Computer Vision Approaches to Pedestrian Detection: Visible Spectrum Survey

David Gerónimo, Antonio López, and Angel D. Sappa

Computer Vision Center, Universitat Autònoma de Barcelona
Edifici O, 08193 Bellaterra, Barcelona, Spain
{dgeronimo,antonio,asappa}@cvc.uab.es
www.cvc.uab.es/adas

Abstract. Pedestrian detection from images of the visible spectrum is a high relevant area of research given its potential impact in the design of pedestrian protection systems. There are many proposals in the literature but they lack a comparative viewpoint. According to this, in this paper we first propose a common framework where we fit the different approaches, and second we use this framework to provide a comparative point of view of the details of such different approaches, pointing out also the main challenges to be solved in the future. In summary, we expect this survey to be useful for both novel and experienced researchers in the field. In the first case, as a clarifying snapshot of the state of the art; in the second, as a way to unveil trends and to take conclusions from the comparative study.

1 Introduction

Pedestrian accidents are the second source of traffic injuries and fatalities in the European Union. In this sense, advanced driver assistance systems (ADAS), and specifically pedestrian protection systems (PPS), have become an important field of research to improve traffic safety. Of course, in order to avoid collisions with pedestrians they must be detected, being camera sensors key due to the rich amount of cues and high resolution they provide.

Currently there are two main lines of work, one based on images of the visible spectrum, and the other, mainly motivated by nighttime, based on thermal infrared. The former has accumulated more literature because the easier availability of either CCD or CMOS sensors working in the visible spectrum, their cheaper price, better signal–to–noise ratio and resolution, and because most of the accidents happen at daytime. Therefore, we restrict the discussion presented in this paper to works based on images of the visible spectrum.

In this context, difficulties of the pedestrian detection task for PPS arise both from working with a mobile platform in an outdoor scenario, recurrent challenge in all ADAS applications, and from dealing with a so aspect–changing class like pedestrians. Difficulties can be summarized in the followings: *(a)* targets have a very high intra–class variability; *(b)* background can be cluttered and changes in milliseconds; *(c)* targets and camera usually follow different unknown

J. Martí et al. (Eds.): IbPRIA 2007, Part I, LNCS 4477, pp. 547–554, 2007.

movements; and *(d)* fast system reaction together with a very robust response is required.

The high social relevance of PPS and the above mentioned difficulties have given rise to a number of works. However, due to the lack of common datasets for validation and the complexity of the different proposals, most of the papers present their own approach without comparison with others. Thus, a *comparative review* is of high relevance both for novel and experienced researchers in the field. In this paper, a survey of works with images of the visible spectrum is presented.

Addressing such a review just by summarizing the most relevant papers one by one in isolation would make difficult the comparative viewpoint. Thus, we propose first (Sect. 2) a common framework (i.e., a system architecture) in which the main works of the literature are fitted. This framework is based on the main subtasks of the pedestrian detection for PPS, then we will use it also for providing a critical overview of the described techniques together with the main challenges for the future (Sect. 3). Finally, conclusions are presented in Sect. 4.

2 Proposed Architecture and Literature Review

Figure 1 presents an architecture of modules used in the sequel as common framework to review the literature. This architecture–based review is summarized in Table 1 from the viewpoint of each individual work, while Table 2 provides some relevant details of the previous systems. Although the PPS architecture has six modules here we focus only on the most active ones due to the lack of space: Foreground Segmentation, Object Classification, Verification/Refinement and Tracking. Refer to Fig. 1 to see each module's responsibility.

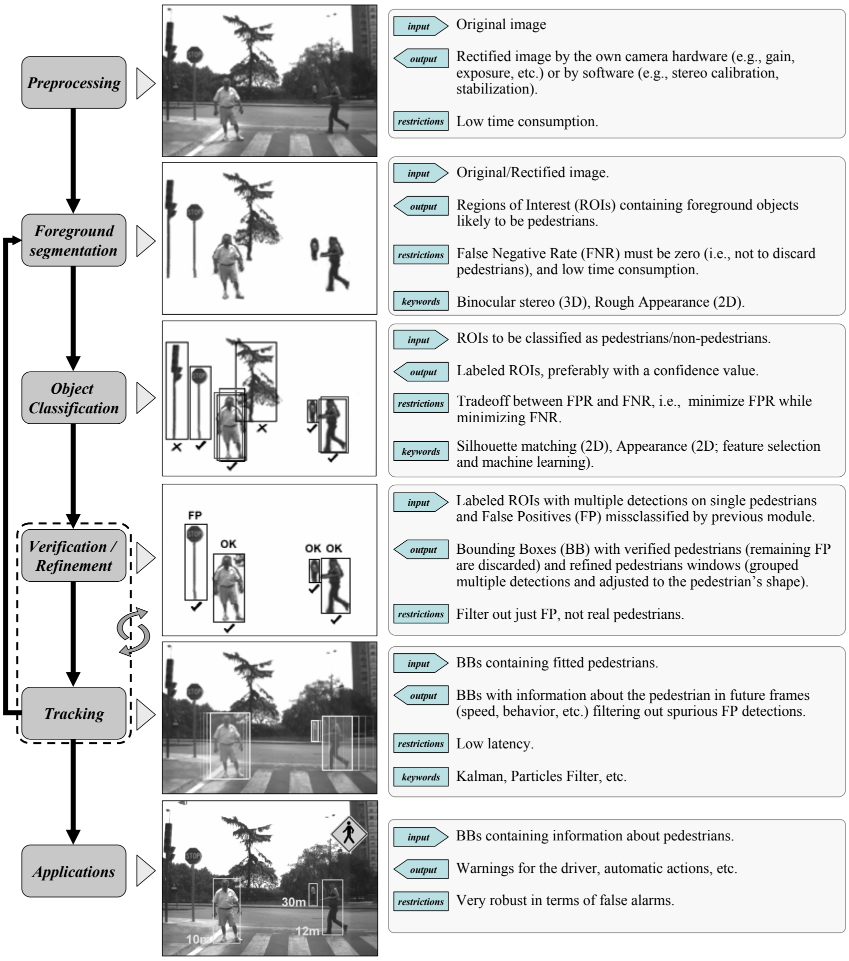
2.1 Foreground Segmentation

Binocular Stereo. The use of stereo in this module aims to provide 2D ROIs corresponding to 3D vertical objects fitting some pedestrian size constraints (PSC). *Gavrila et al.* [1] scan the depth map with PSC–sized ROIs laying in the assumed ground plane. A ROI is accepted if its depth distribution agrees with the expected. *Zhao et al.* [2] apply thresholding, morphological operations and blob analysis to the depth map, selecting remaining PSC–sized blob bounding boxes. *Broggi et al.* [3] use the v–disparity [4] to distinguish between ground, background and vertical objects in the scene.

Rough Appearance. These are 2D approaches. In several works [5,3] by *Broggi et al.* vertical symmetry, derived from grey level and vertical gradient magnitude, is used to select PSC–sized ROIs around each relevant symmetry axis. *Shashua et al.* [6] select PSC–sized ROIs with an expected texture.

2.2 Object Classification

All found approaches fitting object classification are purely 2D, thus they only use the 2D information of the ROIs provided by the foreground segmentation.



input Original image
output Rectified image by the own camera hardware (e.g., gain, exposure, etc.) or by software (e.g., stereo calibration, stabilization).
restrictions Low time consumption.

input Original/Rectified image.
output Regions of Interest (ROIs) containing foreground objects likely to be pedestrians.
restrictions False Negative Rate (FNR) must be zero (i.e., not to discard pedestrians), and low time consumption.
keywords Binocular stereo (3D), Rough Appearance (2D).

input ROIs to be classified as pedestrians/non-pedestrians.
output Labeled ROIs, preferably with a confidence value.
restrictions Tradeoff between FPR and FNR, i.e., minimize FPR while minimizing FNR.
keywords Silhouette matching (2D), Appearance (2D; feature selection and machine learning).

input Labeled ROIs with multiple detections on single pedestrians and False Positives (FP) missclassified by previous module.
output Bounding Boxes (BB) with verified pedestrians (remaining FP are discarded) and refined pedestrians windows (grouped multiple detections and adjusted to the pedestrian's shape).
restrictions Filter out just FP, not real pedestrians.

input BBs containing fitted pedestrians.
output BBs with information about the pedestrian in future frames (speed, behavior, etc.) filtering out spurious FP detections.
restrictions Low latency.
keywords Kalman, Particles Filter, etc.

input BBs containing information about pedestrians.
output Warnings for the driver, automatic actions, etc.
restrictions Very robust in terms of false alarms.

Fig. 1. Proposed module–based architecture and responsibility of each module

Silhouette Matching. In [5] a ROI is considered as containing a pedestrian if there is a good matching with a head–shoulders like binary pattern. A more sophisticated technique is the *Chamfer system* [1,7] by *Gavrila et al.*, where a distance transform of the ROI is calculated and used for a coarse–to–fine template matching in a space of pedestrian shapes hierarchically organized.

Appearance. The methods included here start by defining a space of image features, and then learning a classifier by using ROIs containing examples (pedestrians) and counter–examples (non–pedestrians). A common detail is to normalize the size and aspect ratio of the incoming ROIs, as well as discarding color information (because the variability of clothes).

Two approaches can be found: holistic or parts–based. In the holistic approach a classifier uses different image features to determine if a ROI contains a full

Table 1. Most relevant systems. Each row shows the proposal of the corresponding authors for each module of the architecture of Fig. 1, following our interpretation (note: Q.SVM/L.SVM states for Quadratic/Linear Support Vector Machine).

	Foreground Segmentat.	Object Classification	Verification/ Refinement	Tracking
[1] [7] [8] [9]	Stereo +PSC	Silhouette + Chamfer System Texture + NN-LRF	Stereo, gait pattern also tested	Kalman; particle filters; α-β tracker. Silhouette, texture, stereo, CAN data
[10]	Symmetry +PSC	*(here another possibility was to think about PSC as segmentation and symmetry as classification)*	Stereo + PSC + adhoc image filters, 3D curves matching as well as autonomous agents were also tested	Kalman. Grey-level, Stereo.
[5] [3]	Stereo (v-d) Symmetry +PSC	Silhouette of head and shoulders	Stereo+PSC+entropy	
[6]	Texture +PSC	Components: Gradient Or.*Mag + RR-AdaBoost Different training per pose and illumination conditions	Multi frame after tracking: gait, classification goodness over time, etc., multi-class help suggested	*(used but not detailed)*
[11]	Stereo (v-d)	Gradient magnitude + Q.SVM Different training per pose (Front/Rear or Side viewed)	Classification goodness over time with the help of tracking	Kalman. Stereo
[2]	Stereo+PSC	Gradient magnitude + NN		
[12]	Horizon Line estimation +PSC	Haar + EOH + Real AdaBoost		
[13] [14]		Holistic: Basic Haar + Q.SVM Parts–based: Basic Haar + Q.SVM-L.SVM		Heuristic integration through time
[15]		HOG/Fixed Blocks + L.SVM		

pedestrian. In the parts–based, there is a first stage that searches for predefined different parts (e.g., head, legs and arms) inside of a ROI using different classifiers based on image features. Next, a second stage uses the output of such classifiers as input of a final full pedestrian classifier.

Following the holistic approach, in [1] *Gavrila et al.* propose a classifier that, for the ROIs preclassified as pedestrian for the Chamfer System, uses texture as feature and learning with a Neural Network of Local Receptive Fields (NNLRF, further study in [8]). In [2], the feature used by *Zhao et al.* is gradient magnitude, and a three–layer Feed Forward Neural Network (FFNN) is the learning machine. In [13] a preestablished set of Haar wavelets is used by *Papageorgiou et al.* as features to learn a classifier for front/rear viewed pedestrians with a quadratic Support Vector Machine (SVM). *Dalal and Triggs* [15] present a human classification algorithm that uses Histograms of Oriented Gradients (HOG) as features and a linear SVM as learning method. In order to obtain a classifier for front, rear and side viewed pedestrians *Gerónimo et al.* [12] use a Real AdaBoost learning method to select the best features among a set of Haar wavelets and Edge Orientation Histograms (EOH) that cover all possible scales of a ROI.

Haar wavelets are also used by *Mohan et al.* in a parts–based classification [14]. In this case, each ROI is divided in four parts (head, legs, right and left arms), and for each part a classifier is learned using a quadratic SVM. Then, the final

Table 2. Details of the most relevant classifier based approaches. (DR: Detection Rate, FPR: False Positive Rate, FPPW: False Positives Per Window, n/a: information not available/applicable). Note that training and testing sets are different in each system.

	Learning ROI size	Classifier Train Set	Classifier Test Set	Classifier Performance	System Test Set	System Performance	Detection Range
[1] [7] [8] [9]	(Chamfer) 70–102 pix wide (NN-LRF) 18 × 36	(Chamfer) 1,250pos (NN-LRF) 14,400 pos 15,000 neg	(Chamfer) 900 images (NN-LRF) 9,600 pos 10,000 neg	(Chamfer) 60–90% DR n/a FPR (NN-LRF) 90% DR 10% FPR	24 min driving	(all) 52–76% DR 30% precision (risky) 80–90% DR 75% precision	5–25m
[6]	12 × 36	25,000 pos 25,000 neg	15,244 in total	93.5% DR 8% FPR	5hr driving	(inward moving) 96%DR, 1FPPW (statio. inpath) 93%DR, 3FPPW (statio. outpath) 85%DR,102FPPW	3–25m
[11]	–	1,500 pos 20,000 neg	150 pos 2,000 neg	75% DR 2% FPR	2,500frame (14 pedest.)	83.5% DR 0.4% FPR	up to 30m
[2]	30 × 65	1,012 pos 4,306 neg	254 pos 363 neg	85.4% DR 0.05 % FPR	FGS ROIs	85.2% DR 3.1% FPPW	n/a
[12]	no downscale	700 pos 4,000 neg	300 pos 1,000 neg	90% DR 1% FPR	n/a	n/a	5–50m
[13]	64 × 128	1,848 pos 11,361 neg	123 images scan	(color) 93% DR 0.1% FPPW (grayscale) 83% DR 0.1% FPPW	n/a	n/a	n/a
[14]	64 × 128	889 pos 3,106 neg	12 images scan	96% DR 10^{-4}% FPPW	n/a	n/a	n/a
[15]	64 × 128	2,478 pos 12,180 neg	images scan	85% DR 10^{-4}% FPPW	n/a	n/a	n/a

ROI classification combines the parts–classifiers responses by using a linear SVM. In [6] *Shashua et al.* use thirteen overlapping parts described by SIFT inspired features and ridge regression (RR) to learn the classifier of each part. Moreover, to deal with the high intra–class variability, the training set is divided in nine clusters according to pose and illumination conditions, thus getting $9 \times 13 = 117$ classifiers. The outputs of the 117 classifiers are fed as *weak rules* to an AdaBoost machine that sets the final classification rule.

2.3 Verification/Refinement

In many systems, the methods used in this module take advantage of previously exploited techniques. For instance, in [1], a cross–correlation using the left image and the isolated silhouette computed by the Chamfer System in the right image is used to refine the location of the pedestrian. In [9], the authors suggest to analyse the gait pattern for pedestrians crossing perpendicular to the camera. In this case, the target must be tracked before applying this method, thus verification/refinement and tracking modules are interchanged (Fig. 1). In [5], the head and shoulders silhouette matched during classification is taken as reference to refine the detection until the feet by making use of the vertical edges computed for the symmetry detection. Additionally, since no stereo reasoning is done in the segmentation module, refinement can be improved by this cue.

In [6], *Shashua et al.* propose a *multi–frame approval process*, which consists in validating the pedestrian–classified ROIs by collecting information from several frames: gait pattern, inward motion, confidence of the single–frame classification, etc. In this case, verification comes after tracking too.

2.4 Tracking

Kalman filter is the most used technique for tracking. Two examples are [10], where a Kalman–filter tracker is used to reject spurious detections as well as computing the trajectory of the pedestrian. In [11], Kalman filters are used to maintain pedestrian estimates and Bayesian probability to provide an estimate of pedestrian classification certainty over time and a targets' trajectory and speed.

3 Discussion

In spite of the high number of works in the field and the clear progress achieved, pedestrian detection is still an open area of research. The difficulties this problem carries are so wide that the methods exploited in each module must still improve their robustness before expecting convincing results for the complete system. Next, some discussion is made at each stage of the proposed architecture in order to emphasize the strengths and weaknesses of the described techniques.

Foreground segmentation based on stereo has several advantages: 1) robustness to illumination changes; 2) the provided distances are useful to determine ROIs at the foreground segmentation itself, for tracking and as associated information of the detected pedestrians; 3) stereo information can be shared by different ADAS applications. The main drawbacks come from the high computation time needed to extract depth (considerable improvements are being achieved in this matter [16]) and the problems of the technique when uniform areas appear. Despite these problems, stereo is a very reliable technique. Rough–appearance is not so promising. [6] claims to select just 75 ROIs per frame, but details about the technique are not provided. Moreover, works exploiting vertical symmetry [5,3] tend to be supported by stereo information, so we guess that symmetry alone is not sufficient.

Referring to object classification, it seems clear that silhouette matching methods are not applicable in a stand–alone fashion. Even the very elaborated Chamfer System needs an extra step that follows the appearance–based classification idea. On the other hand, appearance–based seem to be a promising line of research, nowadays still being explored in computer vision. However, despite the improvements in generalization achieved with SVM or AdaBoost, and the more and more faster–to–extract and meaningful features presented in recent years (e.g., Haar, HOG, EOH, etc.), there is still much work to do.

Next we illustrate this with a simple example inspired by [17]. Let us assume 10,000 ROIs per image by using PSC to be classified by the best classifier in the

PPS literature, i.e., a 95% Detection Rate at 0.1% FPR [6] (a full scan would imply at least one million of ROIs for a 640 × 480 image). This means that if 95% of pedestrians have to be detected, then in the worse scenario we could have about 1,000 FP per image, i.e., 25,000 FP per second at 25 fps. Thanks to additional procedures like foreground segmentation or detection clustering, we can assume that this number can be reduced to only 75 ROIs to check *per* frame as suggested in [6], but still this would represent 187.5 FP/s. A tracking module could filter out spurious and not coherent detections to reduce the final number to, say, 1 FP/s. Anyway, even 1 FP/s (i.e., 60 FP/minute) is still useless for a PPS. From this example we see the importance of improving all the system modules, specially the classification rate.

As can be noticed, state–of–the–art classifiers like the HOG–based [15] still need a verification and refinement step. Two points can be highlighted from the this module. First, stereo information tends to be used as long as the classification has been based on the 2D image. However, it is unclear for us why some works do not exploit this 3D information during the foreground segmentation. Second, using verification after tracking seems to be an interesting approach since common movement–based techniques (e.g., gait pattern analysis) used in surveillance could be applied. This

Up to now, the tracking module in PPS has not received as much attention as other modules like segmentation or classification. Each paper has its own proposal and no comparisons have been made. It is clear that tracking could provide useful information for other modules (e.g., trajectory information for applications, potential ROIs for segmentation, etc.).

4 Conclusions

We have presented a review of on–board pedestrian detection works based on images of the visible spectrum. A general module–based architecture is proposed so the reviewed techniques can be fitted and compared according to their objectives and responsibilities in the system, thus providing an comparative snapshot the state of the art. Regarding the future trends, it can be said that object classification is subject to the most active and fruitful research. However, as can be appreciated, the absence of comparisons with common benchmarks and the constant improvement of learning algorithms and features make it hard to state which is the best approach. Finally, it is worth to say in order to achieve commercial performance (e.g., detection rates and timings), the other modules must also be further developed. In this sense, sensors fusion (e.g., visible spectrum cameras with radar) seem to be a promising approach.

Acknowledgments. This work was supported by the Spanish Ministry of Education and Science under project TRA2004-06702/AUT, BES-2005-8864 grant (first author) and Ramón y Cajal Program (third author).

References

1. Gavrila, D., Giebel, J., Munder, S.: Vision–based pedestrian detection: The PROTECTOR system. In: IV, Parma, Italy (2004)
2. Zhao, L., Thorpe, C.: Stereo and neural network–based pedestrian detection. TITS 1(3), 148–154 (2000)
3. Broggi, A., Fascioli, A., Fedriga, I., Tibaldi, A., Del Rose, M.: Stereo–based preprocessing for human shape localization in unstructured environments. In: IV, Columbus, OH, USA, pp. 410–415 (2003)
4. Labayrade, R., Aubert, D., Tarel, J.: Real time obstacle detection in stereovision on non flat road geometry through v–disparity representation. In: IV, Versailles, France (2002)
5. Broggi, A., Bertozzi, M., Fascioli, A., Sechi, M.: Shape–based pedestrian detection. In: IV, Dearborn, MI, USA (2000)
6. Shashua, A., Gdalyahu, Y., Hayun, G.: Pedestrian detection for driving assistance systems: single–frame classification and system level performance. In: IV, Parma, Italy (2004)
7. Gavrila, D.: Pedestrian detection from a moving vehicle. In: ECCV. vol. 2. Dublin, Ireland, pp. 37–49 (2000)
8. Munder, S., Gavrila, D.: An experimental study on pedestrian classification. TPAMI 21(11), 1863–1868 (2006)
9. Franke, U., Gavrila, D.: Autonomous driving goes downtown. IS 13(6), 40–48 (1999)
10. Bertozzi, M., Broggi, A., Fascioli, A., Tibaldi, A., Chapuis, R., Chausse, A.: Pedestrian localization and tracking system with Kalman filtering. In: IV, Parma, Italy, pp. 584–589 (2004)
11. Grubb, G., Zelinsky, A., Nilsson, L., Rilbe, M.: 3D vision sensing for improved pedestrian safety. In: IV, Parma, Italy (2004)
12. Gerónimo, D., Sappa, A., López, A., Ponsa, D.: Pedestrian detection using AdaBoost learning of features and vehicle pitch estimation. In: Proc. of the International Conference on Visualization, Imaging and Image Processing, Palma de Mallorca, Spain, pp. 400–405 (2006)
13. Papageorgiou, C., Poggio, T.: A trainable system for object detection. IJCV 38(1), 15–33 (2000)
14. Mohan, A., Papageorgiou, C., Poggio, T.: Example–based object detection in images by components. TPAMI 23(4), 349–361 (2001)
15. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. vol. 2, San Diego, CA, USA, pp. 886–893 (2005)
16. van der Mark, W., Gavrila, D.: Real–time dense stereo for intelligent vehicles. TITS 7(1), 38–50 (2006)
17. Gavrila, D.: Sensor–based pedestrian protection. IS 16(6), 77–81 (2001)