Transportation Research Part C 25 (2012) 226–237

Contents lists available at SciVerse ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Overview Paper

Stereo regions-of-interest selection for pedestrian protection: A survey

D.F. Llorca [a,*], M.A. Sotelo [a], A.M. Hellín [a], A. Orellana [b], M. Gavilán [a], I.G. Daza [a], A.G. Lorente [a]

[a] Computer Engineering Department, Polytechnic School, University of Alcalá, Madrid 28871, Spain
[b] Institute of Automatics, Faculty of Engineering, National University of San Juan, 1109 San Juan, Argentina

ARTICLE INFO

Article history:
Received 12 July 2011
Received in revised form 14 June 2012
Accepted 15 June 2012

Keywords:
Pedestrian detection
Regions of interest selection
Intelligent vehicles
Intelligent transportation systems
Stereo
Survey

ABSTRACT

Vision-based pedestrian detection for intelligent vehicles applications is a crucial and active research area due to the essential benefits in terms of reducing the number of accidents involving pedestrians and vehicles. During the last decade a considerable amount of research studies have been proposed, filling the gap between prototypes and commercial implementations. Pedestrian detection systems can be roughly divided into three main different sub-parts: *Region Of Interest – ROI – selection*, *classification* and *tracking*. Previous surveys have covered the literature in a holistic way. An example would be, analyzing all the solutions proposed for all the stages and including higher level analysis, but in most cases they give more emphasis to the classification stage. Due to the difficulty of this detection task, the variety of solutions, sensor configurations (monocular/stereo; visible/infrared) available in the literature, we propose to break down the variability of the problem by providing exhaustive review of one specific stage: *stereo-based ROI selection*. ROI selection is a key component that has to be designed to provide generic obstacles at lowest false negative rate and maintain a low number of false positives. The number of missed pedestrians has to be approximately equal to 0 since a pedestrian missed by the ROI selection stage would not be detected in further stages. In addition, the number of non-pedestrians obstacles should be as low as possible to reduce both the number of false alarms and the computational costs of further stages. In contrast to monocular approaches, stereo ROI selection determines the relative distance between the pedestrian and the vehicle, assuring that the reported candidates are related with real physical objects. The stereo-based ROI selection step is also divided into different components that are independently analyzed, increasing visibility for future proposals and developments. Discussion is finally presented highlighting the current problems for obtaining a global overview of the actual performance of the different approaches and analyzing future trends.

1. Introduction

Pedestrian detection is a fundamental task for a variety of important applications, especially in the context of intelligent vehicles (IVs) and intelligent transportation systems (ITSs), since it clearly enhances the pedestrian safety. Every year, according to the statistics estimated by the World Health Organization (Peden et al., 2004), 1.2 million people are known to die in road accidents worldwide. A majority of the deaths and injuries involve motorcyclists, cyclists and pedestrians. Only in the European Union about 8000 pedestrians and cyclists are killed and about 300,000 injured each year. In North America, approximately 5000 pedestrians are killed and 85,000 injured. In Japan approximately 3300 pedestrians and cyclists are killed and 27,000 injured (UNECE, 2005). Over the last decade, this topic has attracted an extensive amount of interest from

* Corresponding author. Tel.: +3491 885 6641; fax: +3491 885 6682.
E-mail address: llorca@aut.uah.es (D.F. Llorca).

0968-090X/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trc.2012.06.006

national and international authorities, the automotive industry and the scientific community, aiming at improving the safety of the most vulnerable road users.

Active sensors such as acoustic-based, radar-based and laser-based have been proposed for pedestrian detection. Refer to the survey presented by Gandhi and Trivedi (2007), to have a broad overview of these active sensors-based approaches. However, during the last years, passive sensors, and more specifically optical sensors, have attracted most of the attention of the research community as well as the industry, due to two main aspects: inexpensive costs and new potential applications such as *Lane Departure Warning*, *Traffic Signs Recognition*, and *Adaptive Cruise Control*.

Pedestrian detection is a difficult task from the computer vision perspective. Large variations in pedestrian appearance, e.g. pose, clothing, size, etc., and environmental conditions, e.g. lighting, moving background, etc., make this problem particularly challenging. Vision-based detection systems can be classified with respect to the number of cameras, monocular or stereo, as well as depending on the spectrum, visible or infrared. In addition, vision-based pedestrian detection systems can be roughly divided into three main stages (see Fig. 1). The first stage consists of identifying generic obstacles as regions of interest, ROI selection, using prior scene knowledge: camera calibration, stereo information, ground plane constraint, etc. Subsequently, a more expensive pattern recognition step is applied: classification or verification. The lack of explicit models leads to the use of machine learning techniques, where an implicit representation is learned from features obtained from thousands, or millions, of samples. Finally, temporal integration or tracking stage is applied to improve single-frame detection performance and smooth the relative vehicle-to-pedestrian trajectory.

As depicted in Fig. 1, the different sub-parts are sequentially linked, that is, ROI selection outputs are fed to the classifier and classifier outputs are used as inputs for the tracking step. Accordingly, the performance of each stage is related to some extent with the performance of previous stages. For example, if the classifier fails when recognizing pedestrians, tracking stage would not be able to follow them. However, if a pedestrian has been detected and tracked during a considerable number of frames, tracking can absorb spurious classification errors. Thus, tracking performance has to be evaluated in the context of a ROI selection and classification ensemble. Additionally, the classifier results are strongly correlated with the type of samples provided by the ROI selection module, not only in terms of computational costs, i.e. the higher (lower) the number of samples to classify, the greater (lesser) the time needed, but in terms of both detection rate and false positive rate (Alonso et al., 2007). Actually, it is recommended to train the classifier with samples generated by the specific ROI selection mechanism in order to optimize the detection performance (Alonso et al., 2007). If the ROI selection algorithm usually provides a specific set of false positives (e.g., poles, trees, etc.), classifier should be boosted using these samples as negative samples. Single-frame analysis is usually carried out by using specific training and test databases to obtain information about the classifier performance and to define the working point of the classifier. However, the actual performance of the classifier can only be measured in real applications working in parallel with the ROI selection algorithm.

The early stage of a pedestrian detection system (ROI selection), does not depend on previous stages (see Fig. 1) and it is probably the key component due to one of the most critical requirements: the number of false negatives has to be approximately equal to 0. If the ROI selection does not detect a pedestrian as a candidate, this one would be neither classified nor tracked by further stages. The number of false positives provided by this stage is not as critical as the number of false negatives since non-pedestrian samples can be rejected by the classifier. However the classification computational cost defines an upper bound for this number. In addition, one of the desirable features of this stage is to provide both pedestrian and non-pedestrian candidates that correspond to real physical objects, that is, to avoid ghost targets that mainly appear due to reflections and shadows.

Among the surveys in the context of pedestrian detection available in the literature we remark (Gandhi and Trivedi, 2007; Gavrila, 1999; Moeslund and Granum, 2006; Poppe, 2007; Enzweiler and Gavrila, 2009; Gerónimo et al., 2010a). Most of the work concerning human motion has been summarized in (Gavrila, 1999; Moeslund and Granum, 2006; Poppe, 2007). Focusing on the pedestrian protection application in the context of intelligent vehicles, we have found three main surveys in the literature (Gandhi and Trivedi, 2007; Enzweiler and Gavrila, 2009; Gerónimo et al., 2010a). Gandhi and Trivedi (2007),

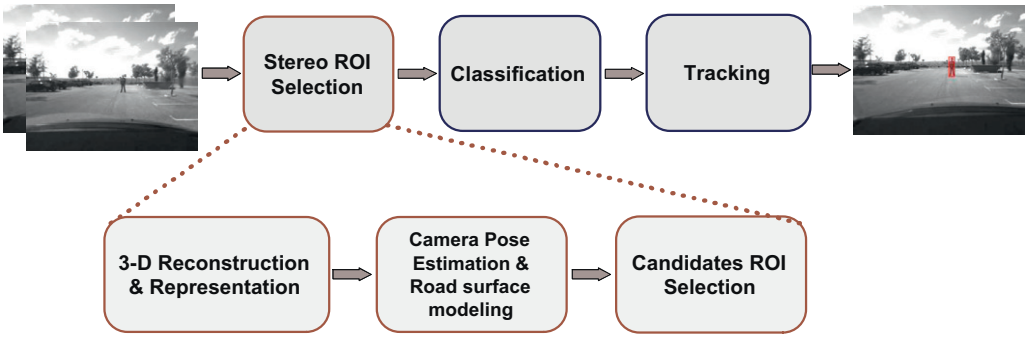

Fig. 1. Overview of the stages of a stereo vision-based pedestrian detection system. The presented survey covers the ROI selection module by means of stereo vision-based algorithms.

provide a broad review covering both active and passive safety technologies, discussing different approaches for collision risk assessment. Enzweiler and Gavrila (2009), provide a detailed overview of the current state of the art for monocular approaches from both methodological and experimental perspective. In addition, a large data set with tens of thousands of manually labeled pedestrians was made public for benchmarking. In a recent survey by Gerónimo et al. (2010a), a detailed discussion concerning benchmarking, public databases, and performance comparisons is also included.

The aforementioned stages have not been covered with the same level of detail by previous surveys. Most of the work is focused on the classification stage whereas ROI selection and tracking stages lack a comprehensive review. Obviously, this is mainly due to the fact that the different approaches proposed in the literature for each one of the sub-parts are unbalanced somewhat. In addition, the amount of available studies related with pedestrian detection makes unfeasible to provide exhaustive review of all the sub-parts in a unique review paper. On top of that, the amount of cameras used (monocular or stereo) as well as the spectrum (visible and/or infrared) introduce more specific variables, increasing the difficulty when providing in-depth analysis. Previous statements motivate to carry out the separate extended and comprehensive survey of each one of the pedestrian detection sub-parts, attempting to break-down the complexity of the problem and increasing visibility.

In this paper, we propose to enrich the insights of the early stage of a pedestrian detection system (ROI selection) in the context of intelligent vehicles, surveying the main approaches available in the literature. Instead of focusing in all ROI selection mechanisms, we are mainly concerned with stereo-based approaches. Monocular techniques that obtain initial object hypothesis are mainly based on the sliding window approach, where detector windows are shifted over the image at various scales and locations. This technique is usually combined with a classifier cascade of increasing complexity or by restricting the search area based on known camera geometry, prior information about the target object dimensions (pedestrian height or aspect ratio) and application-specific constraints such as the flat-world assumption. Other 2D-based attention focusing strategies are based on motion information and interest-point detectors. Most of the references are provided in (Enzweiler and Gavrila, 2009; Gerónimo et al., 2010a).

Although there are some monocular pedestrian detection systems that report very good performance (Shashua et al., 2004), monocular attention mechanisms cannot ensure that the selected candidates correspond to real physical objects (ghost targets). In addition, monocular systems provide ranges estimates after applying flat-world assumption, but they are not range sensors. These features compromise the use of monocular approaches for collision avoidance applications. A second camera can be used to solve these problems since depths of scene points based on disparity between images can be obtained. The depth information offers valuable cues for the ROI selection stage. Flat-world assumption can be overruled, and more complex models can be used to improve the scene understanding. Note that monocular approaches can always be used and combined with stereo-based approaches using images from both cameras. In addition, stereo cues can be useful in further stages such as classification (Keller et al., 2011b) and tracking (Alonso et al., 2007).

Left and right cameras are mounted on a single rigid frame which involves that calibration remains fixed. Real-time hardware computation of disparity and 3D points is commercially available. Depth estimation accuracy has its limitations due to the discrete nature of the stereo vision. However, by choosing a convenient set of system parameters (focal length, images size and baseline), a sufficient range estimation accuracy can be obtained, even for collision avoidance applications (Llorca et al., 2010).

As seen in Fig. 1, stereo-based ROI selection involves three main blocks: (1) 3D-reconstruction and representation; (2) Camera pose estimation and road surface modeling; and (3) Candidates ROI selection. In the following, we review the different approaches according to these three main sub-systems. However, note that we are not concerned with pedestrian detection approaches that use stereo as a verification step, providing depth measurements and adding robustness to the detection process (Bertozzi et al., 2005; Stein et al., 2010). In addition from the perspective of this survey, the use of infrared or visible spectrum is not relevant; to our knowledge there are not significant differences reported in the literature between infrared stereo and visible stereo.

The remainder of this paper is organized as follows: Section 2 provides and overview of the different approaches for representing the 3-D measurements provided by stereo vision systems. Section 3 specifically focuses on how to model the extrinsic relationship between the camera and the road, as well as the methods concerning road modeling. Section 4 surveys the object detection stage and, finally, in Section 6, we conclude this paper, discussing future challenges and needs.

2. 3D map representation

Stereo vision refers to the ability to obtain three-dimensional (3-D) measurements from two images taken from slightly different viewpoints (left and right cameras) (Forsyth and Ponce, 2003). From a computational standpoint, a stereo system must solve two problems. The first consists in determining which item in the left camera corresponds to which item in the right one (correspondence problem). The second problem is stereo reconstruction; if we solve the first problem, stereo reconstruction is straightforward by using a pure algebraic approach (Forsyth and Ponce, 2003). Accordingly, most of the efforts have to be focused on finding the correct correspondences between image points from the left and right cameras. If the geometry of the stereo pair is known for each point the searching area in the other image is constrained to a single epipolar line. Most of the approaches to stereo assume that the epipolar lines run parallel to the image lines. This situation can be forced by means of stereo pair rectification (Fusiello et al., 1998).

If only a subset of image pixels (feature-based stereo) is used for solving both the correspondence and the reconstruction problems, the stereo system will provide a sparse representation of the scene (Franke and Joos, 2000; Broggi et al., 2003; Gavrila and Munder, 2007; Alonso et al., 2007). The subset of image pixels is usually defined by edge features. This approach can be implemented very efficiently, meeting real time processing requirements, and with remarkable accuracy, since the matching process has to find correspondences between textured regions. However, with recent hardware advances, real-time dense stereo is becoming totally feasible. Dense stereo provides a more detailed description of the scene since it estimates 3-D measurements at all pixels, including low-textured regions by interpolation (Zhao and Thorpe, 2000; Grubb et al., 2004; van der Mark and Gavrila, 2006; Krotosky and Trivedi, 2007; Nedevschi et al., 2009; Keller et al., 2009, 2011b).

The output provided by a stereo vision system has to be arranged and prepared for further processing stages. Note that in this section we are just concerned with describing the different 3-D map representations used for pedestrian detection. The methods applied to select generic obstacles (ROIs) using a specific 3-D map will be detailed in Section 4.

The simplest representation of the 3-D measurements is the so-called *disparity map* which was first proposed for pedestrian detection by Zhao and Thorpe (2000). A *disparity map* I_Δ is a 2D image whose pixels (x,y) describe the displacement of every pixel in the right image with respect to the corresponding pixel in the left image. In other words, the grey level of each pixel of I_Δ represents the disparity value. The higher the grey level the higher (lower) the disparity (the depth). Depth Z (in meters) and disparity d (in pixels) are inversely proportional:

$$Z(x,y) = \frac{fB}{d(x,y)} \text{ at pixel } (x,y) \tag{1}$$

where f is the focal length and B is the distance between the cameras (baseline). See Fig. 2b and c for an example of both sparse and dense disparity images.

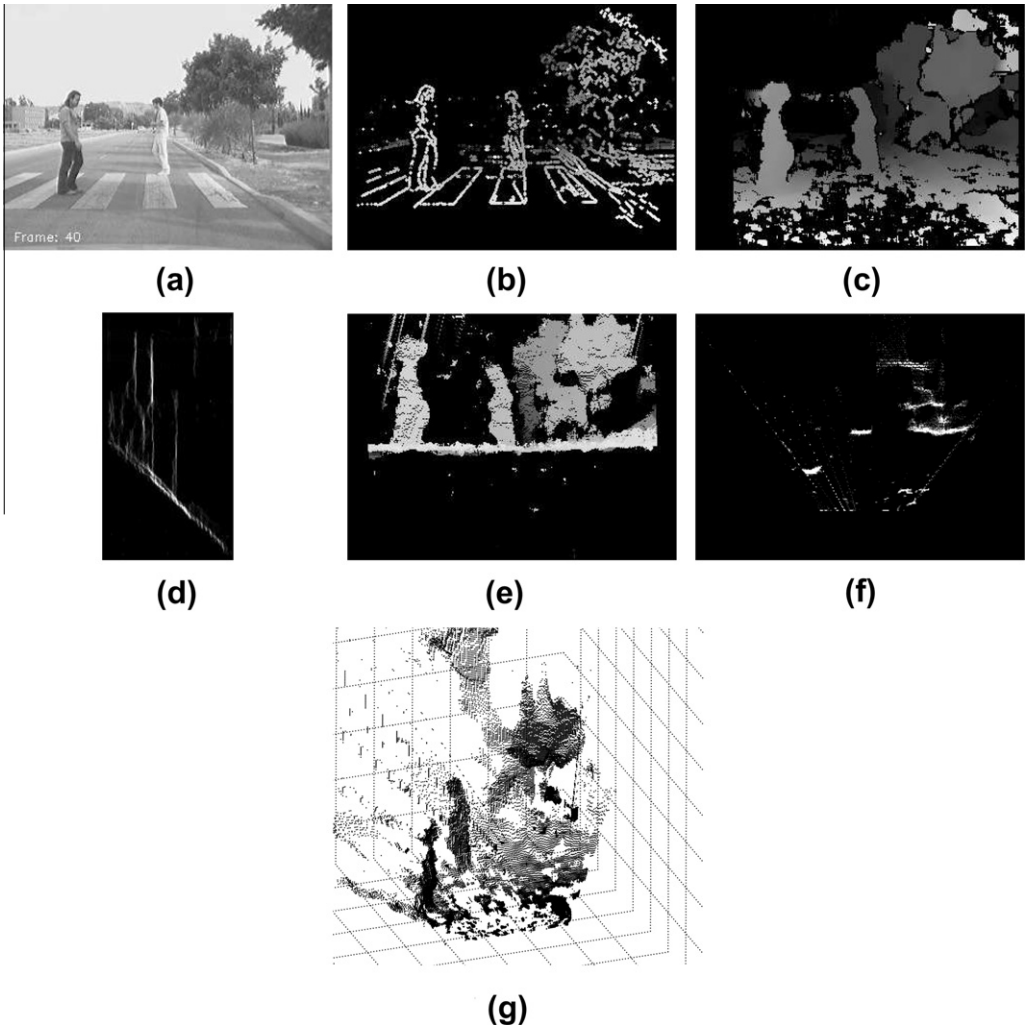

Fig. 2. (a) Left original gray level image; (b) sparse *disparity map*; (c) dense *disparity map*; (d) *v-disparity map*; (e) *virtual disparity map*; (f) *xOz* or *bird's eye* map; and (g) 3-D map.

One of the 3-D transformations most used in the context of pedestrian protection is the so-called *v-disparity map*. It was first proposed by Labayrade et al. (2002) for obstacle detection. Let F be the function of the disparity image I_Δ such that $F(I_\Delta) = I_{v\Delta}$ ($I_{v\Delta}$ is called the *v-disparity image*). F accumulates the points with the same disparity that take place on a given image line i. For the same line i, the abscissa u_M of a point $M(u_M, i)$ in $I_{v\Delta}$ corresponds to the disparity Δ_M and its grey level to the number of points with the same disparity Δ_M on the line i. This method provides scene understanding since ground surface and 3-D obstacles are clearly recognized (see Fig. 2d) and it has been widely used for road modeling and object detection in the context of intelligent vehicles (Labayrade et al., 2002; Broggi et al., 2003; Grubb et al., 2004; Zhencheng et al., 2005). A similar approach can be obtained by accumulating the points with the same disparity that occur on a given image row j. This approach leads to the so-called *u-disparity map* (Zhencheng et al., 2005; Krotosky and Trivedi, 2007).

According to Suganuma and Fujiwara (2007), the so-called *virtual disparity image* is proposed for obstacles detection for intelligent vehicle applications. This image is computed by virtually locating the camera on the ground plane. A rigid transformation involving a rotation to correct pitch and roll with respect to the ground plane, and a translation directly related with the camera height is applied to the 3-D points. The corrected points are then "virtually" backprojected on the image plane (using camera intrinsic parameters). In the *virtual disparity image* the ground plane is clearly visible (see Fig. 2e) making it possible to model vertical and even lateral surface changes (Suganuma and Fujiwara, 2007; Llorca et al., 2009).

Previous representations are based on disparity values, (inversely proportional to the depth, Eq. (1)), in image space where the resolution drastically decreases with increasing distance. To overcome this effect, other approaches propose to use measurements in world coordinates, instead of image space, to get a well-defined depth resolution (with the limitations of the stereo accuracy (Llorca et al., 2010)). 3-D Euclidean sparse space has been directly used by Alonso et al. (2007) (see Fig. 2g). 2D-based representations containing depth information can be also obtained by projecting the 3-D points on both the *yOz* plane (Llorca et al., 2009) and the *xOz* plane (Nedevschi et al., 2009) (see Fig. 2f). In addition, a two-dimensional polar-perspective grid map has been also proposed by Bajracharya et al. (2009). Unlike traditional *xOz* Cartesian maps, the polar-perspective map provides information with specific angular resolution and variable range resolution in polar (r, θ).

3. Camera pose and road surface modeling

Most pedestrian detection approaches detect obstacles under the assumption of a planar road and no camera pose variations, i.e. static camera height, pitch and roll with respect to the ground plane. However, this simplification is not applicable to real world application given the road variability in urban scenarios and the changes in vehicle dynamics. Dynamic camera pose estimation is needed due to car acceleration/breaking, road imperfections, speed bumps, etc., that introduce variations in the extrinsic parameters between the cameras and the road plane just under the camera. Fig. 3 depicts an example where a the vehicle is driving on a speed bump. As can be observed the speed bump considerably affects both camera height and pitch parameters. On the other hand, road surface should be modeled to deal with uphill, downhill and undulating hill (note that the pedestrian appearing in Fig. 3 is located on an uphill).

The first and simplest approach to deal with camera pose estimation and road modeling consists in considering fixed but known camera pose (after some calibration process) and perform linear fitting using the Hough transform in the *v-disparity image* (Broggi et al., 2003; Grubb et al., 2004; van der Mark and Gavrila, 2006; Krotosky and Trivedi, 2007). The road surface profile can be easily extracted from the *v-disparity image* since the road appears as a dominant line feature (see Fig. 1d). This method was first proposed by Labayrade et al. (2002) by fitting the bounding line of piecewise linear functions including the estimation of the camera pitch and height. Linear fitting has been also applied using the Euclidean space in world coordinates, e.g., the *yOz* plane (Nedevschi et al., 2004; Sappa et al., 2008; Llorca et al., 2009), as well as the *virtual disparity image* (Suganuma and Fujiwara, 2007; Llorca et al., 2009) for both road surface modeling and camera pitch estimation. In some cases, linear fitting is only applied to model the vehicle vicinity (Nedevschi et al., 2004; Oniga and Nedevschi, 2010) or for obtaining a first guess of the camera tilt angle (Nedevschi et al., 2004; Wedel et al., 2009).

Some authors have reported estimation problems when fitting in the *v-disparity* space due to the non-linear relationship between depth and disparity (Sappa et al., 2008; Wedel et al., 2009). To overcome this non-linear effect they propose to fit measurements in world coordinates using the aforementioned *yOz* projection maps.

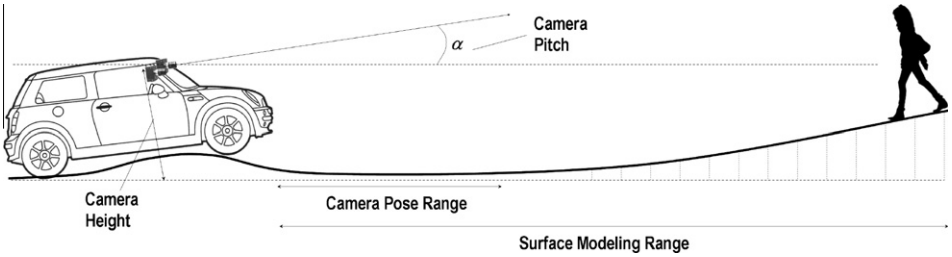

Fig. 3. Example of variations in the camera parameters (pitch angle and camera height) due to a speed bump. The pedestrian ahead is on an uphill. A short area is usually used to estimate camera pose variations. A longer range is used when modeling the road surface profile.

A quadratic approximation of the road surface was proposed by Oniga and Nedevschi (2010) by using a RANSAC approach in the *yOz* plane. According to Nedevschi et al. (2004), a clothoid approximation of the ground surface is proposed. These techniques can only model road surface slope changes into one direction, failing when approximating the road surface if the road is undulating. A general technique that represents the road surface as a parametric B-spline curve has been presented by Wedel et al. (2009), and tested for pedestrian detection by Keller et al. (2009, 2011b), where a Kalman filter is used to track the surface parameters over time.

Concerning lateral surface modeling or camera roll variations by means of stereo vision two main approaches have been proposed by means of several projections of the data provided by the *v-disparity image* (Labayrade and Aubert, 2003) and the *virtual disparity map* (Suganuma and Fujiwara, 2007). However, most authors agree when considering roll camera changes as negligible compared with other extrinsic parameters.

Fig. 4 depicts some examples of the different road modeling approaches computed in the Euclidean space (*yOz* plane). According to what is observed linear fitting yields accurate estimation but only in the vehicle vicinity (see Fig. 4a). The envelope of the surface by means of piecewise linear functions generates non-continuous and abrupt slope changes (Labayrade et al., 2002) (see Fig. 4b). The quadratic approximation only allows slope changes in one direction (Oniga and Nedevschi, 2010) (see Fig. 4c). That is not the case of the clothoid approximation (Nedevschi et al., 2004) (see Fig. 4d). However, in both cases a Hough transform in the vehicle vicinity has to be used since these approximations provide unstable results in the vehicle vicinity (Oniga and Nedevschi, 2010; Nedevschi et al., 2004). The use of a general parametric B-spline curve (Wedel et al., 2009) provides accurate estimation in the vehicle vicinity as well as for large distances, allowing road surface modeling in cases where the road is undulating (see Fig. 4e). Note that these approximations require a dense reconstruction of the scene for an accurate estimation of the road profile.

Common stereo vision-based obstacle detection algorithms detect obstacles (including pedestrians) by evaluating the height above ground, since obstacles can be considered as something that stands in the way. Once the camera pose is estimated and the road surface modeled, 3-D measurements corresponding to obstacles are obtained by removing points under the actual road profile and points under the actual road profile plus the maximum pedestrian height which has to be defined taking into account that pedestrians may appear over raised crosswalks, speed bumps, sideways, etc. The benefits of the different camera pose estimation and road surface modeling approaches for pedestrian detection come from the fact that more accurate road-obstacle segmentation can be achieved since the assumptions of no camera pose changes and a planar road surface are usually violated especially in urban scenarios. However, if previous approaches fail in providing unrealistic results, pedestrian detection performance would be drastically reduced. Accordingly, reliability on the road profile estimation is an important issue which has to be considered for real implementations. This topic has been addressed by Keller et al. (2009, 2011b), by rejecting measurements with high variance and measurements outside the limits of a pre-defined region of interest. If a reliable estimate is not possible, the planar ground assumption will be used instead.

4. Regions of interest: selection mechanisms

ROI selection stage, which is also referred to as candidate selection stage, is responsible for providing regions of interest (bounding boxes) from the image to be sent to further classification or verification modules. The selected ROIs have to correspond to possible pedestrians to ensure the highest detection rate at the lowest false positive rate. In other words, this stage has to assure that no pedestrians are missed. Otherwise the subsequent modules will not be able to detect these pedestrians. In addition, it has to provide a reduced number of candidates avoiding regions such as shadows or reflections as well as regions where it is not possible to find pedestrians (the sky, buildings, etc.). A reduced number of generated ROIs implies fewer computations in later stages of subsequent detection modules, and thus faster processing speed (approximately linear in number of ROIs).

Monocular approaches 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). The simplest monocular candidate selection technique is the sliding window technique, where detector windows at various locations and scales are shifted over the whole image. This approach does not fulfill real-time requirements. However, significant speed-ups can be obtained by either restricting the search space, based on known camera geometry and prior knowledge about the target size, or coupling the sliding window approach with a cascade classifier. Other monocular techniques obtain ROIs employing features derived from the image data such as optical flow or interest points. For a more detailed description and references concerning monocular ROIs selection approaches, we refer to (Gandhi and Trivedi, 2007; Enzweiler and Gavrila, 2009; Gerónimo et al., 2010a).

The first stereo approach for ROI selection was proposed by Zhao and Thorpe (2000) where a connected components analysis is used to segment objects in the *disparity map*. In this case the segmentation algorithm provides very little information about the scene and the road surface which is not recovered from the 3-D measurements.

The *v-disparity map* has been widely used for obstacles detection in the context of intelligent vehicles (Broggi et al., 2003; Grubb et al., 2004; van der Mark and Gavrila, 2006; Krotosky and Trivedi, 2007). Once the estimated road surface is removed from the *v-disparity image*, generic obstacles can be detected by scanning each column and summing the histogram value above the ground plane. If this sum is greater than a configurable threshold, the region is selected spans from the ground plane to the highest point in the column where the histogram entry exceeds the given threshold (Krotosky and Trivedi,

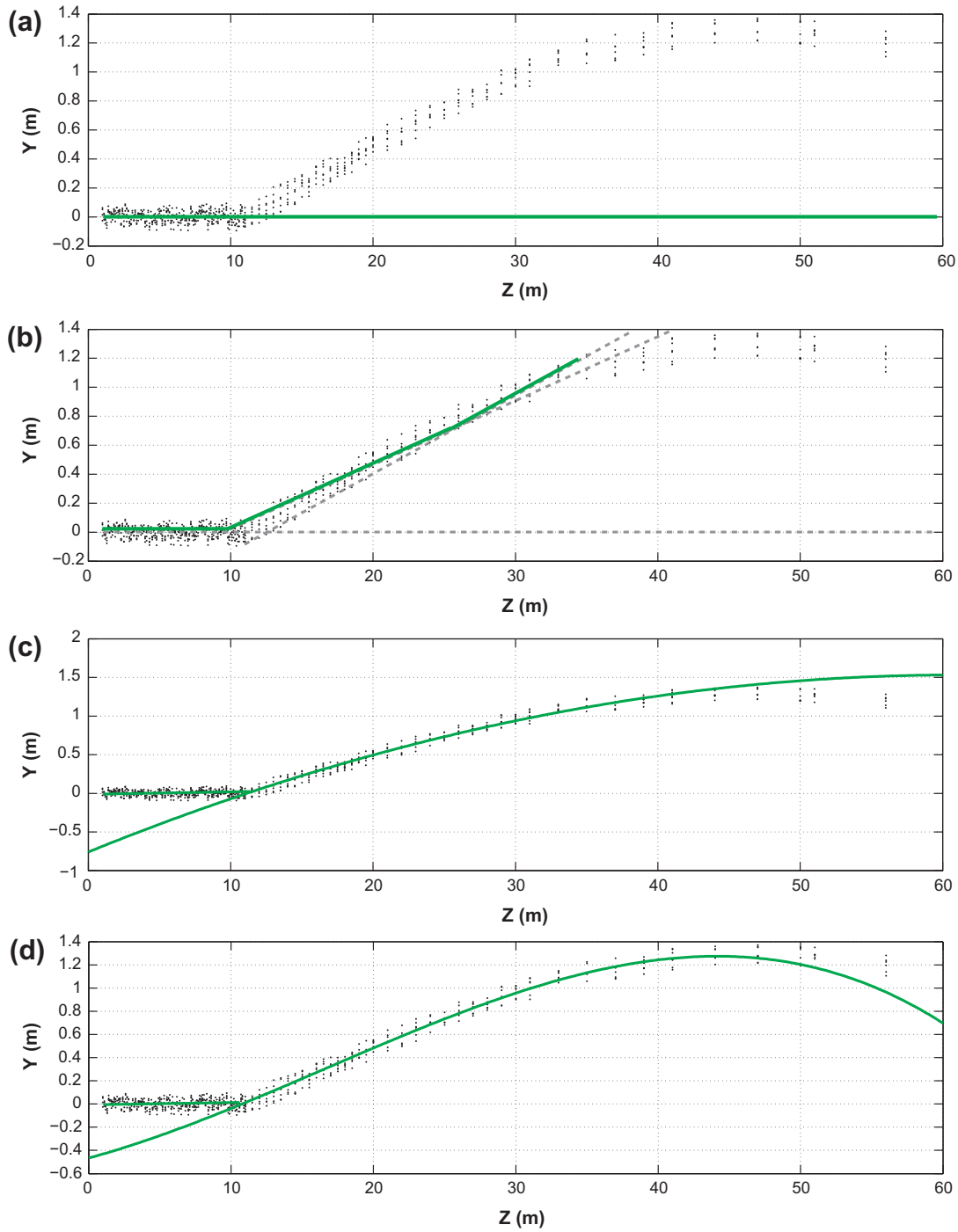

Fig. 4. Different road surface modeling approaches applied in the Euclidean space. (a) Linear fitting in the vehicle vicinity (Broggi et al., 2003; Grubb et al., 2004; van der Mark and Gavrila, 2006; Krotosky and Trivedi, 2007); (b) piecewise planar modeling (Labayrade et al., 2002); (c) Hough in the vicinity and a quadratic fit for large distances (Oniga and Nedevschi, 2010); (d) Hough in the vicinity and a clothoid (cubic) fit for large distances (Nedevschi et al., 2004); and (e) B-spline road surface modeling (Wedel et al., 2009).

2007). However, each vertical line in the *v-disparity map* may correspond with several obstacles located at the same depth, as can be observed in Fig. 5a. Accordingly, further analysis is needed.

According to Broggi et al. (2003), the depth features selected as an obstacle in the *v-disparity map* are projected in the original image. Then, horizontal and vertical histograms are used to identify the borders of the obstacles. According to Hu et al. (2005) and Krotosky and Trivedi (2007), ROIs are also detected in the *u-disparity map* by identifying continuous spans

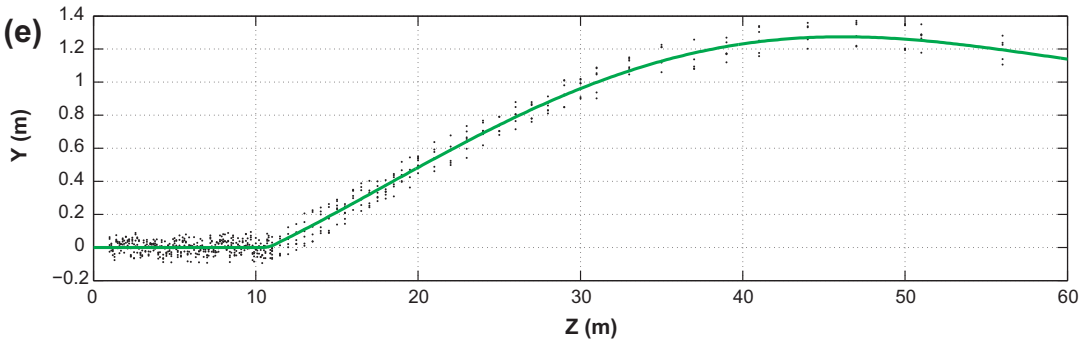

Fig. 4 (*continued*)

along rows where the histogram values exceed a given threshold. Candidate bounding boxes are finally selected by associating the ROIs in the *u*- and *v-disparity* images based on their disparity values (see Fig. 5a). A similar approach using the *virtual disparity image* is presented in Suganuma and Fujiwara (2007) for obstacle detection.

In the context of the well-known PROTECTOR project (Gavrila and Munder, 2007), the following stereo ROI selection process is utilized (Franke and Joos, 2000): the disparity image is multiplexed into N discrete depth ranges. The associated images are scanned with windows related to minimum and maximum extents of pedestrians while taking into account the ground plane location at a particular depth range and appropriate pitch tolerances. The locations where the number of depth features exceed a threshold are added to the ROI list for the subsequent modules. This approach, which was applied using sparse disparity maps by Gavrila and Munder (2007), has been used by Keller et al. (2009) with dense disparity maps and the road modeling method described by Wedel et al. (2009), obtaining a detection performance improvement of factor five. The same scanning window technique can be also applied directly to the 3-D points (Gerónimo et al., 2010b) (see Fig. 5c). In this case the question of how to distribute windows all over the detection range arises. Image-based uniform scanning schemes involve non-uniform scanning in the world coordinates and vice versa. In Gerónimo et al. (2010b) a non-uniform sampling is proposed as a trade-off between uniform scanning in image and world coordinates. Other approaches utilize a polar-perspective map and obstacles are segmented, finding peaks in the map (Bajracharya et al., 2009).

Based on the idea that obstacles (including pedestrians) have a higher density of 3-D points than the road surface, ROI selection can be carried out by determining those positions in world coordinates where there is a high concentration of

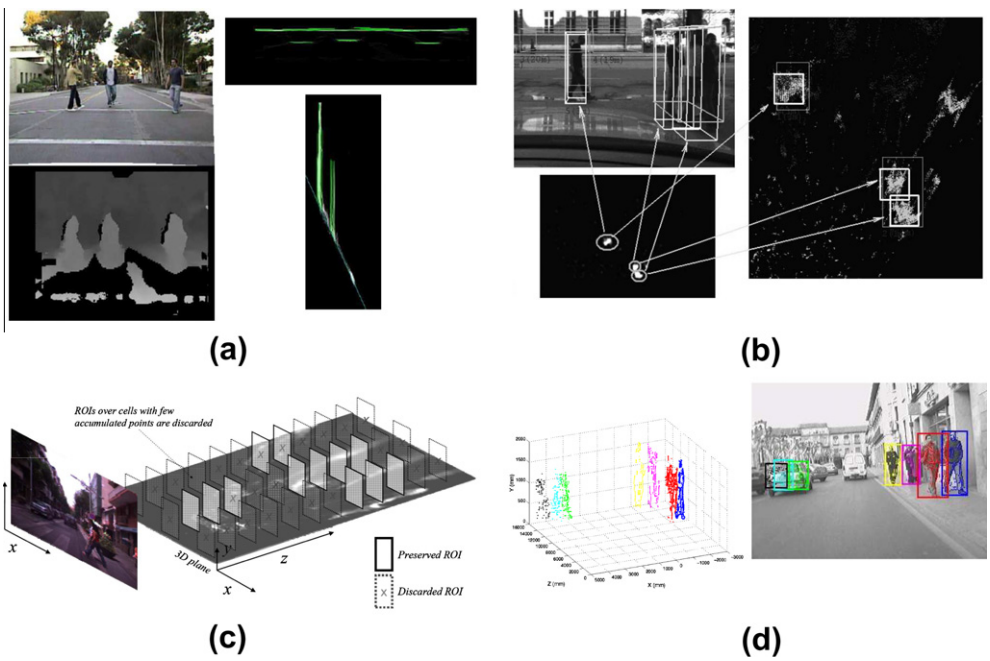

Fig. 5. (a) ROI selection using *u*- and *v-disparity* images (figure from Krotosky and Trivedi (2007)). (b) Region-growing using *xOz* projection map (figure from Nedevschi et al. (2009)). (c) ROI scanning using 3-D coordinates distributed in cells (figure from Gerónimo et al. (2010b)). (d) Adaptive 3-D subtractive clustering for ROI selection.

3-D points. According to Alonso et al. (2007) and Llorca et al. (2009), a *3-D subtractive clustering* method is proposed to deal with the ROI selection stage using sparse data. The idea is to find high-density regions, which are roughly modeled by a single 3-D Gaussian distribution, in the Euclidean space. The parameters of each Gaussian distribution are defined according to minimum and maximum extents of pedestrians. Thus, whereas pedestrians are correctly selected, bigger obstacles such as vehicles or groups of pedestrians are usually split in two or more parts (see Fig. 5d).

According to Nedevschi et al. (2009), and based on the same density idea, a region-growing algorithm is performed on a density map which is computed using the *xOz* projection map (bird's eye view). The density map is an accumulation buffer that integrates and smoothes the *xOz* projection points. The region-growing threshold is defined according to the total amount of 3-D points (see Fig. 5b).

The discrete nature of the stereo sensor generates a depth accuracy that is directly proportional to the focal length and the distance between the cameras, and inversely proportionate to the squared of the disparity. Thus, 3-D points that correspond to an obstacle become much more sparse along the depth axis as we move away from the stereo pair. If ROI selection mechanisms that use world coordinates to detect generic obstacles are not adapted, they are prone to report false negatives with distant obstacles. The introduction of the stereo accuracy in the ROI selection process was first proposed by Fernández et al. (2007) by means of the so-called *adaptive 3-D subtractive clustering*, where the Gaussian variance in the *z*-axis of the density computation is adapted to the stereo uncertainty at a given 3-D point. Thus, cluster densities are corrected to enhance the density of distant obstacles. A similar approach including the correlation accuracy in the stereo uncertainty was proposed by Gerónimo et al. (2010a), to re-weight the number of points distributed in world coordinates cells. According to Nedevschi et al. (2009), the cell size when computing the density maps, is pre-defined depending on the range. As they declare, a better approach would be to consider a probabilistic error model for stereo reconstruction and compute the cell size depending on that model.

Other relevant problems are how to handle partial occlusions, e.g., a pedestrian behind an object such a car, and low contrast depth pedestrians, e.g., a pedestrian in front of a stationary object such a wall. These topics have been studied focusing on the classification stage (Wang et al., 2009; Enzweiler et al., 2010). However, from the ROI selection perspective they have been somewhat neglected in the literature. When these pedestrians are moving, they can be detected by fusing stereo and optical flow simultaneously (Rabe and Franke, 2007). Generic obstacles are then identified as groups of contiguous coherent 3D motion vectors using the Mahalanobis distance as a similarity measure in the cluster analysis (Keller et al., 2011a). Detection of partially occluded or low contrast depth static pedestrians remains as one of most difficult tasks for the ROI selection stage.

5. Discussion

Stereo-based region of interest selection approaches for pedestrian detection are summarized and grouped in Table 1, where a taxonomy of approaches is shown. As can be observed, this specific topic, stereo-based ROI selection, which corresponds to one specific stage of a wider system, vision-based pedestrian protection systems, raises a considerable number of research studies. The first stereo-based approach (Zhao and Thorpe, 2000) is dated from 2000, so more than a decade has passed since this topic appeared in the literature. These statements demonstrate the importance of the subject.

Due to computational limitations, the first real-time stereo-based studies were developed using sparse stereo. However, with recent hardware and software advances, real-time dense stereo is now feasible providing a more detailed description of the scene and introducing new benefits, especially in the context of road surface modeling. Most of the works that use sparse stereo can be applied using dense stereo without losing performance. However, approaches that use dense stereo may not work with a sparse representation of the scene.

Concerning dynamic camera pose computation and road surface modeling, there is not a well established framework for evaluating the different methods available in the literature. First approaches were all based on linear fitting on the *v-disparity map* by means of Hough transform, Least-squares or RANSAC. In recent works the use of the Euclidean space in world coordinates has provided very good results. Based on our experience, the Wedel et al. (2009) approach seems to be the most general and accurate for road surface modeling, especially when it is enriched with the outlier removal step (Keller et al., 2011b). However, we cannot provide quantitative comparison between the different approaches. Up to our knowledge, there are only two works in the literature that report a comparative study using different system configurations: sparse/stereo, fixed/adaptively computed camera pose, and flat world assumption/road B-Spline modeling (Keller et al., 2009, 2011b). For the case of flat world assumption and fixed camera parameters, sparse and dense stereo provided equal ROI generation performance. The estimation of the camera pose (camera height and pitch angle) resulted in a performance improvement of factor three: reduced false positives at same detection rate. When estimating road surface as well, the benefit increased by a factor of five. These are more or less intuitive results that should experiment variations depending on the type of roads, but they clearly demonstrate the importance of using camera pose estimation and road modeling for pedestrian detection. Note that reducing the number of false positives implies fewer computations in later classification and tracking stages.

It is important to remark that although there are a considerable number of stereo-based approaches for pedestrian detection for intelligent vehicles applications, there is a strong lack of comparisons due to two main reasons: the lack of public stereo benchmarking and the difficulty of reproducing many of the proposed methods. Up to now, all the stereo ROI selection approaches have been developed and tested without any comparison to other state-of-the-art proposals. This is not the case of the classification stage, where a considerable number of databases are available with tens of thousands of training and test

Table 1
Taxonomy of stereo-based ROI selection approaches for pedestrian detection.

A. *3-D reconstruction*
   a. Sparse (Franke and Joos, 2000; Broggi et al., 2003; Gavrila and Munder, 2007; Alonso et al., 2007; Llorca et al., 2009)
   b. Dense (Zhao and Thorpe, 2000; Grubb et al., 2004; Krotosky and Trivedi, 2007; Nedevschi et al., 2009; Bajracharya et al., 2009; Keller et al., 2009; Gerónimo et al., 2010b; Keller et al., 2011b)

B. *3-D representation*
   a. *Disparity* image (Zhao and Thorpe, 2000; Gavrila and Munder, 2007; Keller et al., 2011b)
   b. *u-* and *v-disparity* image (Broggi et al., 2003; Grubb et al., 2004; Krotosky and Trivedi, 2007)
   c. *Virtual disparity* image (Suganuma and Fujiwara, 2007; Llorca et al., 2009)
   d. *xOz* projection maps (Nedevschi et al., 2009)
   e. *yOz* projection maps (Nedevschi et al., 2004; Fernández et al., 2007; Sappa et al., 2008; Llorca et al., 2009; Nedevschi et al., 2009; Gerónimo et al., 2010b)
   f. 3-D polar-perspective maps (Bajracharya et al., 2009)
   g. 3-D Euclidean space (Alonso et al., 2007; Llorca et al., 2009; Gerónimo et al., 2010b)

C. *Camera pose and road surface modeling*
   a. Linear (Broggi et al., 2003; Grubb et al., 2004; Gavrila and Munder, 2007; Alonso et al., 2007; Krotosky and Trivedi, 2007; Sappa et al., 2008; Llorca et al., 2009; Nedevschi et al., 2009; Bajracharya et al., 2009; Gerónimo et al., 2010b)
   b. Piecewise planar (Labayrade et al., 2002)
   c. Quadratic (Oniga and Nedevschi, 2010)
   d. Clothoid (Nedevschi et al., 2004)
   e. B-Spline (Wedel et al., 2009; Keller et al., 2009, 2011b)

D. *Stereo ROI selection*
   a. *Disparity* image segmentation (Zhao and Thorpe, 2000)
   b. *u-* and *v-disparity* image segmentation (Broggi et al., 2003; Grubb et al., 2004; Hu et al., 2005; Krotosky and Trivedi, 2007; Gavrila and Munder, 2007)
   c. *Virtual disparity* image segmentation (Suganuma and Fujiwara, 2007)
   d. Multiplexed depth map (Franke and Joos, 2000; Gavrila and Munder, 2007; Keller et al., 2009, 2011b)
   e. Non-uniform multiplexed 3-D Euclidean map (Gerónimo et al., 2010b)
   f. Region-growing on the *xOz* projection map (Nedevschi et al., 2009)
   g. Adaptive 3-D Subtractive clustering in the Euclidean space (Alonso et al., 2007; Llorca et al., 2009; Llorca et al., 2011; Milanés et al., 2012)
   h. Fusion of stereo and optical flow (Rabe and Franke, 2007; Keller et al., 2011a)
   i. Stixel representation (Enzweiler et al., 2012)

samples (Dalal and Triggs, 2005; Papageorgiou and Poggio, 2000; Munder and Gavrila, 2006; Gerónimo et al., 2007), including partially occluded pedestrians (Enzweiler et al., 2010). However, classification results are not clarifying when we want to know the overall performance of a pedestrian protection system. ROI selection stage plays a key role in the pedestrian detection pipeline since it is responsible for not losing a pedestrian (false negative rate approximately equal to zero), and maintaining a reasonable low level of false positives. In the last years, Enzweiler and Gavrila (2009) and Dollar et al. (2009) made different benchmarks public that also include monocular sequential images to evaluate both hypothesis generation and tracking components. Both monocular data sets appeared in 2009, but up to our knowledge the lack of monocular ROI-selection performance comparison studies still remains. This is mainly caused by the aforementioned difficulty of reproducing the proposed algorithms.

The need of similar stereo-based benchmarking has been recently amended thanks to the work presented by Keller and Gavrila (2011), where the same 27-min sequence provided by Enzweiler and Gavrila (2009) has been made publicly available including stereo image pairs to allow the computation of distance data using different stereo implementations. Vehicle velocity and yaw-rate measurements are also provided to enable integration into a tracking and decision making scheme. Nevertheless, the same difficulty related with reproducing stereo ROI algorithms applies here.

Despite the problems stated above, in the following years, considerable research is expected in the context of stereo-based pedestrian protection, as it happened in other areas like face detection or document analysis once well-established databases and benchmarking protocols were available. Future trends will probably include a considerable number of new studies that will be focused on performance comparisons. Fast implementations of camera pose estimation, road surface modeling and candidates ROI selection algorithms will be likely available in the near future. In addition, pedestrian detection systems will benefit from recent improvements in the stereo accuracy (Gehrig et al., 2012). The fusion of stereo with optical flow seems to be a relevant approach for future improvements, especially for difficult cases such as partially occluded and low contrast depth pedestrians. Finally, stereo-based pedestrian protection systems will be a fundamental component of new pedestrian collision avoidance applications (Llorca et al., 2011; Keller et al., 2011a).

6. Concluding remarks

Vision-based pedestrian detection for intelligent vehicles applications is a crucial topic due to the potential benefits in terms of reducing the number of accidents involving pedestrians and vehicles. During the last decade a considerable amount of research studies have been proposed and the technology is almost prepared for being applied in commercial vehicles.

In this paper we presented a survey covering recent work on stereo-based pedestrian detection. In contrast with available pedestrian protection surveys (Gandhi and Trivedi, 2007; Enzweiler and Gavrila, 2009; Gerónimo et al., 2010a), we have focused on one specific stage of the pipeline (stereo-based ROI selection) to provide a more detailed and exhaustive review of the different approaches published in the literature. Stereo-based ROI selection involves different components that have been analyzed independently, increasing visibility for future proposals and developments.

Although the technology is very mature, new studies are still needed to better understand the actual performance of the different approaches. In the following years and thanks to the appearance of new stereo databases, a substantial number of studies, including benchmarking and comparison analysis, are expected to appear. This is clearly a hot topic in which there is still room for improvements. We expect that the gap between current systems performance and their implementation in commercial vehicles will be definitively closed in the short term.

Acknowledgment

This work has been supported by the Spanish Ministry of Science and Innovation by means of Research Grant ONDA-FP TRA2011-27712-C02-02.

References

Sappa, A., Dornaika, F., Ponsa, D., Gerónimo, D., López, A., 2008. An efficient approach to onboard stereo vision system pose estimation. IEEE Transactions on Intelligent Transportation Systems 9 (3), 476–490.

Alonso, I.P., Llorca, D.F., Sotelo, M.A., Bergasa, L.M., de Toro, P.R., Nuevo, J., na, M.O., Garrido, M.A.G., 2007. Combination of feature extraction methods for SVM pedestrian detection. IEEE Transactions on Intelligent Transportation Systems 8 (2), 292–307.

Bajracharya, M., Moghaddam, B., Howard, A., Brennan, S., Matthies, L.H., 2009. Results from a real-time stereo-based pedestrian detection system on a moving vehicle. In: Proc. of the IEEE International Conference on Robotics and Automation.

Bertozzi, M., Binelli, E., Broggi, A., Rose, M.D., 2005. Stereo vision-based approaches for pedestrian detection. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05).

Broggi, A., Fascioli, A., Fedriga, I., Tibaldi, A., Rose, M.D., 2003. Stereo-based preprocessing for human shape localization in unstructured environments. In: Proc. IEEE Intelligent Vehicles Symposium.

Keller, C., Enzweiler, M., Gavrila, D.M., 2011. A new benchmark for stereobased pedestrian detection. In: Proc. IEEE Intelligent Vehicles Symposium.

Rabe, C., Franke, U., Gehrig, S., 2007. Fast detection of moving objects in complex scenarios. In: Proc. IEEE Intelligent Vehicles Symposium.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: International Conference on Computer Vision and Pattern Recognition, pp. 886–893.

Dollar, P., Wojek, C., Schiele, B., Perona, P., 2009. Pedestrian detection: a benchmark. In: Proc. of the IEEE Conf. Computer Vision and Pattern Recognition.

Enzweiler, M., Eigenstetter, A., Schiele, B., Gavrila, D.M., 2010. Multi-cue pedestrian classification with partial occlusion handling. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition.

Enzweiler, M., Gavrila, D.M., 2009. Monocular pedestrian detection: survey and experiments. IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (12), 2179–2195.

Enzweiler, M., Hummel, M., Pfeiffer, D., Franke, U., 2012. Efficient stixel-based object recognition. In: Proc. IEEE Intelligent Vehicles Symposium.

Fernández, D., Parra, I., Sotelo, M.A., Revenga, P., Álvarez, S., Gavilán, M., 2007. 3D candidate selection method for pedestrian detection on non-planar roads. In: Proc. IEEE Intelligent Vehicle Symposium.

Forsyth, D.A., Ponce, J., 2003. Computer Vision: A Modern Approach. Prentice Hall.

Franke, U., Joos, A., 2000. Real-time stereo vision for urban traffic scene understanding In: Proc. IEEE Intelligent Vehicles Symposium.

Fusiello, A., Trucco, E., Verri, A., 1998. Rectification with unconstraint stereo geometry. In: Proc. of the Eight British Machine Vision Conference.

Gandhi, T., Trivedi, M.M., 2007. Pedestrian protection systems: Issues, survey, and challenges. IEEE Transactions on Intelligent Transportation Systems 8 (3), 413–430.

Gavrila, D.M., 1999. The visual analysis of human movement: a survey. Computer Vision and Image and Understanding 73 (1), 82–89.

Gavrila, D.M., Munder, S., 2007. Multi-cue pedestrian detection and tracking from a moving vehicle. International Journal of Computer Vision 73 (1), 41–59.

Gehrig, S.K., Badino, H., Franke, U., 2012. Improving sub-pixel accuracy for long range stereo. Computer Vision and Image Understanding 116 (1).

Gerónimo, D., López, M.L., Sappa, A.D., Graft, T., 2010a. Survey on pedestrian detection for advanced driver assistance systems. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (7), 1239–1258.

Gerónimo, D., Sappa, A.D., López, A., Ponsa, D., 2007. Adaptive image sampling and windows classification for on-board pedestrian detection. In: Proc. of Fifth International Conf. Computer Vision Systems.

Gerónimo, D., Sappa, A.D., López, A.M., 2010b. Stereo-based candidate generation for pedestrian protection systems. In: Binocular Vision: Development, Depth Perception and Diorders. NOVA Publishers (Chapter 9).

Grubb, G., Zelinsky, A., Nilsson, L., Rilbe, M., 2004. 3D vision sensing for improved pedestrian safety. In: Proc. IEEE Intelligent Vehicles Symposium.

Hu, Z., Lamosa, F., Uchimura, K., 2005. A complete u–v-disparity study for stereovision based 3D driving environment analysis. In: Proc. of the IEEE Fifth International Conference on 3-D Digital Imaging and Modelling.

Keller, C.G., Dang, T., Fritz, H., Joos, A., Rabe, C., Gavrila, D.M., 2011a. Active pedestrian safety by automatic braking and evasive steering. IEEE Transactions on Intelligent Transportation Systems 12 (4), 1292–1304.

Keller, C.G., Enzweiler, M., Rohrbach, M., Llorca, D.F., Schnör, C., Gavrila, D.M., 2011b. The benefits of dense stereo for pedestrian detection. IEEE Transactions on Intelligent Transportation Systems. http://dx.doi.org/10.1109/TITS.2011.2143410.

Keller, C.G., Llorca, D.F., Gavrila, D.M., 2009. Dense stereo-based ROI generation for pedestrian detection. In: Proc. DAGM Symp. Pattern Recog. LNCS 5749, pp. 81–90.

Krotosky, S., Trivedi, M., 2007. On color-, infrared-, and multimodal-stereo approaches to pedestrian detection. IEEE Transactions on Intelligent Transportation Systems 8 (4), 619–629.

Labayrade, R., Aubert, D., 2003. A single framework for vehicle roll, pitch, yaw estimation and obstacles detection by stereovision. In: Proc. IEEE Intelligent Vehicle Symposium.

Labayrade, R., Aubert, D., Tarel, J.P., 2002. Real time obstacle detection in stereovision on non-flat road geometry through v-disparity representation. In: Proc. IEEE Intelligent Vehicles Symposium.

Llorca, D.F., Milanés, V., Alonso, I.P., Gavilán, M., Daza, I.G., Pérez, J., Sotelo, M.A., 2011. Autonomous pedestrian collision avoidance using a fuzzy steering controller. IEEE Transactions on Intelligent Transportation Systems 12 (2), 390–401.

Llorca, D.F., Sotelo, M.A., Parra, I., na, M.O., Bergasa, L.M., 2010. Error analysis in a stereo vision-based pedestrian detection sensor for collision avoidance applications. Sensors 10 (4), 3741–3758.

Llorca, D.F., Sotelo, M.A., Parra, I., Naranjo, J.E., Gavilán, M., Álvarez, S., 2009. An experimental study on pitch compensation in pedestrian-protection systems for collision avoidance and mitigation. IEEE Transactions on Intelligent Transportation Systems 10 (3), 469–474.
Milanés, V., Llorca, D.F., Villagrá, J., Pérez, J., Fernández, C., Parra, I., González, C., Sotelo, M.A., 2012. Intelligent automatic overtaking system using vision for vehicle detection. Expert Systems with Applications 39, 3362–3373.
Moeslund, T.B., Granum, E., 2006. A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image and Understanding 103 (2–3), 90–126.
Munder, S., Gavrila, D.M., 2006. An experimental study on pedestrian classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (11), 1863–1868.
Nedevschi, S., Bota, S., Tomiuc, C., 2009. Stereo-based pedestrian detection for collision-avoidance applications. IEEE Transactions on Intelligent Transportation Systems 10 (3), 380–391.
Nedevschi, S., Danescu, R., Frentiu, D., Marita, T., Oniga, F., Pocol, C., Schmidt, R., Graf., T., 2004. High accuracy stereo vision system for far distance obstacle detection. In: Proc. IEEE Intelligent Vehicles Symposium.
Oniga, F., Nedevschi, S., 2010. Processing dense stereo data using elevation maps: road surface, traffic isle, and obstacle detection. IEEE Transactions on Vehicular Technology 59 (3), 1172–1182.
Papageorgiou, C., Poggio, T., 2000. A trainable system for object detection. International Journal of Computer Vision 38 (1), 15–33.
Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A.A., Jarawan, E., Mathers, C., 2004. World Report on Road Traffic Injury Prevention. World Health Organization, Geneva, Switzerland.
Poppe, R., 2007. A vision-based human motion analysis: an overview. Computer Vision and Image and Understanding 108 (1), 4–18.
Shashua, A., Gdalyahu, Y., Hayun, G., 2004. Pedestrian detection for driving assistance systems: single-frame classification and system level performance. In: Proc. IEEE Intelligent Vehicles Symposium.
Stein, G.P., Gdalyahu, Y., Shashua, A., 2010. Stereo-assist: top-down stereo for driver assistance systems. In: Proc. of the IEEE Intelligent Vehicles Symposium.
Suganuma, N., Fujiwara, N., 2007. An obstacle extraction method using virtual disparity image. In: Proc. IEEE Intelligent Vehicles Symposium.
UNECE, 2005. Pedestrian safety global technical regulation preamble. In: Transport Division. World Forum for Harmonization of Vehicle Regulations.
van der Mark, W., Gavrila, D.M., 2006. Real-time dense stereo for intelligent vehicles. IEEE Transactions on Intelligent Transportation Systems 7 (1), 38–50.
Wang, X., Han, T.X., Yan, S., 2009. A HOG-LBP human detector with partial occlusion handling. In: Proc. of the IEEE 12th International Conference on Computer Vision.
Wedel, A., Badino, H., Rabe, C., Loose, H., Franke, U., Cremers, D., 2009. B-spline modeling of road surfaces with an application to free-space estimation. IEEE Transactions on Intelligent Transportation Systems 10 (4), 572–583.
Zhao, L., Thorpe, C.E., 2000. Stereo and neural network-based pedestrian detection. IEEE Transactions on Intelligent Transportation Systems 1 (3), 148–154.
Zhencheng, H., Lamosa, F., Uchimura, K., 2005. A complete u–v-disparity study for stereovision based 3D driving environment analysis. In: Proc. Fifth International Conference on 3-D Digital Imaging and Modelling.