-disparity Representation*. In Proc. IEEE Intelligent Vehicles Symposium, 2002.

[Liu 04] X. Liu & K. Fujimura. *Pedestrian Detection Using Stereo Night Vision*. IEEE Transactions on Vehicular Technolgy, vol. 53, no. 6, pages 1657–1665, November 2004.

[Lucas 81] B. D. Lucas & T. Kanade. *An Iterative Image Registration Technique with an Application to Stereo Vision*. In Proceedings of the 7th International Joint Conference on Artificial Intelligence, pages 674–679, April 1981.

[Mahalanobis 36] P. C Mahalanobis. *On the generalized distance in statistics.* In Proc. of the National Institute of Science of India. 12: 49-55, 1936.

[Mahlisch 05] M. Mahlisch, M. Oberlander, O. Lohlein, D. Gavrila & W. Ritter. *A Multiple Detector Approach to Low-resolution FIR Pedestrian Recognition.* In Proc. IEEE Intelligent Vehicles Symposium, 2005.

[Matlab. 07] Matlab. *Camera Calibration Toolbox for Matlab*, 2007. http://www.vision.caltech.edu/bouguetj/calib_doc/.

[Meinecke 03] M. M. Meinecke, M. A. Obojski, D. Gavrila, E. Marc, R. Morris, M. Tons & L. Letellier. *Strategies in Terms of Vulnerable Road User Protection.* In EU Project SAVE-U - Deliverable D6., 2003.

[Meinecke 05] M. M. Meinecke & M. A. Obojski. *Potentials and Limitations of Pre-Crash Systems for Pedestrian Protection.* In Second International Workshop on Intelligent Transportation. Hamburg, Germany, March, 2005.

[Miled 07] W. Miled, J.C. Pesquet & M. Parent. *Robust Obstacle Detection based on Dense Disparity Maps.* In Proc. Eleventh International Conference on Computer Aided Systems Theory EUROCAST 2007, 2007.

[Mohan 01] A. Mohan, C. Papageorgiou & T. Poggio. *Example-based object detection in images by components.* IEEE Transactions on Pattern Analisis and Machine Intelligence, vol. 23, no. 4, pages 349 –361, April 2001.

[Muhlmann 02] K. Muhlmann, D. Maier, J. Hesser & R. Manner. *Calculating Dense Disparity Maps from Color Stereo Image, an Efficient Implementation.* International Journal of Computer Vision, vol. 47, no. 1-3, pages 79–88, April-June 2002.

[Munder 06] S. Munder & D. M. Gavrila. *An Experimental Study on Pedestrian Classification.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 11, pages 1863–1868, November 2006.

[Nanda 02] H. Nanda & L. Davis. *Probabilistic template based pedestrian detection in infrared videos.* In Proc. IEEE Intelligent Vehicles Symposium, 2002.

[Nedevschi 04] S. Nedevschi, R. Danescu, D. Frentiu, T. Marita, F. Oniga, C. Pocol, R. Schmidt & T. Graf. *High accuracy stereo vision system for far distance obstacle detection.* In Proc. IEEE Intelligent Vehicles Symposium, 2004.

[Nedevschi 06] S. Nedevschi, F. Oniga, R. Danescu, T. Graf & R. Schmidt. *Increased Accuracy Stereo Approach for 3D Lane Detection*. In Proc. IEEE Intelligent Vehicles Symposium, 2006.

[Nuevo 05] J. Nuevo. *TestBuilder Tutorial. Technical Report*, 2005. [Online] Available: ftp://www.depeca.uah.es/pub/vision/SVM/manual.pdf.

[Oren 97] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna & T. Poggio. *Pedestrian Detection Using Wavelets Templates*. In Proc. Computer Vision and Pattern Recognition, 1997.

[Papageorgiou 00] C. Papageorgiou & T. Poggio. *A trainable system for object detection*. International Journal of Computer Vision, vol. 38, no. 1, pages 15 –33, June 2000.

[Parra 07] I. Parra, D. Fernández, M. A. Sotelo, L. M. Bergasa, P. Revenga de Toro, J. Nuevo, M. Ocaña & M. A García Garrido. *Combination of Feature Estraction Methods for SVM Pedestrian Detection*. IEEE Transactions on Intelligent Transportation Systems, vol. 8, no. 2, pages 292–307, June 2007.

[Premebida 06] C. Premebida & U. Nunes. *A Multi-Target Tracking and GMM - Classifier for Intelligent Vehicles*. In Proc. of the IEEE Intelligent Transportation Systems Conference, 2006.

[Premebida 07] C. Premebida, G. Monteiro, U. Nunes & P. Peixoto. *A Lidar and Vision-Based Approach for Pedestrian and Vehicle Detection and Tracking*. In Proc. of the IEEE Intelligent Transportation Systems Conference, 2007.

[PReVENT. 07] PReVENT. *PReVENTive and Active Safety Applications*, 2007. http://www.prevent-ip.org/.

[Scharstein 02] D. Scharstein & R. Szeliski. *A taxonomy and evaluation of dense two-frame stereo correspondence algorithms*. International Journal of Computer Vision, vol. 47, no. 1, pages 7–42, April 2002.

[Se 98] S. Se & M. Brandy. *Stereo Vision-based Obstacle Detection for Partially Sighted People*. In Proc. Asian Conference on Computer Vision, 1998.

[Shashua 04] A. Shashua, Y. Gdalyahu & G. Hayun. *Pedestrian detection for driving assistance systems: single-frame classification and system level performance*. In Proc. IEEE Intelligent Vehicles Symposium. pp. 1-6, 2004.

[Shi 94] J. Shi & C. Tomasi. *Good Features to Track*. In IEEE Conference on Computer Vision and Pattern Recognition, pages 593–600, 1994.

[Siemens 07] Siemens, 2007. http://www.siemensvdo.com/.

[Sobel 68] I. Sobel & G. Feldman. *A 3x3 isotropic gradient operator for image processing.*, 1968. Never published but presented at a talk at the Stanford Artificial Project.

[Sotelo 04a] M.A. Sotelo, F.J. Rodríguez & L. Magdalena. *Virtuous: Vision-based road transportation for unmanned operation on urban-like scenarios.* IEEE Transactions on Intelligent Transportation Systems, vol. 5, no. 2, pages 69–83, June 2004.

[Sotelo 04b] M.A. Sotelo, F.J. Rodríguez, L. Magdalena, L.M. Bergasa & L. Boquete. *A color vision-based lane tracking system for autonomous driving on unmarked roads.* Autonomous Robots, vol. 16, no. 1, pages 95–116, January 2004.

[Sotelo 06] M. A. Sotelo, I. Parra, D. Fernández & E. Naranjo. *Pedestrian Detection Using SVM and Multi-Feature Combination.* In Proc. IEEE Intelligent Transportation Systems Conference, 2006.

[Sotelo 07] M. A. Sotelo, D. Fernández, I. Parra & E. Naranjo. *Improved Vision-based Pedestrian Detection System for Collision Avoidance and Mitigation.* In Proc. IEEE ICRA 2007 Workshop: Planning, Perception and Navigation for Intelligent Vehicles, 2007.

[Stefano 02] L. Di Stefano, M. Marchionni, S. Mattoccia & G. Neri. *A Fast Area-Based Stereo Matching Algorithm.* In 15th IAPR/CIPRS International Conference on Vision Interface, 2002.

[Suard 05] F. Suard, A. Rakotomamonjy, A. Bensrhair & V. Guigue. *Pedestrian Detection using stereo vision and graph kernels.* In Proc. IEEE Intelligent Vehicle Symposium, 2005.

[Suard 06] F. Suard, A. Rakotomamonjy, A. Bensrhair & A. Broggi. *Pedestrian Detection using Infrared images and Histograms of Oriented Gradients.* In Proc. IEEE Intelligent Vehicles Symposium, 2006.

[Suganuma 07] N. Suganuma & N. Fujiwara. *An Obstacle Extraction Method Using Virtual Disparity Image.* In Proc. IEEE Intelligent Vehicles Symposium, 2007.

[Sun 02] C. Sun. *Fast Stereo Matching Using Rectangular Subregioning and 3D Maximum-Surface Techniques.* International Journal of Computer Vision, vol. 47, no. 1, pages 99–117, May 2002.

[Szarvas 05] M. Szarvas, A. Yoshizawa, M. Yamamoto & J. Ogata. *Pedestrian Detection with convolucional neural networks.* In Proc. IEEE Intelligent Vehicle Symposium, 2005.

[Trucco 98] E. Trucco & A. Verri. Introductory techniques for 3-d computer vision. Prentice Hall PTR, 1998.

[Tsuji 02] T. Tsuji, H. Hattori, M. Watanabe & N. Nagaoka. *Development of night-vision system.* IEEE Transactions on Intelligent Transportation Systems, vol. 3, no. 3, pages 1524–9050, September 2002.

[UNECE] UNECE. *Pedestrian Safety Global Technical Regulation Preamble.* In Transport Division. World Forum for Harmonization of Vehicle Regulations.

[Unibrain. 07] Unibrain. *Fire-i Digital Camera*, 2007. http://www.unibrain.com/Products/VisionImg/Fire_i_DC.htm.

[van der Mark 06] W. van der Mark & D. M. Gavrila. *Real-time dense stereo for intelligent vehicles.* IEEE Transactions on Intelligent Transportation Systems, vol. 7, no. 1, pages 38–50, March 2006.

[Vapnik 95] V. Vapnik. The nature of statistical learning theory. Springer Verlag, 1995.

[Viola 03] P. Viola, M.J. Jones & D. Snow. *Detecting Pedestrians using patterns of motion and appearance.* In Proc. IEEE International Conference on Computer Vision, 2003.

[Wang 90] L. Wang. *Texture unit, texture spectrum and texture analysis.* In IEEE Transactions on Geosciences and Remote Sensing. Vol. 28, No 4, pp. 509-512 (90-19), 1990.

[Xu 96] G. Xu & Z. Zhang. Epipolar geometry in stereo, motion and object recognition: A unified approach. Kluwer Academic Publishers, Dordrecht, Boston, London, 1st edition, 1996.

[Xu 05] F. Xu, X. Liu & K. Fujimura. *Pedestrian Detection and Tracking with Night Vision.* IEEE Transactions on Intelligent Transportation Systems, vol. 6, no. 1, pages 63–71, March 2005.

[Yu 03] Q. Yu, H. Araujo & H. Wang. *Stereo-Vision Based Real time Obstacle Detection for Urban Environments.* In Proc. International Conference on Advanced Robotics, 2003.

[Zhao 00] L. Zhao & C. E. Thorpe. *Stereo and Neural Network-Based Pedestrian Detection.* In Proc. IEEE Transactions on Intelligent Transportation Systems, vol. 1, no. 3, pages 148 –154, September 2000.