Region of Interest Generation for Pedestrian Detection using Stereo Vision

Korra Abhishek Chauhan

Department of Computer Science and Engineering
National Institute of Technology Rourkela

Region of Interest Generation for Pedestrian Detection using Stereo Vision

Thesis submitted in partial fulfillment

of the requirements of the degree of

Master of Technology

in

Computer Science and Engineering
(*Specialization*: *Computer Science*)

by

Korra Abhishek Chauhan

(Roll Number: 711CS1122)

based on research carried out

under the supervision of

Prof. Pankaj Kumar Sa

May, 2016

Department of Computer Science and Engineering
National Institute of Technology Rourkela

Department of Computer Science and Engineering
National Institute of Technology Rourkela

Prof. Pankaj Kumar Sa
Professor

May 24, 2016

Supervisor's Certificate

This is to certify that the work presented in the thesis entitled *Region of Interest Generation for Pedestrian Detection using Stereo Vision* submitted by *Korra Abhishek Chauhan*, Roll Number 711CS1122, is a record of original research carried out by him under my supervision and guidance in partial fulfillment of the requirements of the degree of *Master of Technology* in *Computer Science and Engineering*. Neither this thesis nor any part of it has been submitted earlier for any degree or diploma to any institute or university in India or abroad.

Pankaj Kumar Sa

Dedication

Dedicated to friends and family.

Signature

Declaration of Originality

I, *Korra Abhishek Chauhan*, Roll Number *711CS1122* hereby declare that this thesis entitled *Region of Interest Generation for Pedestrian Detection using Stereo Vision* presents my original work carried out as a postgraduate student of NIT Rourkela and, to the best of my knowledge, contains no material previously published or written by another person, nor any material presented by me for the award of any degree or diploma of NIT Rourkela or any other institution. Any contribution made to this research by others, with whom I have worked at NIT Rourkela or elsewhere, is explicitly acknowledged in the thesis. Works of other authors cited in this dissertation have been duly acknowledged under the sections “Reference” or “Bibliography”. I have also submitted my original research records to the scrutiny committee for evaluation of my thesis.

I am fully aware that in case of any non-compliance detected in future, the Senate of NIT Rourkela may withdraw the degree awarded to me on the basis of the present thesis.

May 24, 2016
NIT Rourkela

Korra Abhishek Chauhan

Acknowledgment

I have taken endeavors in this project. Be that as it may, it would not have been conceivable without the kind backing and help of numerous people and associations. I would like to extend my sincere thanks to every one of them.

I am very obligated to Dr. Pankaj K Sa for their direction and steady supervision and for giving vital data with respect to the task and additionally for their backing in finishing the project.

Other than my supervisor, I want to thank the Ph.D Scholar Suman Ku. Choudhury for his wise remarks and support, additionally for the hard question which incented me to enlarge my research from different points of view.

I might want to express my appreciation towards my folks and individual from NIT Rourkela for their kind co-operation and support which help me in finishing of this project.

My thanks and thanks likewise go to individuals who have energetically bailed me out with their capacities.

May 24, 2016
NIT Rourkela

Korra Abhishek Chauhan
Roll Number: 711CS1122

Abstract

Pedestrian detection is an active research area in the field of computer vision. The sliding window paradigm is usually followed to extract all possible detector windows, however, it is very time consuming. Subsequently, stereo vision using a pair of camera is preferred to reduce the search space that includes the depth information. Disparity map generation using feature correspondence is an integral part and a prior task to depth estimation. In our work, we apply the ORB features to fasten the feature correspondence process. Once the ROI generation phase is over, the extracted detector window is represented by low level histogram of oriented gradient (HOG) features. Subsequently, Linear Support Vector Machine (SVM) is applied to classify them as either pedestrian or non-pedestrian. The experimental results reveal that ORB driven depth estimation is atleast seven times faster than the SURF descriptor and ten times faster than the SIFT descriptor.

***Keywords*: *ROI generation*; *ORB*; *adaptive windowing*; *disparity map*; *stereo vision*; *pedestrian detection*.**

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Pedestrian detection has various applications in many fields such as video surveillance, advanced driver assistance system (ADAS), robotic navigation, augmented reality, traffic management and intelligent transport system. Pedestrian detection is a very intriguing task due to many complexities like cluttered backgrounds, change in articulated pose, occlusions between pedestrians, various styles of clothing, presence of other occluding accessories, illumination variation, weather condition and camera motion. Many pedestrian detection methodologies compose of 3 major steps: Region of interest(ROI) generation, feature generation and classification. Tracking and refinement steps are also added to enhance the rate of detection in some cases. The early stage of a pedestrian detection system (ROI generation) does not depend on previous stages and it is the primary component due to one of the most critical requirements. If a pedestrian escapes from the ROI, it is impossible to detect in the next levels. Hence, such ROI should be chosen which doesn't miss any candidate regions. The detection systems are grouped into two categories based on the sensors used to collect data. In the first aggregation, active sensors like lasers and radars are used and in the other, passive vision sensors are used. A major disadvantage of active sensors is that colors and patterns cannot be detected, making recognition of object very difficult. According to the comparative analyses performed by Foix et al. [1], Kim et al. [2], and Zhang et al. [3], active sensor devices perform satisfactorily only up to a maximum distance of approximately 5–7 meters and are too sensitive to be used in outdoor environments, especially in very bright areas. Because of these limitations of active sensors, passive sensors (i.e., vision sensors) are more reliable and robust; they are able to produce high-resolution disparity maps and are suitable for both indoor and outdoor environments. Passive sensors are preferred to active sensors for this reason.

1.1 Motivation

Several methods have been proposed for detecting humans with a static camera. However, pedestrian detection from a moving platform is a more complex task. The challenge of change in articulated pose and clothing can be solved using large data samples with feature representation (currently available) and change in light illumination can be solved by using invariant pixel representation and cluttered background problem using additional texture details. Early pedestrian detection methods used a single camera and techniques such as contour extraction and chamfer distance. However, pedestrian detection and tracking done using monocular cameras and only RGB data is not reliable. There are some monocular detection systems that estimate very good performance, but those methodologies cannot guarantee that the obtained candidates represent real physical objects. They are not range sensors but provide range estimates, applying flat-world assumption. Adding another camera can solve these problems since distance of scene points from camera based on disparity between images can be obtained. The disparity information provides important cues for the ROI selection stage. Using stereo vision provides the possibility to take the advantage of depth of object for detection and tracking.

1.2 Objective

Disparity map is found by solving correspondence problem and then reconstructing the disparity image from the disparity values. In traditional methods, in order to solve the correspondence problem, the images are first rectified using the feature vectors obtained by using the descriptors like SIFT and SURF. But, the computation time of SURF is costly, therefore, some improved algorithm ORB [4] with lower computation cost have been put forward. Determining the features points is the first step of finding the disparity map, which in turn is the first step of ROI generation. In this thesis, we take advantage of ORB(Oriented FAST and Rotated BRIEF) to find the key-points in both the images in order to rectify the images. This helps in increasing the speed of ROI generation.

In conventional mechanisms, pedestrian detections were found by an exhaustive search. However, these mechanism have enormous computational cost and many drawbacks in detection of humans with different scales, lighting fixtures and in the presence of occlusion. Also, due to camera movement, the scale of every object is varying frame to frame and is very necessary to have a method of generating scale invariant ROI to be able to sight objects in each frame. Driven by these observations, an adaptative windowing for generating disparity map using ROI is presented. The regions of interest detected are applied to a function such as Histogram Oriented Gradients and then a linear support vector machines (SVM) for classification.

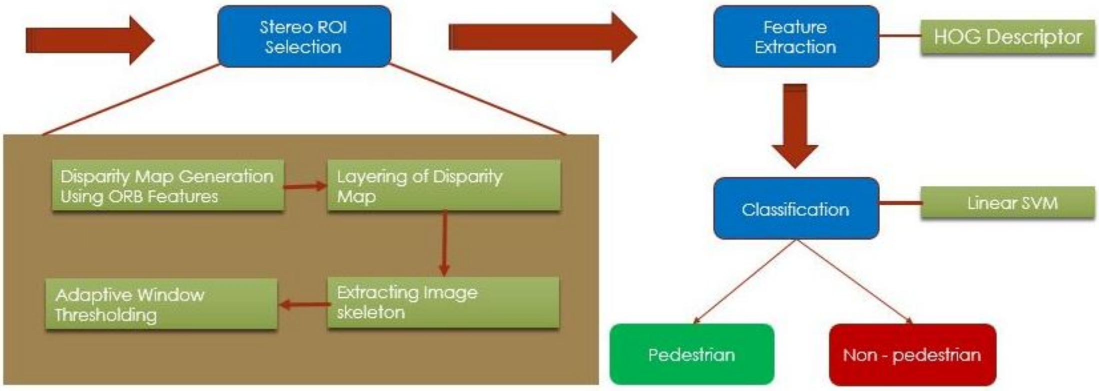

Stereo ROI Selection
Disparity Map Generation Using ORB Features
Layering of Disparity Map
Adaptive Window Thresholding
Extracting Image skeleton
Feature Extraction
HOG Descriptor
Classification
Linear SVM
Pedestrian
Non - pedestrian

Figure 1.1: Framework of pedestrian detection.

1.3 Thesis Layout

The remainder of this thesis is organized as follows:

i. **Chapter 2**
This chapter gives the details of existing methods for Pedestrian detection, their advantages, shortages. This chapter depicts the picture of advancement of technology and improvement in detection algorithms.

ii. **Chapter 3**
This chapter explains the methodology which has been implemented for the generation of disparity map. It gives a clear view on how the ORB algorithm is used for faster determination of the disparity map. The development of block matching techniques SGBM/SBM is presented in this chapter.

iii. **Chapter 4**
This chapter describes the ROI generation from dense depth map obtained. It includes 3 major steps of layering the depth map, extracting the image skeleton and introduces the implementation of adaptive windowing over the skeletons of layered depth map.

iv. **Chapter 4**
This chapter includes the feature extraction and classification from the candidate ROI.

Chapter 2

Literature Review

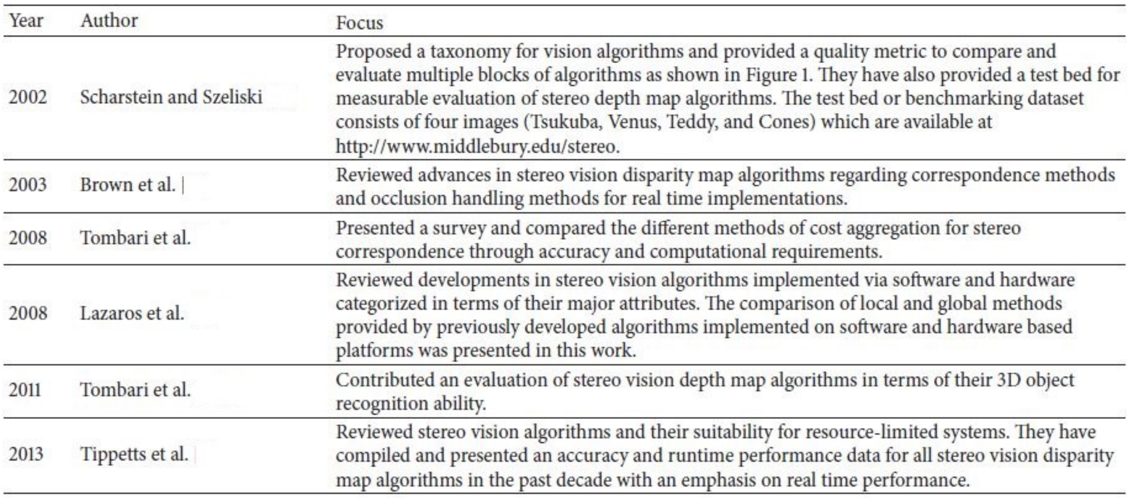
Many interesting approaches to detect pedestrians have followed a basic training approach, by overcoming an attitude of recovery and description of pedestrians visibility in terms of low level signal features of a region of interest (ROI). There are many ways for these region of interest. Background subtractions are prominently used in monocular video surveillance applications, is not suitable for non-static camera. Other methodologies like sliding windows object identification and independent movements based on the detection of music barriers . sliding window mechanism ROI slides the window of all sizes of every possible size over on the every image locations in a frame, as the moving as the moving feature extraction and pattern classification is performed . Method of brute force is combined with strong classifiers [5][6][7] is currently too computations expensive for real-time application. However,[8] showed efficient technical variation sliding weindow ,which includes a cascade identifier with normal features are placed before the waterfall and later high complicated identifiers. In every stage of the detector, AdaBoost[9] select incrementally these features of the less error waited in the training set,till a acceptable and wrong identification rates that user has supplied a set of validation.

One more approach for getting ROIs includes identifying non static objects independently in the monocular images. With this general approach translator movement takes camera and identifies movements in the field of optical flow motion expected back-ground [10][11][12]. A third impact approach for getting ROI was the stereoscopic vision. Zhao and Thorpe [7] get a foreground region of the cluster in the space of disparity. Broggi et al [13] and Grubb et al [14] consider the X and Y projections technical space below disparity “V-disparity” once you have established ROI. Variety of combinations of features and classifier pattern will be make the different between pedestrians and no pedestrians.

Broggi et al.[13] employs vertical symmetry characteristics, Zhao and Thorpe[7] apply a high-pass filter and normalize the return on investment for size.then apply a feedforward neural networks. Papageorgiou and Poggio [6] took advantage of the features of Haar wavelet-and combined with a support vector machine (SVM); This method was later adapted by Elzein et al. [10] and others. components based approaches have been used to reduce the complexity of the detection of pedestrians.

Shashua et al. [16], extracts the features from each of the 9 fixed subregions. Many methodologies attempt to detect particular human parts (part based methods). Mohan et al. (2001) [5] , work of Papageorgiou and Poggio (2000) extends [6] with four heads detection classifiers components, legs, left and right hands separated. The individually obtained results are composed of a second classifier, after ensuring appropriate geometric conditions. Most recently, there have been many attempts to reduce the complexity of sorting, separating the pedestrian training set of subsets (ie, based on the starting address for pedestrians). [16] , Grubb et al. [14] and Shimizu and Poggio [17]. It has no separate comprehensive form on the control of pedestrians. Models as representations of Assets form (Baumberg and Hogg, 1994;. Coote et al, 1995 ;. Philomin et al, 2000); Stenger et al, 2003; Toyama and Blake, 2001), examples of the form and colored dots (Heisele and Wohler, 1998) were occluded with Kalman- (Baumberg and Hogg, 1994; Cootes et al, 1995) or particulate filter (Philomin et al, 2000; .. Stenger et al., 2003; Toyama and Blake, 2001) approaches return the temporary investment given people, former form has found walking sideways to the orientation of view, either using the taco periodicity (Cutler and Davis, 2000; Polana and Nelson 1994) or by the side gait pattern function (Heisele and Wohler, 1998) perpendicular to the formet, there has been increased curiosity in the domain of the FIR (Broggi et al, 2004 learning .. fang et al, 2003; Liu and Fujimura, 2003) driven by the arrival of cheaper cameras, uncooled. pedestrian detection by the heat of your body is attractive, certainly, when examining images taken on a cold winter night, where pedestrians stand out as white areas in front of a black background. However, the situation is less attractive on sunny summer days, when there are a lot of additional hot spots. In the latter case, we must resort to a similar set of detection techniques as in the visible domain.

Figure 2.1 summarizes the main systems of pedestrian detection, distinguished by the type of image, coverage, detection capability, usage tracking, speed of processing and testing equipment, whenever specified by the respective authors sensors. Performance comparisons are dangerous, because the data sets are usually small (and diverse). (Mohan et al., 2001), and Papageorgiou Poggio (2000), Viola et al. (2005), Zhao and Thorpe (2000) often refer to the individual components of calculation (eg classification). Even when test data is more abundant, many important test criteria remain unspecified (eg exact coverage area, location tolerances, data allocation rule). The problem of automatic optimization of parameters of the system so far has not been discussed in the context of pedestrian detection, to our knowledge. A lot of literature on numerical methods dealing with the minimization of nonlinear objective functions. Such approaches are not practical for optimization problem given a complex system because (a) a single evaluation of the objective function is computationally expensive, and (b) the number of parameters involved is relatively high. Since production, the authors have attempted either to model system behavior by the simplest functions, for example, Bayesian networks (Sarkar and Chavali, 2000), or specific solutions developed for the subject matter. Cascaded classifiers (Viola et al., 2005) have recently received



Year	Author	Focus
2002	Scharstein and Szeliski	Proposed a taxonomy for vision algorithms and provided a quality metric to compare and evaluate multiple blocks of algorithms as shown in Figure 1. They have also provided a test bed for measurable evaluation of stereo depth map algorithms. The test bed or benchmarking dataset consists of four images (Tsukuba, Venus, Teddy, and Cones) which are available at http://www.middlebury.edu/stereo.
2003	Brown et al.	Reviewed advances in stereo vision disparity map algorithms regarding correspondence methods and occlusion handling methods for real time implementations.
2008	Tombari et al.	Presented a survey and compared the different methods of cost aggregation for stereo correspondence through accuracy and computational requirements.
2008	Lazaros et al.	Reviewed developments in stereo vision algorithms implemented via software and hardware categorized in terms of their major attributes. The comparison of local and global methods provided by previously developed algorithms implemented on software and hardware based platforms was presented in this work.
2011	Tombari et al.	Contributed an evaluation of stereo vision depth map algorithms in terms of their 3D object recognition ability.
2013	Tippetts et al.	Reviewed stereo vision algorithms and their suitability for resource-limited systems. They have compiled and presented an accuracy and runtime performance data for all stereo vision disparity map algorithms in the past decade with an emphasis on real time performance.

Figure 2.1: Previous review papers on stereo vision disparity map algorithms.

increasing interest due to its computational efficiency, and a series of publications aimed at optimization. For example, Sun et al. (2004) and Luo (2005) found that the overall performance is optimal cascade if the slope of the logarithmic scale of the ROC curve is the same for all nodes, since individual nodes in cascade are statistically independent. That hypothesis by Huo and Chen (2004), which recently proposed a ”border surveillance” heuristic ROC to establish thresholds in a classifier thought cascade.The to analyze the optimal points against the Republic of China was first used by Provost and Fawcett (2001) in the context of comparison classifier. They showed that any misclassification cost, optimal classifiers are convex hull in the ROC.

Chapter 3

Disparity Map

In the first stage, our system performs stereo camera calibration to extract the intrinsic matrix, distortion parameter, and length of the baseline. For each input image sequence, image undistortion is applied. Then the images are to be rectified to find the corresponding points in the stereo pair images. In order to rectify the images we need to know the pose of cameras, which can be obtained from the fundamental matrix. Fundamental matrix is determined by matching of feature points of both the images.

3.1 Image Feature Points Extraction and Matching for Rectification

Image feature extraction is very important in image reconstruction step. The characteristic extraction and matching algorithms are continuously updated in recent years. There are a lot of feature extraction methods, which are used in various fields of real life scenarios, such as the acquisition of images of space, weather and pedestrian detection. The SIFT (Scale Invariant Transform functions) [18] proposed by David G.Lowe made significant improvements in feature extraction algorithm. SIFT has the rotational and scale invariance and great anti-noise capability. SIFT is better than different strategies and has been broadly utilized; Now it has turned into the standard of the algorithm. In 2006, Bay, et al., Proposes an enlistment algorithm in view of SIFT algorithm after hearty element brisk SURF (speeded up strong components) [19] algorithm. This technique has the power of the scale and pivot as SIFT algorithm however has significantly enhanced the rate of computation. Edward Rosten initially presented the FAST (Feature Accelerated Segment Test) algorithm in 2006, which significantly enhances the speed, yet exactness is not high and has scale invariance. Keeping in mind, the end goal to streamline the estimation, there is a two fold technique algorithm for removing highlight points, utilized as a part of late years. The ORB is a run of the mill double algorithm proposed by Ethan Rublee, et al., In ICCV (IEEE International Conference on Computer Vision) 2001. ORB algorithm is advanced in view of FAST and BRIEF algorithm, not just the pace has enhanced a great deal , additionally has the revolution invariance when contrasted with single Brief algorithm. Firstly this proposal utilizes middle channel to dispose of the clamor in the picture to lessen the interference, then uses the ORB algorithm to concentrate descriptors.

3.1.1 The Principle of ORB Algorithm

ORB algorithm [4] takes the advantage of the fastest detector and descriptor techniques. ORB first finds the position of the key points by the FAST feature point detector and then uses Harris method to select top N points among those feature points. since FAST is not scale invariant and rotation invariant, ORB uses Scale-pyramid transform to enforce scale invariance. In order to enforce rotation invariance, it adds the direction of the points along the intensity Centroid. In the later stage, it extracts the binary descriptor of those key points using BRIEF algorithm. Yet again, BRIEF is not rotation invariance, so, ORB implements rBRIEF (rotated-BRIEF). It receives a 256-bit descriptor for each feature point, which can be easily matched with other potential corresponding points.

FAST Feature Points Detection

Feature Accelerated Test Segment (FAST algorithm) [20] has become one of the finest mechanisms for the feature detection. The concept behind the FAST algorithm is that if the surrounding pixels across a pixel ”A” contains adequate points in distinct gray value with the “point A”, then the pixel “point A” is considered as a feature curve. Applying it on the gray scale images, the neighborhood FAST curve has enough number of pixels, and the gray values of these pixels is larger or smaller than pixel of the point gray values around ’A’. In general, the choice of a random pixel as the center to form a circular area, and taking into account of this area as the neighborhood of pixel point. Figure (3.1) shows a unique circle of the radius 3, the central pixel is ’P’ and peripheral pixels clockwise given numbers from 1 to 16. If in those 16 pixels, there are n pixels and they satisfy the equation (3.1), Considering P as a point of candidate feature.

$$\mid I_x - I_p \mid > t \tag{3.1}$$

where t is the given threshold, I_x is the intensity value of each of the n pixels, I_p is the intensity value of pixel P. To obtain fast, we can choose n as 12 or 9, etc., It depends on the particular necessities, and the value of n is taken as 9 in ORB algorithm.

We can calculate the four points 1,9,5,13 first. The point may be considered as a candidate corner point if we can find at least three points, among those four points, to satisfy this point and we continue to compare the other points. But both FAST detector and BRIEF descriptor algorithms are scale variant, hence we use scale pyramid to solve this problem. For candidate characteristic points, we must carry out rapid tests in each layer.

ORB algorithm uses Harris mechanism [21] to identify the best suited N interesting points in the feature points that have been detected.

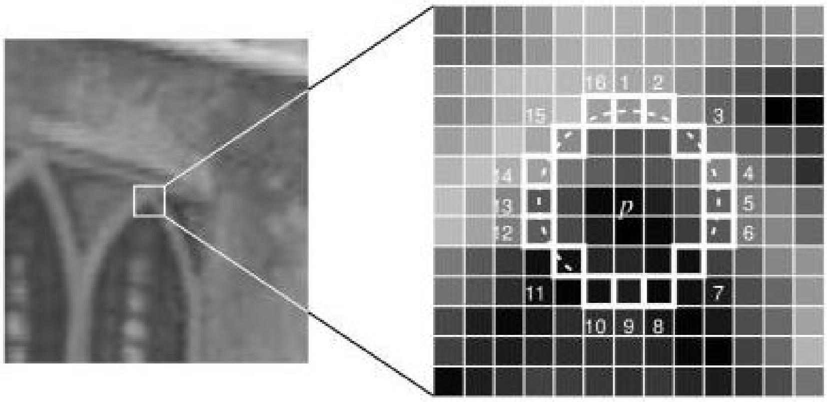


Figure 3.1: FAST feature point selection

FAST has the characteristics of translation and motility invariance and insensitive to noise, dependableness of feature points, conveniently computing, but feature points have no directional orientation. ORB utilizes the O-FAST method to find feature points, which is the FAST operator that has directional orientation. To obtain the direction of the corner, we achieve the feature point that has a inclination using intensity centroid technique.

BRIEF Feature Points Descriptor

DBRIEF(Binary Robust Independent Elementary Features) descriptors [22] are built after obtaining the FAST feature points with orientation. Descriptors around feature points are extracted using BRIEF by binary coding method. The BRIEF descriptors are simpler and space required to store these descriptors is smaller than SIFT and SURF.

The image patch P is SxS over the feature pixel, selecting nd pairs of pixels and defining it as τ

$$f(x) = \begin{cases} 1, & \text{if p(x) < p(y).} \\ 0, & \text{otherwise.} \end{cases} \tag{3.2}$$

p(x) is the intensity of the pixel at x=(u ,v)^T which is in the image patch P after filtering process. A set of points can uniquely identify using one binary detection τ. It defines nd-dimensional binary string just as equation (2.3) as the BRIEF descriptor.

$$f_{n_d} = \sum_{1<i<n} 2^{i-1}(p, x_i, y_i) \tag{3.3}$$

nd can be 128, 256 and so on.

Brief algorithm does not have scale invariance. Besides, there are drawbacks, like noise sensitivity and orientation invariance. Primarily, solve the issue of noise sensitivity. BRIEF selects random nd points for comparison around the feature point in the pixel dots 31x31 while the ORB makes the equivalent point compared to the sum of pixels. The window of 5x5 pixels takes the location of a single original point, the ORB has smoothen the image blocks and improves stability and repetition of the descriptor. Accelerate the overall image, so this step will not be a great loss in performance.

The ORB algorithm can rectify the issue of rotational invariance which the BRIEF algorithm couldn't, the technique is to add a orientation for BRIEF descriptors. First defining the moments of neighborhood by

$$m_{pq} = \sum_{x,y} x^p y^q I(x,y) \tag{3.4}$$

x, y is in the location of the FAST feature point with circular neighborhood radius r, x, y $\in [-r, r]$.

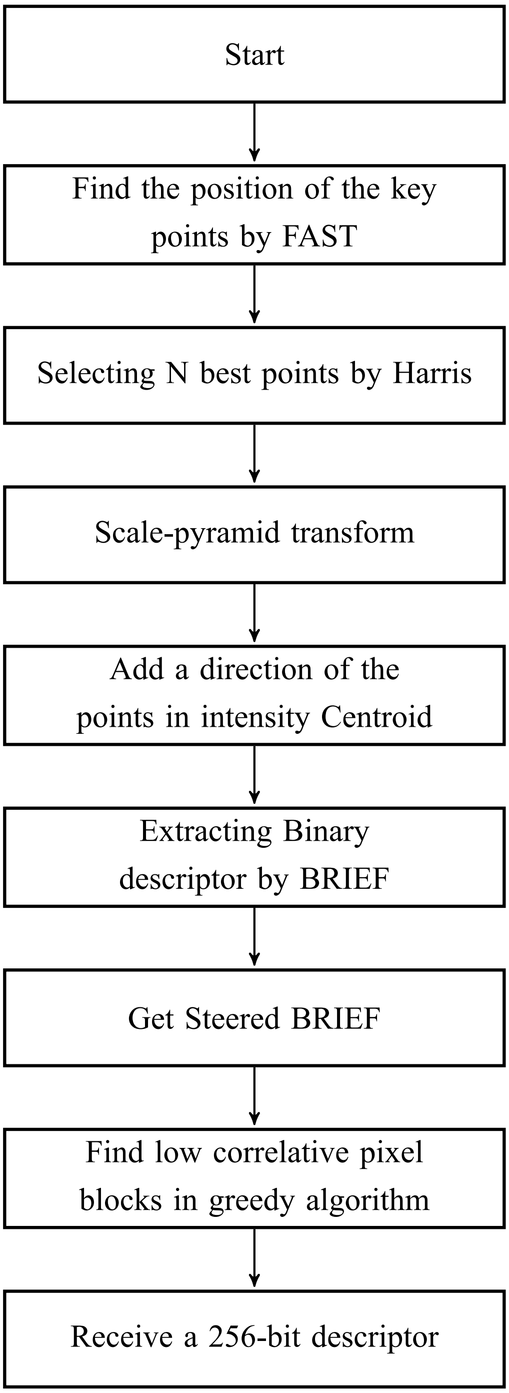
Then calculate the intensity centroid, as shown in equation (3.5).

$$C = \Big(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}\Big) \tag{3.5}$$

The direction which is obtained by feature point through intensity centroid has been defined as, for the FAST feature point direction:

$$\theta = arctan\Big(\frac{m_{01}}{m_{10}}\Big) = arctan\left(\frac{\sum_{x,y} yI(x,y)}{\sum_{x,y} xI(x,y)}\right) \tag{3.6}$$

The ORB algorithm extracts the descriptors according to the orientation given by equation (2.6) Due to environmental factors and noise, the orientation of the feature points will change, the correlation of the blocks of random pixel will be comparatively huge, thereby decreasing discrimination descriptors. ORB randomized algorithms take an avid to find the block random pixel with low correlation algorithm, generally select block 256 pairs of pixels with the smallest correlation function shape descriptor 256 bits, called rBRIEF.


Start
Find the position of the key points by FAST
Selecting N best points by Harris
Scale-pyramid transform
Add a direction of the points in intensity Centroid
Extracting Binary descriptor by BRIEF
Get Steered BRIEF
Find low correlative pixel blocks in greedy algorithm
Receive a 256-bit descriptor

Flow chart of ORB Algorithm.

3.1.2 Hamming Distance Matching

The description of the features obtained using ORB is a binary string. We can easily match the corresponding feature points in both the images using Hamming distance. Hamming distance between two strings of same length can be obtained by calculating the number of similar bits at the corresponding index of the strings. More the number of similar bits, more similar the descriptors are. We set a threshold value, in order to match the points. This threshold is based on the ratio of the distance T and the second neighbour distance, where T<0.8. This matching technique gives the simpler and faster way of matching two descriptors.

3.1.3 Eliminating the False Matching Points

Random sample consensus (RANSAC) [23] and least squares adjustment methods are most widely utilized to identify false corresponding points. But the least squares adjustment technique is easily influenced by false correspondence with large errors. RANSAC is an reiterative mechanism to calculate the parameters of a mathematical framework from a set of determined data. RANSAC algorithm is supported on a set containing aberrant data, the calculation of the mathematical model parameter data, obtain values from the sample. RANSAC algorithm uses fewer points to determine the model, then check the model of the leftover points. Default is the general method of parameter approximation, then reducing the influence of anomalous data.

The method has been implemented as described in the previous section provides better and faster feature points and thus helps to accelerate the generation of the depth map.

3.2 Dense Disparity Generation

The major task to compute the disparity map is stereo correspondence problem. Several stereo matching techniques have been implemented to solve the correspondence problem in the stereo pair images. These techniques are mainly categorized into two categories: global and local. Global techniques depend on repetitious schemes that produce disparity assignments depending on the diminution of a global cost function. Global techniques produce more accurate and dense disparity measurements but consume a lot of computational cost that makes them unsuitable for the real time applications. Local techniques (area-based techniques) calculate disparity value at each pixel on the basis of intensity gradients of the neighbouring pixels. Local techniques are significantly less accurate than global techniques but they can run fast enough to be applied in real time applications.

We first rectify the images using the transformation matrix which can be obtained by using the feature points we extracted using ORB in last step in such a way that the epipolar lines of the corresponding points in both the images intersect each other. This makes the left image point has a related pixel in the right on the same level line. Finding the same points in two images such that the matched points are the same projections of a point in the scene is called stereo correspondence matching and is the basic computational task underlying disparity map determination. To determine the correspondence of a pixel in the left image we use Semi block global matching(SGBM).

3.2.1 Correspondence Matching

Correlation based stereo matching has long been the preferred method for dense reconstruction, especially if runtime is an issue, due to real-time constraints or huge amounts of data. Correlation methods compute the similarity between pixels by comparing windows around pixels of interest. This is done for all possible correspondences for finding the pixel with the highest similarity. Correlation methods implicitly assume that all pixels within the window have the same distance from the camera, i.e. lie on a fronto-parallel surface. Slanted surfaces and abrupt changes, as caused by depth discontinuities, result in wrongly including non-corresponding image parts in the calculation. Slanted surfaces can be handled by permitting affine transformations of corresponding windows, i.e. Least Squares However, this is computationally expensive and does not work at fine structures or depth discontinuities. Therefore, the computer vision community often uses the sum of absolute or squared differences and optionally limits differences (i.e. truncated costs as discussed by Scharstein and Szeliski (2002)) for reducing errors due to matching non corresponding pixels. parametric costs like Rank and Census further reduce the problem (Zabih and Woodfill, 1994) since outliers have a much lower weight. The problem can also be reduced by adapting the size and shape of windows (Kanade and Okutomi, 1994; Fusiello et al., 1997; Hirschmüller et al., 2002). However, a real solution of this problem is only possible by matching pixels individually instead of matching windows. Of course, individual pixels do not contain enough information for unique matching. Therefore, global methods additionally use a smoothness constraint that penalizes discontinuities. This is typically formulated in a cost function,

$$E(D) = \sum_{p}(C(p, D_p) + \sum_{q\ N_p} PT[|D_p - D_q| \geq 1]) \tag{3.7}$$

The first term of the function sums all pixel-wise matching costs over the whole image, while the second term adds a penalty for all pixels with neighbors that have a different disparity. In this way, discontinuities are permitted if pixel-wise matching is stronger than the penalty, i.e. if the texture indicates a discontinuity. It is noteworthy that the second term indirectly connects all pixels with each other in the image and makes the function global. In this formulation, the disparity image D is sought that minimizes equation (1).

Unfortunately, this is an NP complete problem (Boykov et al., 2001). Famous approximate solutions to this problem are known as Graph Cuts (Kolmogorov and Zabih, 2001) and Belief Propagation (Felzenszwalb and Huttenlocher, 2004). The drawback of these and many other global methods are the speed and memory consumption, which often does not scale well with image size.

3.2.2 Disparity Generation with Semi block global matching(SGBM)

Semi-Global Matching (Hirschmüller, 2005 and 2008) combined with successful concepts of global and local stereo methods establishes the exact match, pixel-wise at low computational time. The central algorithm considers pairs of images with the known intrinsic and extrinsic orientation. The method has been applied to the undistorted rectified images. In the latter case, the epipolar lines are calculated and followed explicitly while efficient matching (Hirschmüller et al., 2005). SGM uses a slightly different global cost function as shown in equation (2.7) for penalizing disparity steps, that are frequent components of inclined surfaces, less than real discontinuities:

$$E(D) = \sum_{p}(C(p, D_p) + \sum_{q\ N_p} P_1 T[|D_p - D_q| = 1] + \sum_{q\ N_p} P_2 T[|D_p - D_q| > 1]) \tag{3.8}$$

The novel idea of SGM is calculated along various paths, symmetrically from all directions through the image. Each path carries information about the cost to reach a pixel with a certain amount disparity. For each pixel and each gap, the costs add up during the eight paths. Thereafter, at each pixel, the disparity with the lowest cost is chosen. This formulation ignores occlusions. Therefore, arbitrary results in occluded areas occur. However, occlusions can be identified by calculating disparities separately for the left and right image and comparison of the results of a check of left-right consistency. Other subsequent processing steps are possible for cleaning image disparity.s

Chapter 4

ROI Generation Using Dense Depth

A depth map is a map in which the value of each pixel gives the depth of that point in the real world. These depth values can be used to estimate the human size at each depth. In our proposed method, first depth map is segregated into different layers and the size of the rectangular windows is defined that vary along the each layer. Therefore, comparing the number of pixels that fall in the window with a predefined threshold, deciding whether the region contains a pedestrian or not. Our method of generating ROI can be divided into three main stages: the layering of depth map, extracting skeleton for each layer of depth and defining a window for each adaptation layer and sliding on the skeleton of objects in each layer.

4.1 Layering Depth Map

After generating disparity map, our system secretes the image based on the disparity values. In order to detect pedestrians, the disparity range for projecting human image has to be estimated. In this thesis we have considered this range to be 50. The depth image obtained is divided into several sub-images in which each sub-image contains objects with similar depth values (range 50). For this purpose, depth histogram is obtained and pixel values belonging to each histogram bin represents a sub-image.

4.2 Extracting image skeleton

In methods based on exhaustive search, all pixels are considered for search but most of these pixels are not essential for calculating the disparity. To reduce the computational cost, consider the pixels of the image of the skeleton. For each image layer, medial axis skeleton is removed and the search window slides only on these pixels. Image skeleton, S(X), can be extracted using morphological operations given in Eqns. (4.1) and (4.2):

$$S(X) = \bigcup S_n(X) \tag{4.1}$$

$$S_n(X) = (X \otimes nB) - (X \otimes nB) \circ B \tag{4.2}$$

where $\otimes$ and $\circ$ are morphological erosion and opening, respectively, n = 0,1,2,... N, X is the input image, B is the structuring element and

$$nB = B \oplus ..B(ntimes) \tag{4.3}$$

4.3 Adaptive windowing

Since the size of an object is perceived more strongly influenced by the object distance from the camera, we determine the detection window size for each ROI search image based on the depth sub value. Specifically, we first proposed size detection window of the first layer depth that is closest to the camera as the height/2 x width/6 given the approximate shape of a human. Then, adjust the height and width of the detection window in the layer depth nth using Eqs. (4.4) and (4.5).

$$h_n = \frac{D - d_n}{D - d_1} * h_1 \tag{4.4}$$

$$w_n = \frac{D - d_n}{D - d_1} * w_1 \tag{4.5}$$

where d_1, h_1 and w_1 are the disparity value of the first layer and height and width at that layer, D is the maximum depth range (255 in our experiments), and d_n is the depth value in the n-th depth layer.

we slipped the window along the medial axis of objects and calculate the magnitude of convergence between the detection window and the candidate region. If the magnitude of convergence is greater 1/3, the region is determined as a candidate region of interest.

4.4 Experimental Results

To test the performance of our method, we used a data set based DAS stereoscopic vision provided by [24]. classification was used and system packages this data set that provides both the formation and test sets and multiple video streams. Its test equipment contains 250 frames with 640 * 480 pixels and range from 0 to 50 meters captured by stereo vision camera from a moving vehicle in various lighting conditions and with different size and number of pedestrians. The training set contains 7649 negative and 1015 positive samples. Since depth maps in the data set have poor quality and limited range, we applied linear image inpainting and contrast extends to improving the quality depth maps.

Chapter 5

Pedestrian Classification

The candidate region of interest areas are mapped over the original color image and is exposed to feature extractions. The ROI contains both pedestrian and non-pedestrian regions, an pedestrian classification step is necessary to classify these ROI.

5.1 Feature Extraction

Feature extraction process may either use explicit features or use implicit features. Features such as edges, symmetry shadow etc., are referred as explicit features. Implicit features are extracted using techniques such as Histogram of gradients (HOG), Haar wavelet (henceforth referred as Haar features), Scale Invariant Feature Transform (SIFT) features and Gabor filters. Both HOG and Haar features are extremely-well utilized in various pedestrian detection researches.

5.1.1 Haar Based Extraction

Haar features are simplified form of Haar waelets. Three categories of Haar features are edge features, linear features and center features. Haar features can be computed fast because integral image is being used. Haar features are computed similar to the co-efficients of Haar wavelet transform. A template connecting black and white rectangles, and their relative coordinates to the origin of search window is what form Haar like features. Like HOG features, Haar features are independent to illumination variation. HOG features are much heavier for computation compared to Haar features. For the detection of horizontal, vertical, and symmetric structures, Haar-like features are found to be well suitable.

5.1.2 HoG Based Extraction

Among several shape and appearance based methods, we use Histogram of Oriented Gradients (HOG) for classifying ROI since it gives the best results in human detection (knowledge obtained from literature survey). HOG is one of the popular ways of deriving descriptor for a bounding box of an object. HOG is a feature descriptor inspired by SIFT

descriptor and is based on calculating orientation gradients in sub-blocks of a region. It is based on identifying intensity gradients of object shape within an image.
HOG is calculated with the following steps.

1. Compute image gradients of each pixel.
2. Accumulate weighted votes into orientation bins.
3. Contrast normalization for each block.
4. Collect HOGs for all blocks.

To calculate the HOG descriptors of an image, the image is first divided into small regions and each region of the gradient histogram edge directions or guidance be computed. The combination of all these histograms forms the descriptor at last. In descriptor HOG, the region is divided into smaller cells and orientation histograms are calculated for each cell. histograms They can be created by calculating the orientation of gradients to form the intervals of the histogram and its magnitude as a vote for bin using equations (5.1) and (5.2)

$$m(x, y) = \sqrt{{L_x}^2 + {L_y}^2} \tag{5.1}$$

$$\theta(x, y) = arctan\frac{L_y}{L_x} \tag{5.2}$$

where m is the magnitude and θ is the orientation at each pixel and L_x and L_y are gradients in x and y direction. These cells are re-grouped into blocks and histograms of these blocks are normalized to improve lighting invariance. The standard blocks are concatenated to obtain vector features.

HOG feature is not sensitive to variance of light, direction, and size whereas it will not work well in scale invariant and rotation invariant scenarios. But since we already have the sub images which don't need scale invariant feature extractor, the HoG feature extraction provides very good results, with the system.

For the HOG descriptor, the size of blocks and cells are set to 3x3 and 9x9, respectively, and 9 bins are used for the histograms.

5.2 Classification

Classification or Hypothesis verification can be either based template based or based on appearance based. Predefined patterns from vehicles are used in template based classifications. But appearance based classification learns characteristics of the vehicle from set of training images. Training is a supervised learning and a large set of pedestrian and non-pedestrian images are used. Features selection is very important to achieve good classification results. Classification for appearance based vehicle detection uses artificial neural network concepts. Support vector machines (SVM) based classifier training is a popular classification method for vision based pedestrian detection. Adaboost classification is also widely used for vehicle detection due to its adoptability for cascade classification. General trend in vehicle detection approaches are HOG+SVM and Haar+Adaboost. HOG-SVM formulation is effective to calculate vehicle orientations whereas Haar-Adaboost is useful for detecting rear faces of vehicles. Differences between SVM and Adaboost classifier is given in Table below.

5.2.1 Discussion

Two stage process including HG & HV is a widely accepted process by various authors to derive a robust pedestrian detection process. Background subtraction may be useful for static background whereas, it may not be suitable where background keeps changing. Tail light detection is a simple method but works well for night time and when the leading vehicle travels in the same plane. Optical flow has been used for monocular vehicle detection. This method finds applications in overtaking vehicle detection in blind spot. This method may detect persisted interest points over long periods of time. Also, this method can’t work for slow moving objects. Radar sensor used for minimizing ROI suffers with time synchronization issues. Lidar based vehicle detection yields robust results under various scenarios but they are expensive. Stereo vision techniques are highly resource consuming and may need high speed hardware. This two stage process will bring down the computing time and computing resource needs to major extent because this avoids exhaustive searches over all possible variations. This method achieves relatively better accuracy compared to other methods which use less exhaustive random sampling . There are various state-of-the art methods that are matured, robust and efficient. They are partially available in market and they can detect the presence of other vehicles with high levels of accuracy. One of the main challenges to extend the application of pedestrian detection is the computing needs. Though these methods provide better detection results, the computation is much heavier, and it is hard to migrate to embedded architectures currently used in automobile industry. Therefore it becomes necessary to reduce the computation time with minimum resources without compromising the accuracy. Driver assistance systems seem promising to play important role in the near future automotive systems with the evolution of vision based detection schemes. Considerable effort has been spent by researchers to derive techniques for robust and reliable vehicle detection techniques. However many of these methods have been verified mostly with nonurban environments. Very few papers attempted to tackle the challenges posed by urban scenarios.

S.No	SVM	Adaboost
1	Used as classifier for HOG based detectors	Used as classifier for channel based detectors
2	Majority detectors are HOG based	Recent detectors employ channel features
3	Part-based models or occlusion handling strategy have been explored for HOG based detectors	Efficient to detect most discriminative channel features

5.2.2 Linear Support Vector Machine(L-SVM)

We use a linear SVM to classify the feature vectors. To do this, the first to use SVM trained with 1,000 pedestrians and 1,000 non-pedestrian frames. Based on this table support, regions of test frames extracted interest are classified into classes and no crosswalk. The regions of interest extracted from the 250 frames of the test equipment using our detection algorithm ROI below has been classified by the classifier. In the application of detector ROI, the initial selection of frame sizes is set to 240x107 and the threshold for the filling area in the bounding box is set to 1/3.

Chapter 6

Results

The pedestrian detection system has been improved in the virtue of speed and detection performance. The advantage of ORB algorithm in the ROI selection and of adaptive window technique for feature extraction decreased the computational time by 4 times over exhaustive HOG-SVM technique. Comparing the results of our proposal with the HOG pedestrian classification technique, you can see that our technique outperforms the method of prior state of the art, except in a small range. The main reason for this improvement is due to the use of depth maps for generating ROI that are invariant to lighting conditions. Our system is also scale invariant due to use of adaptive window for ROI extraction. We have also performed another experiment by changing the initial window size adjustment and the results show that by reducing the size of the window, the detection rate is reduced.

Image size(pixels)	**Time taken by SURF (s)**	**Time taken by ORB (s)**
480 × 640	0.301	0.032
370 × 1226	0.417	0.030
376 × 1241	0.401	0.046

Table 6.1: Comparison of computational speed of SURF and ORB

System configuration

The above results are observed with OpenCV-python installed on windows 10 with i5 processor.

Chapter 7

Conclusion

Pedestrian detection is a trending field because it has many applications in the real world scenarios, which can be used in everyday life. With the advancement of technology, the need for faster and accurate pedestrian detection methods are necessary. The Stereo ROI generation methods are widely used in various other image processing, computer vision and machine learning applications.

The sliding window paradigm consumes a lot of time. Occlusions can only be lined out using stereo vision. In Stereo ROI generation method, since the features required to rectify the image are extracted using the ORB algorithm, it can be noticed that the pedestrians can be detected with less computational cost. This ROI selection method can also be used in many other application which can help in fasten the classification or detection.

Scope for Further Research

The ORB algorithm mentioned in this thesis is not by self, a scale variant. ORB also leads to many wrong matching points. In order to solve that RANSAC algorithm and homography matrix can be used. This could give better feature point extraction. No perceptive transformation matrix is used to correct the images. If had been, better image matching can be achieved.

References

[1] S. Foix, G. Alenya, and C. Torras, “Lock-in time-of-flight (tof) cameras: a survey,” *Sensors Journal, IEEE*, vol. 11, no. 9, pp. 1917–1926, 2011.

[2] M. Y. Kim, S. M. Ayaz, J. Park, and Y. Roh, “Adaptive 3d sensing system based on variable magnification using stereo vision and structured light,” *Optics and Lasers in Engineering*, vol. 55, pp. 113–127, 2014.

[3] S. Zhang, C. Wang, and S. Chan, “A new high resolution depth map estimation system using stereo vision and depth sensing device,” in *Signal Processing and its Applications (CSPA), 2013 IEEE 9th International Colloquium on*. IEEE, 2013, pp. 49–53.

[4] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: an efficient alternative to sift or surf,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2564–2571.

[5] A. Mohan, C. Papageorgiou, and T. Poggio, “Example-based object detection in images by components,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 4, pp. 349–361, 2001.

[6] C. Papageorgiou and T. Poggio, “A trainable system for object detection,” *International Journal of Computer Vision*, vol. 38, no. 1, pp. 15–33, 2000.

[7] L. Zhao and C. E. Thorpe, “Stereo-and neural network-based pedestrian detection,” *Intelligent Transportation Systems, IEEE Transactions on*, vol. 1, no. 3, pp. 148–154, 2000.

[8] P. Viola, M. J. Jones, and D. Snow, “Detecting pedestrians using patterns of motion and appearance,” *International Journal of Computer Vision*, vol. 63, no. 2, pp. 153–161, 2005.

[9] X. Xie, S. Wu, K.-M. Lam, and H. Yan, “Promoterexplorer: an effective promoter identification method based on the adaboost algorithm,” *Bioinformatics*, vol. 22, no. 22, pp. 2722–2728, 2006.

[10] H. Elzein, S. Lakshmanan, and P. Watta, “A motion and shape-based pedestrian detection algorithm,” in *Intelligent Vehicles Symposium, 2003. Proceedings. IEEE*. IEEE, 2003, pp. 500–504.

[11] R. Polana and R. Nelson, “Low level recognition of human motion (or how to get your man without finding his body parts),” in *Motion of Non-Rigid and Articulated Objects, 1994., Proceedings of the 1994 IEEE Workshop on*. IEEE, 1994, pp. 77–82.

[12] W. B. Thompson and T.-C. Pong, “Detecting moving objects,” *International journal of computer vision*, vol. 4, no. 1, pp. 39–57, 1990.

[13] M. Bertozzi, A. Broggi, P. Grisleri, T. Graf, and M. Meinecke, “Pedestrian detection in infrared images,” in *Intelligent Vehicles Symposium, 2003. Proceedings. IEEE*. IEEE, 2003, pp. 662–667.

[14] D. M. Gavrila and S. Munder, “Multi-cue pedestrian detection and tracking from a moving vehicle,” *International journal of computer vision*, vol. 73, no. 1, pp. 41–59, 2007.

[15] R. Labayrade, D. Aubert, and J.-P. Tarel, “Real time obstacle detection in stereovision on non flat road geometry through” v-disparity” representation,” in *Intelligent Vehicle Symposium, 2002. IEEE*, vol. 2. IEEE, 2002, pp. 646–651.

[16] A. Shashua, Y. Gdalyahu, and G. Hayun, “Pedestrian detection for driving assistance systems: Single-frame classification and system level performance,” in *Intelligent Vehicles Symposium, 2004 IEEE*. IEEE, 2004, pp. 1–6.

[17] H. Shimizu and T. Poggio, “Direction estimation of pedestrian from multiple still images,” in *Intelligent Vehicles Symposium, 2004 IEEE*. IEEE, 2004, pp. 596–600.

[18] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[19] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” in *Computer vision–ECCV 2006*. Springer, 2006, pp. 404–417.

[20] M. Trajković and M. Hedley, “Fast corner detection,” *Image and vision computing*, vol. 16, no. 2, pp. 75–87, 1998.

[21] C. Harris and M. Stephens, “A combined corner and edge detector.” in *Alvey vision conference*, vol. 15. Citeseer, 1988, p. 50.

[22] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, “Brief: Binary robust independent elementary features,” *Computer Vision–ECCV 2010*, pp. 778–792, 2010.

[23] F. Tombari, S. Mattoccia, L. D. Stefano, and E. Addimanda, “Classification and evaluation of cost aggregation methods for stereo correspondence,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[24] D. Gerónimo, A. D. Sappa, D. Ponsa, and A. M. López, “2d–3d-based on-board pedestrian detection system,” *Computer Vision and Image Understanding*, vol. 114, no. 5, pp. 583–595, 2010.