TRAFFIC SIGN DETECTION AND TRACKING USING ROBUST 3D ANALYSIS

Javier Marinas, Luis Salgado, Jon Arróspide and Massimo Camplani

Abstract

In this paper we present an innovative technique to tackle the problem of automatic road sign detection and tracking using an on-board stereo camera. It involves a continuous 3D analysis of the road sign during the whole tracking process. Firstly, a color and appearance based model is applied to generate road sign candidates in both stereo images. A sparse disparity map between the left and right images is then created for each candidate by using contour-based and SURF-based matching in the far and short range, respectively. Once the map has been computed, the correspondences are back-projected to generate a cloud of 3D points, and the best-fit plane is computed through RANSAC, ensuring robustness to outliers. Temporal consistency is enforced by means of a Kalman filter, which exploits the intrinsic smoothness of the 3D camera motion in traffic environments. Additionally, the estimation of the plane allows to correct deformations due to perspective, thus easing further sign classification.

1. Introduction

Within the field of Advanced Driver Assistance Systems (ADAS), automatic detection and recognition of road signs is one of the most appealing areas. In particular, traffic sign detection is critical as it paves the way for further sign classification, whose performance depends strongly in the quality of the detections. This is indeed very challenging due to the dynamic nature of the environment, which results in continuous pose, appearance and illumination changes.

The methods reported in the literature in the field of video-based ADAS can be broadly classified into monocular and stereo approaches. Monocular approaches have been extensively proposed, showing good results [1][2], despite the fact that they need a considerable amount of assumptions based on a priori information about the scene geometry. In highly dynamic environments, if these assumptions are not adequately met, the performance decreases significantly.

In turn, stereo systems provide 3D information about the scene analyzed, as well as the position and size of detected objects (e.g., vehicles [3] or pedestrians [4]). As a consequence, these systems do not require a priori geometry assumptions, outdoing monocular systems in terms of detection performance. Particularly, in the field of traffic sign detection, 3D analysis has been proposed in several works, although very few works have been proposed to implement pure 3D-based solutions. For instance, in [5] it is utilized to confirm detected traffic signs, and in [6] to help classification using GPS combined with the analysis of the road sign reflectance evolution. However, these approaches suffer from sensitivity to outliers in the 3D reconstruction, and do not take full advantage of the estimated 3D information. For instance, in [1] a 2D Kalman filter is implemented considering several assumptions about the scene geometry, which could be simplified through an adequate analysis of the 3D information.

In this paper a strategy for on-board road sign detection and tracking is proposed using stereovision. The method is based on the combination of contour-based and SURF feature-based matching, which enables us to attain continuous 3D reconstruction of the signs, as opposed to traditional single-cue based approaches. 3D reconstruction of the plane containing the sign is achieved through RANSAC algorithm, which removes potential outliers, thus enhancing the reliability of the method. In addition, the knowledge of this plane also allows us to rectify the deformations caused by the perspective effect and the camera projection parameters in the traffic sign images. Finally, in contrast to traditional 2D tracking approaches, a tracking framework is proposed that naturally exploits the smoothness in the 3D evolution of the sign with respect to the camera, thus providing more reliable and accurate results.

2. System overview

The block diagram of the proposed system is depicted in Fig. 1. As can be observed, prior to the 3D analysis, the process is carried out at single camera level. Firstly, the left image, considered as the reference image in this work, is analyzed. After a pre-processing stage (that involves noise filtering, image resizing and RGB to HSV conversion), a Bayesian classifier is used to segment the left image according to the color information, taking into account several illumination conditions. This pixel-wise classification is then used to perform an 8-connected component analysis, in order to determine a set of potential candidates to be a traffic sign (TS) according to an appearance model. At this point we have a set of potential candidates precisely located in the reference image and characterized by their bounding boxes (BB). Taking into account the position of this set of BB's and the known relationship between the projections in both cameras, we can define a new set of coarse BB's in the right image where the corresponding candidates are supposed to be located.

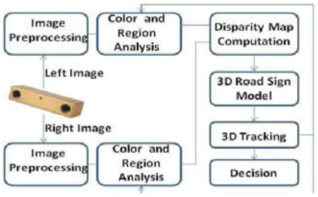

Fig. 1. Block diagram of the proposed TS detection and tracking system.

Once we have the hypothesized position of the candidates in both images, we compute a sparse disparity map for each candidate according to two different techniques (contour-based and SURF-based). Most importantly, the proposed method is able to adapt to the available object resolution by appropriately selecting either contour-based or SURF feature-based matching strategies according to the distance. Therefore a 3D reconstruction of points of the candidate is available at each time step.

Additionally, all the points belonging to a TS should be approximately located in a plane. Thus, robust plane estimation is performed to find the plane that best fits the cloud of 3D points. In particular, the RANSAC algorithm is used, which is robust to outliers due to bad correspondences. Once the 3D information of the sign computed, 3D tracking is performed by means of a Kalman filter. Additionally the relative position between the plane containing the sign and the camera will be used to rectify the view of the object (which may be affected by rotations) in order to provide the subsequent stages with a frontal view of the object, which can ease classification task.

3. Color and region analysis.

This section describes briefly the color and region analysis. A more detailed explanation can be found in [7]. First, the color segmentation stage aims to separate the TSs from the background using the color information of the objects according to the Hue (H) and Saturation (S) components of HSV color space. To perform this task, we use a Bayes classifier working with three classes (Red, Blue and Background). The likelihood functions are modeled using mixtures of Gaussians for both H and S components of red and blue classes. Regarding the Background class, a mixture between a uniform distribution and two Gaussian distributions modeling bricks and sky is used. Indeed, these elements have similar color as TSs, and thus need to be explicitly considered in order to avoid false detections.

After color analysis, the probability of the candidates to be TSs is further assessed through region-level modeling. First, color-segmented images are binarized, and 8-connected components analysis is applied to them for candidate region labeling. TS characterization at region level involves modeling of the following parameters: TS area, pictogram area and aspect ratio. All the three parameters have an acceptable range of values according to the known nature of the traffic signs. If a given object fulfills all the region-level requirements, this object is accepted as a potential TS (final decision is conditioned to temporal coherence analysis), and is characterized by its bounding box.

4. 3D road sign model.

The availability of corresponding regions in both images and the fact that road signs are planar surfaces help us to efficiently exploit 3D information not only to improve the detection accuracy but to render perspective corrected road signs.

4.1 Sparse disparity map computation

Dense disparity maps are typically required to achieve highly accurate perspective corrections. However, as road signs can be accurately approximated by planar surfaces and to estimate them only a set of reliable point correspondences is required, here the computation of sparse disparity maps is proposed. Besides, as their computation and posterior analysis is fast, they are very appealing when real-time constraint is to be enforced. Sparse disparity maps are computed in this work for each stereo candidate by means of two different techniques.

On the one hand, when a road sign is detected in the far distance a contour-based matching strategy is taken. In this case, considering the external contour of each pair of candidates, correspondences are quickly established taking into account the epipolar and ordering constraint: correspondent points of the contours in both images are supposed to be located in the same horizontal line (cameras are aligned). This method is fast and allows us to generate a disparity map with several points.

On the other hand, when the distance between the car and the object is reduced, we apply a robust image descriptor, SURF. Naturally, the closer the vehicle is to the traffic sign, the larger the resolution of the TS image, and hence the more feature points SURF is able to find. By applying this descriptor a set of correspondences between both ROIs is obtained, which constitute the sparse disparity map. We will only consider points belonging strictly to the road sign. Once the disparity map has been calculated, a cloud of 3D points is estimated by means of a stereo triangulation technique.

4.2 Plane estimation

Road signs are planar surfaces, so that the previous cloud of 3D points should fit to a plane in the case they do belong to road signs. RANSAC algorithm is applied in order to get an estimate of the plane that best fits the set of points calculated in each frame. As a result of applying RANSAC estimation, we will be able to determine the equation of the plane, π_w, that contains the road sign. The use of RANSAC prevents the estimation from being misled by the presence of outliers.

4.3 Perspective correction

Normally, road signs are supposed to be located perpendicularly to the direction of the vehicle. However, it is often the case that signs are rotated with respect to their expected plane either due to rotations in the camera (for instance, when the vehicle is turning) or to misalignments in the installation of the signs. In these cases, the estimated plane will not be perpendicular to the car movement. In fact, even in the case that the sign is indeed perpendicular, the projection of the TS into the image depends on the intrinsic camera projection parameters.

In order to undo the effect of perspective and camera projection, we can project the 3D road sign information in an affine plane, π_a, to get a frontal view of the sign, as can be seen in Fig. 2. This involves a change of coordinates: the new orthonormal base is found by applying Gram-Schmidt and imposing that two of its vectors, $\tilde{u}_1$ and $\tilde{u}_2$, be in π_w. These two vectors are the basis of the affine plane π_a. The traffic sign transformed to this plane is undistorted, as shown in Fig. 2. This improved view would be passed to a following classification stage.

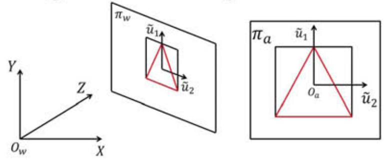

Fig. 2. Example of affine plane computation, π_a, given an estimated 3D plane, π_w.

5. 3D tracking.

Tracking is performed by means of a Kalman filter that gives coherence to the evolution of each candidate along time, ensuring the algorithm robustness. Remarkably, instead of tracking the objects in the image domain as done in traditional approaches, in the proposed method the Kalman filter operates always in the 3D space. Namely, the bounding box that contains the traffic sign in the image is back-projected to the plane π_w expected to hold the sign in the 3D space. Specifically, the Kalman filter tracks the 3D bounding box, characterized by one of its corners, its width and its aspect ratio. Formally, the Kalman filter equations are formulated as follows:

$$\boldsymbol{x}_k = \boldsymbol{A}\boldsymbol{x}_{k-1} + \boldsymbol{w}_{k-1} \qquad (1)$$

$$\boldsymbol{z}_k = \boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{v}_k \qquad (2)$$

$$\boldsymbol{x}_k = (x, y, z, \dot{x}, \dot{y}, \dot{z}, c, ar)^{\mathrm{T}} \qquad \boldsymbol{z}_k = (x, y, z, c, ar)^{\mathrm{T}}$$

$$\mathrm{A} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad \mathrm{H} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

where (1) and (2) represent the state and measurement equations. In these equations, $\boldsymbol{x}_k$ denotes the state vector and $\boldsymbol{z}_k$ the measurement vector, A is the state transition matrix, H is the matrix relating $\boldsymbol{x}_k$ and $\boldsymbol{z}_k$, and k is the time index. The random variables $\boldsymbol{w}_k$ and $\boldsymbol{v}_k$ represent the process and measurement noise. In $\boldsymbol{x}_k$, x, y and z are the coordinates of the upper left corner of the 3D bounding box, $\dot{x}$, $\dot{y}$ and $\dot{z}$ their respective velocities, c represents its width and ar its aspect ratio. These equations involve a first-order linear model for the TS position and a zero-order linear model for its width and aspect ratio.

Most importantly, in this case the linear process modeled by the Kalman filter corresponds to an inherently linear process in the 3D parameters, as opposed to a 2D tracking case, in which an approximation needs to be made to circumvent the non-linearity produced by perspective distortion. Therefore, the 3D tracking is bound to provide more reliable results.

Finally, if the filtered trajectory of the candidate (i.e. fulfilling color-based and region-level analysis) is smooth, as expected, the final decision module accepts it as a TS, otherwise it is discarded.

6. Results.

Our experimental data consisted of more than one hour of stereo video sequences. The stereo pair used is Point Grey Bumblebee II, with baseline 120 mm and resolution 480x360.

The implemented technique strongly relies on the result of the matching process. In Fig. 3 the estimation of the plane in two different moments can be observed: a) when the sign is far and a contour based matching is applied, and b) for a close sign SURF-based matching strategy is used. The green dot represents the position of the left camera, considered as the origin of the world.

The previous results disclose that, especially in contour-based matching technique, the presence of outliers is significant, mainly due to inaccurate contour detection -see right image in Fig. 3(a). The application of RANSAC allows us to be robust to outliers. In the case of SURF-based matching, RANSAC is also applied, even if the presence of outliers is smaller (as can be appreciated in Fig. 3 (b), in the right view)).

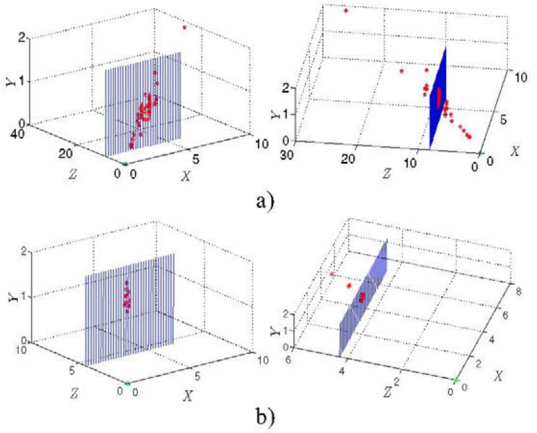

Fig. 3. Example of plane estimation considering a) the contour-based matching strategy and b) the SURF-based technique. In all images, numbers are expressed in meters.

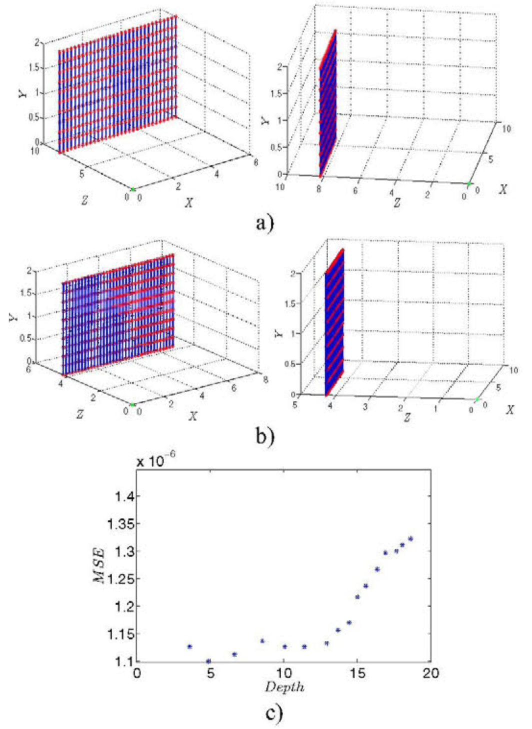

Fig. 4. Error committed in plane estimation: a) far distance, using contour-based technique; b) close distance, using SURF-based technique; MSE for a given sign during for each frame.

In order to prove the accuracy of our proposed technique to estimate the plane, we compare the results of both contour and SURF-based techniques with the manually extracted ground-truth plane. In Fig. 4 (a) the contour-based plane estimation result (in blue) and the ground truth plane (red) can be observed. In Fig. 4 (b) the plane (for a closer view of the same sign) estimated using SURF features is shown in blue. Finally, Fig. 4 (c) shows, for a given TS, the mean square error (MSE) between the projection of all TS points into the estimated and the reference planes as a function of distance. The MSE is very small (although naturally larger in the far distance, when using contour-based matching), which proves the accuracy of our approach.

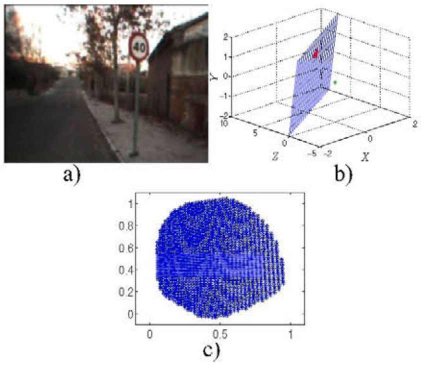

Fig. 5. Example of view correction of a rotated sign: a) original , b) 3D plane that contains the sign, c) rectified view of the distorted sign, which appears more circular than the original shape.

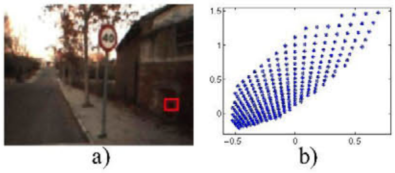

Fig. 6. Example of object rejection due to unexpected shape: a) the original image and b) the correspondent frontal view.

In addition, as explained in Section 4.3, the knowledge of the 3D plane containing the traffic sign allows for rectification of possible deformation resulting of camera or sign rotation. This is exemplified in Fig. 5. Observe that the traffic sign is imaged as an asymmetrical elongated shape in the original image, which complicates its recognition. In Fig. 5 (b) the detected features and the estimated plane in 3D are shown. The affine plane associated to this is shown in Fig 5 (c), in which the actual circular shape of the sign is roughly retrieved (rectification is less accurate in the right side, where the resolution of the original image is smaller).

The knowledge of the plane is also useful to discard false detections. In Fig. 6 we can see an example of false detection in a shady zone. If we compute the plane that contains the object and get the same frontal view as previously was explained, the shape obtained has no relationship with any expected road sign silhouette, therefore the object will not be further considered.

One of the main advantages of 3D analysis compared to 2D is that it allows for more accurate tracking of the traffic signs. The proposed algorithm has proved to provide excellent traffic sign detection and tracking within a 3D framework. Namely, in 3D the linear process modeled by the Kalman filter corresponds to an inherently linear process in the 3D parameters, in contrast to the 2D tracking case. The enhancement is exemplified in Fig. 7, where 2D and 3D tracking examples are shown some example signs in different illumination conditions. The bounding box is more accurately determined in the 3D tracking case, although the difference with 2D results is not dramatic. However, it is important to remark that the 3D analysis, as it has been proved, allow us to perform a richer representation of the TS (plane, orientation, actual position) compared to the traditional monocular systems.

7. Conclusions.

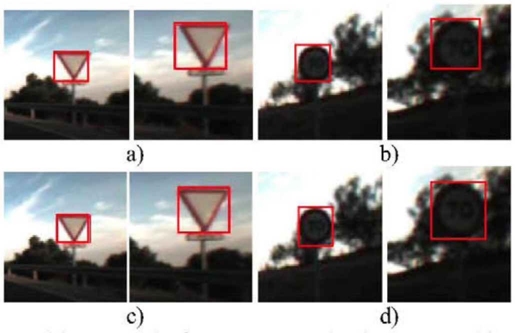

Fig. 7. Tracking example for some example sign: 2D tracking (good illumination conditions in a), shadow in b)). 3D tracking algorithm (good illumination conditions in c), shadow in d)).

The principal contribution of our approach is the continuous use of real 3D information about the objects of interest within a recursive Bayesian decision framework that allows to easily combine information of different nature, such as HS color at pixel level, and temporal and spatial coherence of image regions. In addition, Kalman-based tracking stage provides more accurate results due to the intrinsic linearity of 3D information. Finally, through plane estimation it is possible to get a corrected frontal view of the traffic signs, which can be used to enhance the performance of the posterior recognition stage. Future work is oriented to improve the efficiency of the algorithm to attain real-time operation conditions, in order to be ready to be used in real traffic scenarios, improving the security of road environments.

8. Acknowledgements.

This work has been partially supported by the Ministerio de Economía y Competitividad of the Spanish Government under project TEC2012-20412 (Enhanced 3DTV).

9. References.

[1] C.-Y. Fang, S.-W. Chen and C.-S. Fuh, "Road-sign detection and tracking," IEEE Trans. on Vehicular Technology, vol. 52, pp. 1329–1341, Sept. 2003.

[2] L. D. López and O. Fuentes, "Color-based road sign detection and Tracking," in Proc. International Conf. on Image Analysis and Recognition, LNCS 4633, vol. 4633/2007, pp. 1138–1147, 2007.

[3] M. Bertozzi, E. Binelli, A. Broggi, and M. D. Rose, "Stereo vision-based approaches for pedestrian detection," in Proc. IEEE Comput. Soc. Conf. CVPR Workshops, pp. 16-22, 2005.

[4] M. Bertozzi, A. Broggi, A. Fascioli and S. Nichele. "Stereo vision-based vehicle detection," in Procs. IEEE Intelligent Vehicles Symposium 2000, Detroit, USA, pp. 39–44, Oct. 2000.

[5] Y. Sheng, K. Zhang, C. Ye, C. Liang and J. Li. "Automatic detection and recognition of traffic signs in stereo images based on features and probabilistic neural networks", Proceedings of the SPIE, Volume 7000, pp. 70001I-70001I-12, 2008.

[6] S. D. McLoughlin, C. Deegan, C. Fitzgerald and C. Markham. "Classification of road sign type using mobile stereo vision", Proceedings of SPIE, Volume: 5823, Pages: 133-142, 2005.

[7] J. Marinas, L.Salgado, J.Arróspide and M. Nieto, "Detection and tracking of traffic signs using a recursive bayesian decision framework", IEEE Annual Conf. on Intelligent Transportation Systems, ITSC 2011, Washington (DC), USA, pp. 1942-1947, 5-7 Oct. 2011.