Computer Vision and Image Understanding 114 (2010) 583–595

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

2D–3D-based on-board pedestrian detection system

David Gerónimo *, Angel D. Sappa, Daniel Ponsa, Antonio M. López

Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

ARTICLE INFO

Article history:
Received 29 December 2007
Accepted 10 July 2009
Available online 10 February 2010

Keywords:
Pedestrian detection
Advanced Driver Assistance Systems
Horizon line
Haar wavelets
Edge orientation histograms

ABSTRACT

During the next decade, on-board pedestrian detection systems will play a key role in the challenge of increasing traffic safety. The main target of these systems, to detect pedestrians in urban scenarios, implies overcoming difficulties like processing outdoor scenes from a mobile platform and searching for aspect-changing objects in cluttered environments. This makes such systems combine techniques in the state-of-the-art Computer Vision. In this paper we present a three module system based on both 2D and 3D cues. The first module uses 3D information to estimate the road plane parameters and thus select a coherent set of regions of interest (ROIs) to be further analyzed. The second module uses Real AdaBoost and a combined set of Haar wavelets and edge orientation histograms to classify the incoming ROIs as pedestrian or non-pedestrian. The final module loops again with the 3D cue in order to verify the classified ROIs and with the 2D in order to refine the final results. According to the results, the integration of the proposed techniques gives rise to a promising system.

1. Introduction

Nowadays, traffic accidents represent one of the major causes of death worldwide. According to the World Health Organization, everyday 3000 people die as a result of a road accident [1]. Concretely, in the vehicle-to-pedestrian accidents case the Economic Commission for Europe reported almost 150,000 injuries and 7000 killed pedestrians only in the European Union in 2003, representing the second source of fatalities just after vehicle-to-vehicle accidents [2]. However, contrary to the socially accepted view of traffic accidents as a random and unpredictable consequence of road transportation, these fatalities can be tackled by prevention and sensible measures. As a result, in the last decades such a problem is gaining more attention from both governments and industry, which invest big efforts in traffic safety research.

In last decade, in addition to the improvements in the road infrastructures (e.g., visibility enhancements, roundabouts, speed controls, better signposting, etc.), a new area of research has received a special focus: the Advanced Driver Assistance Systems (ADAS). ADAS are intelligent on-board systems that aim at anticipating and preventing accidents, or at least, minimizing their effects when unavoidable. Examples of ADAS are the Adaptive Cruise Control, which adjusts the own vehicle speed in order to keep a safe gap with the preceding vehicle, or the Lane Departure Warning, which warns the driver in case that the vehicle leaves the lane inadvertently. One of the most complex ADAS applications are the Pedestrian Protection Systems (PPSs), focus of this paper. In this case, the aim is to detect and localize static or moving people in a defined area in front of the vehicle in order both to provide information to the driver and to perform evasive or braking actions. Fig. 1 illustrates the typical risky areas to be tackled by a PPS. In regular conditions, the vehicle stopping distance is about 5 m at 30 km/h , increasing up to 12 m at 50 km/h, thus the systems must intelligently focus their techniques on the danger of detecting a pedestrian in these areas.

Computer Vision, by the use of passive sensors like cameras, plays a key role in most of these systems. For instance, cameras are used in PPSs in order to detect the traffic objects of interest (i.e., pedestrians) taking advantage of their rich amount of cues and high resolution. The topics involved in ADAS are in the frontier of the state-of-the-art since they require real-time interpretation of outdoor scenarios (uncontrolled illumination) from a mobile platform (fast background changes and presence of objects of unknown movement). Furthermore, in the PPSs context, pedestrian detection is even more challenging due to the high variability of their appearance (i.e., different articulated pose, clothes, distance and viewpoint) and the cluttered scenarios usually found in urban environments. It is worth to mention that the moving nature of ADAS makes some well-established techniques from other human detection areas, like background subtraction methods for surveillance, not applicable in our case.

In this paper we present a pedestrian detection system that makes use of Computer Vision cues, specially taking advantage of 3D information to enrich the classification, which is typically based on 2D. The system is divided in three steps. First, 3D data

* Corresponding author. Fax: +34 935811670.
E-mail addresses: dgeronimo@cvc.uab.es (D. Gerónimo), asappa@cvc.uab.es (A.D. Sappa), daniel@cvc.uab.es (D. Ponsa), antonio@cvc.uab.es (A.M. López).
URL: http://www.cvc.uab.es/adas (D. Gerónimo).

1077-3142/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2009.07.008

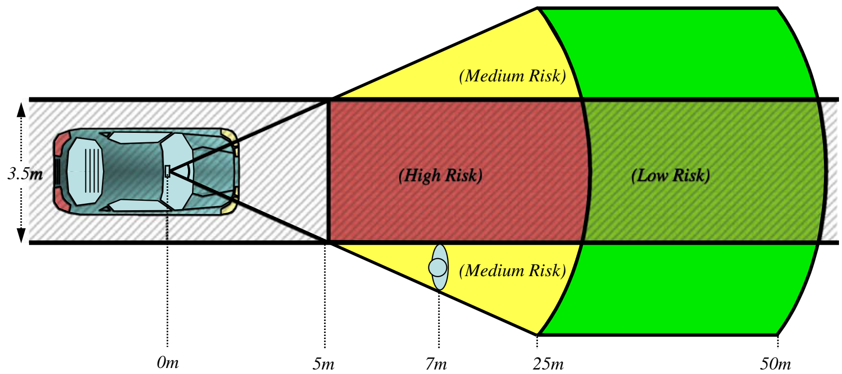

Fig. 1. The different areas of risk when driving. High risk area, in red, corresponds to a big danger of collision with pedestrians, always depending on the speed of the vehicle. Pedestrians in the medium risk area, in yellow, are likely to cross the front road, so typically no imminent is expected but the system must be aware of them. The low risk area, in green, contains pedestrians with no danger of imminent collision but that must be detected in advance since they stand in the vehicle's path. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

computed from a stereo rig are used to estimate the road pose, which is needed to adjust pedestrian sized windows in 3D. These windows, regions of interest (ROIs from now on), are then projected onto the 2D image plane where they are labeled as pedestrians or non-pedestrians by our proposed classifier: Real AdaBoost learning algorithm with Haar wavelets (HW) and edge orientation histogram (EOH) features. The final stage of the system verifies each positive labeled ROI by checking its 3D position and size. A final refinement stage is used to group overlapped redundant detections in 2D.

The remainder of this paper is as follows. After overviewing the related research in Section 2, an introduction to the proposed system is described in Section 3, fitting it to a general PPS architecture presented in [3]. Then, the modules of the current system, which make use of the aforementioned techniques, are placed in this architecture context. The first module, described in Section 4, makes use of the 3D-based adaptive image sampling technique. Section 5 presents the 2D classification module. Section 6 presents the last module, consisting of the 3D verification and the final 2D detections grouping. Finally, Section 7 presents experimental results of each of the three modules and of the whole system. Conclusions are summarized in Section 8.

2. Related research

By having a look at the literature [3] it is seen that most of the systems are based on feature selection and machine learning to perform 2D pedestrian classification. Some examples are the symmetry and binary template based approach by Broggi et al. [4], SVM on gradient images approach by Grubb et al. [5], the hierarchical template matching (*Chamfer System*) and neural networks by Gavrila et al. [6] or the parts-based SVM and AdaBoost approach by Shashua et al. [7]. In fact, PPSs can take advantage from a growing number of general people detection approaches proposed in recent years. For instance, Dalal and Triggs [8] propose *histograms of oriented gradients* (HOG) features and SVM. In [9], Leibe et al. perform the detection in two steps. First, image patches are extracted around difference of Gaussians keypoints [10]. Then, these patches are matched to a pedestrian model, which provides their spatial distribution, later used to cast votes to an hypotheses map. Finally, these hypotheses are verified and refined using template matching inspired in [6]. Tuzel et al. [11] base their classifier on the covariance of different measures (position, first and second order derivatives, gradient module, gradient orientation) in subwindows as features and boosting using Riemannian manifolds. Wu et al. [12] propose a parts-based scheme consisting of four body parts and three view categories to train a boosting-like classifier. They use short edge segments as features. Felzenszwalb et al. [13] use HOG and SVM in a parts-based approach, too. In this case, six different dynamic parts (not constrained to a fixed position in the hypothesis) are used.

Given that these methods are based on processing 2D images, a simple way of applying them is to classify windows of all the possible positions and sizes in the incoming image, which is often referred to as exhaustive window scanning (Fig. 3a). However, although widely used in general human detection approaches [14,8], this procedure not only is too expensive in terms of computational time (millions of windows should be classified) but also potentially increases the number of false positives by providing not relevant ROIs (e.g., sky areas). As a result, prior knowledge of the scene is generally considered to reduce this large amount of windows. For instance, since the system looks for pedestrians, only windows on the road surface should be taken into account for classification. Hence, an intuitive technique often used in ADAS literature is to fix an image row corresponding to the horizon and then assume that all pixels below this row belong to the road surface. As a result, a window laying on each pixel can be generated according to some mean pedestrian size constraints and the geometry of image formation. This approach, used by Gavrila et al. [6], has an implicit assumption: the relative position and orientation between the camera and the road do not change, i.e., the horizon line row is defined for the first frame and kept constant through the whole video sequence. They refer to this constraint as *flat world assumption*. However, due to vehicle movement, road slope and even road surface irregularities, there are many cases where such assumption is not fulfilled, specially in urban scenarios. Therefore, in order to compensate camera changes, many possible different windows per pixel should be considered, which would translate again in a very high processing time and potential false positives.

Some strategies to avoid the *flat world assumption* have been proposed. For instance, Soga et al. [15] propose a dense stereo based candidate window selection step that avoids an exhaustive searching of the whole image. Candidate windows are defined in those regions that contain solid objects (i.e., vertical surfaces) with a height in between 70 cm and 250 cm. Broggi et al. [16] propose first to identify vertical objects using a kind of v-disparity image [17] obtained from a stereo head. Then, further classification stages are focused only on those vertical objects.

Some systems propose a further step to reinforce detections. Gavrila et al. [6] make use of a disparity consistency test from a calibrated stereo rig to validate silhouette-based hypotheses. Ess et al. [18] propose a multi-frame scheme that jointly estimates scene geometry and verifies hypotheses by using a graphical model. In-