Obstacle Detection from Disparity Analysis using an A-contrario Approach

Moez Ammar¹, Sylvie Le Hégarat-Mascle¹, Marius Vasiliu¹, and Hugues Mounier²
¹Université Paris-Sud, Institut d'Electronique Fondamentale, 91405, Orsay
²Université Paris-Sud, Laboratoire des Signaux et Systèmes, 91192, Gif-sur-Yvette, France

Abstract - This study focuses on object detection problem considering data acquired by an on-board stereo pair of cameras. The disparity maps are analyzed in order to detect events salient relative to a disparity model. This model is minimal in the sense that it describes the absence of any object on the road. It is included in the information of the observed disparity image (and thus automatically deduced). Then, the a-contrario modeling allows detecting the regions 'amazingly' coherent with this minimal model. The used a-contrario model has two-levels, namely pixel and window so that the objects are detected as exceptional realizations of clusters of exceptional significantly high pixel values in the image representing the difference with the disparity model image. The 'exceptional' feature is quantified via a Number of False Alarms (NFA). We show that such an approach is successful both on simulated and actual data.

Keywords: A-contrario decision, video processing, change/object detection, embedded camera, ADAS.

1 Introduction

The increase in traffic in recent decades has prompted the emergence of new applications in embedded vision. In particular, there is a strong demand for the detection, localization, and recognition of objects in road scenes, i.e. from images of the road environment acquired by a vehicle in traffic. Then, in driver assistance systems, the image processing focuses on the automatic detection of the objects on the road or at the front of the vehicle, since it is the entry of emergency braking systems or avoidance ones.

Now, in acquired images, the 3D structure of the scene has been degraded by its projection onto the image plane of the camera. Then, stereovision methods have been developed for recovering the three-dimensional information from multiple views. The goal is to calculate the 3D spatial position of points from the coordinates of their projections in two different images. This can be done from the knowledge of the triplet (u,v,δ), where (u,v) is the image (line, column) coordinates of the pixel, and δ is its disparity. Practically, having rectified the stereo image pair, δ is the gap column coordinates of the same 3D point imaged in each of the right and left images [5]. δ is inversely proportional to the depth of the point in the scene. We refer the reader to [1] or [2] for a basic overview of stereovision, or [3] for advanced methods such as stereovision with N views and self-calibration techniques, or [4] for the neighbor approach so-called Structure From Motion approach, based on multiple images of a single moving camera. Now, among the stereovision algorithms proposed for obstacle detection, the so-called υ-disparity ([6], [7]) approach allows the extraction of the road profile and the objects located above this profile. Precisely, the υ-disparity consists in passing from a 3D representation of planes to two 2D ones that are less expensive to analyze; moreover, due to the cumulative basis provided by this representation, object extraction is robust even with noisy disparity maps.

The method proposed in this paper aims at detecting of obstacles without enumerating and specifying them (as pedestrians, bicycles, vehicles, or other objects ...) except as objects or events salient relative to a disparity model, also called background model. This model is minimal in the sense that it describes a simplified scene: either only containing static objects or containing only the road. The objects are then detected as ‘unpredictable’ events by comparison between this background model and the observation at a given instant, the ‘unpredictable’ or ‘amazing’ feature of an event or observation being measured in term of Number of False Alarms. The strength of a-contrario approaches ([8], [9], [10]) is its very simple foundation, namely the use of a ‘naïve model’, instead of requiring precise a-priori model of the data, that is all the more simple as these latter are manifold (as in our case). They have been applied to various detection problems: the detection of point alignments [8], change detection (in remote sensing images [11], or in video scenes [12]). In a previous work [13], we introduce an algorithm for detecting changes relative to a background image that is the previous image projected from time t to time t+1 knowing approximately the camera motion. This algorithm, based on an a-contrario approach, uses a double NFA criterion and is applied to grey level images. Here we generalize this approach showing that the background image can be defined from a model of the scene, either deduced from previous image(s) or extracted from the current image (as the road model here). We also extend the previous NFA measure defined between two images (the current one and the background one) to a measure that applies on an image time series.
The paper is organized as follows: Section 2 presents the two background models used. It recalls the way road model can be extracted from data image and the basis of the stereovision in an embedded system. Section 3 presents the multitemporal extension of [13] approach. Section 4 shows some results and compare them with obstacle detection from more classical stereovision algorithm. Finally Section 5 gathers our conclusions.

2 Context model

The extraction of image information is driven by the salient feature of this information. Saliency is measured relative to an a-priori model of the scene that specifies the so-called ‘background’ image. Two a-priori models can be considered leading to different features for the extracted objects:

- the previous image (or set of previous images) extrapolated to the current time if the camera is moving (as in our case) that boils down to focus on objects having a specific movement, and
- an image model in absence of any object on the road that boils down to focus on objects protruding from the road.

In both case, the disparity map knowledge is necessary. In this study, we assume a fully rectified image flow: no distortions, same resolution, isotropic focal, so both cameras are equivalent to the pinhole model with a focal $f$. We assume a fully epipolar correction: for both cameras the image plane is XOY, where $X$ is the horizontal axis (to the right) and $Y$ is the vertical axis (to the bottom). The stereo pair has parallel epipolar lines looking toward the positive $Z$ axis, the left one is placed at the origin $(0,0,0)$ and the right one is placed at $(b_z,0,0)$ where $b_z$ is the stereo baseline.

Many techniques exist for stereo matching, either local or global techniques (these latter ones offer a greater robustness to occlusions, sparsely textured or periodic patterns, but at the cost of increased computational complexity). In this study the stereo matching is based on the Sum of Absolute Differences (SAD) between local windows (typical size: 12x12 pixels), under the constraint of reciprocity (between left-to-right image matching, and right-to-left image matching) and of a matching cost lower than a given threshold. Let us note $\Delta$ the computed disparity image (from left camera point of view).

If objects with specific movement are researched, the background is deduced from a static scene assumption. In this case, the next disparity can be deduced from the previous one if we have some a-priori knowledge of the 6D camera movement between times $t$ and $t+1$, namely the translational component, $T = (T_x, T_y, T_z)$, and the rotational component, $\Omega = (\Omega_x, \Omega_y, \Omega_z)$, assumed small. For a 3D-point $M$, projected at current time $t$ in $m$ pixel at $(u_m,v_m)$ in the (left) image, we predict its next disparity $\delta_m'$ from current disparity $\delta_m$:

$$\delta_m' = \frac{\delta_m}{f} \frac{v_m}{\Omega_y} \frac{\Omega_x}{\Omega_z} \delta_m + \delta_{m+1}$$

(1)

If objects protruding from the road are researched, the background disparity image is the void scene one. In the absence of any other a-priori we construct it from the road estimation from $uv$-disparity technique briefly recalled here. The $v$-disparity technique [6] is based on the modeling of the road as a succession of portions of horizontal planes and oblique relative to the plane of the stereoscopic sensor, whereas an object on the carriageway (vehicle, pedestrian, tree) is characterized by a nearly vertical plane portion in the scene.

If objects protruding from the road are researched, the background is deduced, in the absence of any other a-priori, from the road plane. The $uv$-disparity technique provides a fast road plane estimation if we use a simple model where the road plane is parallel to $X$ axis, intersects the $Y$ axis at $h$, and has a pitch angle $\theta$ (small rotation around $X$ axis): $Y_m = h - Z_m - Z_x.\sin \theta$. In this case the road disparity must be:

$$\delta_m = \frac{h_y}{h_x} [(v_m - v_0) + f.\sin \theta] \text{ for } v_m \geq (v_0 - f.\sin \theta),$$

(2)

where $v_0$ is the vertical image coordinate of camera optical center.

This road disparity 'projects' in the $v$-disparity histogram $vA$ as a line segment $(\delta_m, v_m)$ between $(0, v_0 - f.\sin \theta)$ on first column and $(\frac{h_y}{h_x} [(v_m - v_0) + f.\sin \theta], v_{\max})$ on the bottom row $v_{\max}$. Thus, in real applications, a generalized Hough transform and a smart tracking algorithm allow us to evaluate the road model parameters $h$ and $\theta$. Then, we are able to compute the background disparity.

3 Proposed Approach

Our problem is as follows. We aim at detecting and localizing, e.g. through bounding box, the objects that are salient relative to the road and/or with an own movement. We assume these objects are large compared with the size of the image pixels, i.e. they include several pixels. Thus, in the absence of any other assumption about the shape of the researched objects, they will be detected as clusters of pixels satisfying a specific feature at pixel level. The considered features at pixel level define the kind of researched objects.

For instance, in change detection application, generally the considered pixel feature is a high squared difference between the pixel current (at time $t$) grey level value and the same pixel grey level value in the background image. Here, we consider that the main information to characterize objects at pixel level is the pixel disparity. Precisely, when researched objects are salient objects above the road, a pixel belonging to an object should correspond to a 3D point having a $Z$ coordinate much lower than the road point that would project at the same pixel (if not hidden). Since disparity is inversely proportional to $Z$ coordinate, the researched pixels should exhibit much higher disparity value than if they were road
pixels. When researched objects are characterized by an own movement, the main effect is that some pixel values change drastically due to the fact that the considered pixel no longer represent the same 3D point either due to occultation phenomenon (the object passes in front the 3D point previously imaged, i.e. projecting on the considered pixel) or due to uncovering (the object that moved does not mask anymore the 3D point now imaged). This occultation / uncovering phenomena induce change in disparity values (in addition to change in grey level values).

In summary, the researched objects are characterized in two ways: at the pixel level, by observing a value (grey level, disparity) ‘unexpected’ relative to a model of the scene projected on the image, and at image level, by the spatial organization of these pixels exhibiting ‘unexpected’ values.

In their work [9], A. Desolneux et al. propose several measures of the significance of some spatial organizations of points (aligned, clustered, etc...). These measures apply to binary images and thus they require previous labelling of the image pixels, in terms of ‘background’ pixels and pixels of interest. In [11] is proposed a measure of the significance of the grey level value of the pixels. This measure identifies the image subdomain (subset of pixels) whose values of grey level are the most amazing compared to a naive model that is assumed a priori and describes the absence of significant phenomenon. This step is thus similar to a detection step at pixel level that produces a pseudo-binary image (two or three labels). It should be applied to an image where the pixel values characterize the researched objects: typically an image of the squared differences of grey level values or of disparity values between the observed image and the background image (deduced from a scene model). Then, the researched objects are assumed to correspond to amazing clusters of points in the pseudo-binary image that can be detected using one of the measures [9].

Let us now describe more precisely the two steps of the detection.

In this study we consider two scene models. A strength of the proposed approach is that it will apply for both. Indeed, adopting an a-contrario approach, the detection is based on the definition of a ‘naive model’, instead of requiring precise a-priori model of the data, that is all the more simple as these latter are manifold (as in our case). The naive model can be only approximate since its part is less to model data than to be contradicted in the case (and only in the case) of structured data. The basic idea is then to detect the part of the data that is structured (e.g. point alignment, predictable pixels, etc.) as a very improbable realization of the naive model.

The first scene model is a virgin scene only including the road. As seen in previous section, the road can be estimated directly from y-disparity analysis, based on the assumption that it is piecewise planes, horizontal and oblique relative to the plane of the stereoscopic sensor. Then, assuming any 3D point of the scene either belongs to the road (i.e. satisfies the road equation) or lies at infinity, and knowing the camera parameters, the disparity image can be computed (cf. Eq.(2)). Let us note $\Delta^2$ the squared difference image, where in any pixel $s$ the value is $\Delta^2(s) = (\delta_t(s) - \delta_{\text{road}}(s))^2$, with $\delta_t(s)$ the disparity value in pixel $s$ according to the observed image at time $t$ ($\delta_t(s) = \delta_m$ in Eqs.(1-2)), and $\delta_{\text{road}}(s)$ the disparity value in pixel $s$ according to the road model.

The second scene model is the scene observed at previous time, $t-1$, and projected at time $t$ assuming the 6D movement of the camera and that the scene elements are static (cf. Eq.(1)). According to this model $\Delta^2(s) = (\delta_t(s) - \delta_{t-1-\text{rt}}(s))^2$ with $\delta_{t-1-\text{rt}}(s)$ the disparity value in pixel $s$ according to $(T, \Omega)$ and the disparity image at $t-1$. Note that missing values of disparity at $t-1$ generate missing values in the projected image at $t$. Hence conversely to previous model, here the background image is sparse (as the observed image).

In the following, $\delta$ is the generic notation to represent both model image, either $\delta_{\text{road}}$ or $\delta_{t-1-\text{rt}}$.

Let $\mathcal{D}$ be the $\delta$ domain of defined values (also called ‘known’ pixels), and $|\mathcal{D}|$ its cardinal (in our case, due to $\delta$ sparseness, $|\mathcal{D}|$ is lower to the image pixel number for $\delta = \delta_{t-1-\text{rt}}$). The significance of the difference image $\Delta$ values is measured choosing as naive model $\mathcal{H}_1$ a random field of $|\mathcal{D}|$ independent Gaussian centered variables with a given variance $\sigma$. The principle of a-contrario approach is to detect by contradicting $\mathcal{H}_1$. Hence, the model $\delta$ is as much correct as the do not follow $\mathcal{H}_1$ distribution.

The residual error between the observed image and $\delta$ can be measured over a sub-domain $D \subset \mathcal{D}$ by the squared Euclidean norm $\epsilon_D^2$ (that is the sum of $\Delta^2$ pixels over the considered image sub-domain $D$). The consistency over a spatial sub-domain $D$ is measured from the probability $\mathbb{P}_{\mathcal{H}_1}(\epsilon_D^2)$ to observe the residual error $\epsilon_D^2$:

$$NFA_1(D) = \eta_1(D) \cdot \mathbb{P}_{\mathcal{H}_1}(\epsilon_D^2) = \eta_1(D) \cdot \frac{1}{\Gamma(\frac{|\mathcal{D}|}{2})} \int_0^{\epsilon_D^2/2\sigma^2} e^{-t} \cdot t^{(|\mathcal{D}|/2-1)} dt,$$

(3)

where $\Gamma$ is the usual Euler function, and $\eta_1(D) = |\mathcal{D}| \cdot \left(\frac{|\mathcal{D}|}{|\mathcal{D}_0|}\right)$ is a normalization term to control the expected number of false alarms. Here, we choose $\eta_1(D) = |\mathcal{D}|$ in order to distribute the risk uniformly with respect to the domain size.

At the end of this first test of significance, we have detected the image sub-domain $\tilde{D}$ the most significant in term of $\Delta^2$ low values. By combining $\tilde{D}$ information with $\mathcal{D}$, we got a 3-label image showing the partition between ‘object’ pixels that belong to $\mathcal{D}$, ‘background’ pixels that belong to $\tilde{D}$, and ‘unknown-label’ pixels ($\epsilon \mathcal{D}$).
Then, the chosen naïve model \( \mathcal{H}_c \) for measuring the significance of a cluster of ‘object’ pixels is such that an object in the 3-label image is detected when the number of ‘object’ pixels within the object does not follow \( \mathcal{H}_c \) distribution. Classically, \( \mathcal{H}_c \) is the binomial distribution of parameter \( p \):

\[
NFA_2(W_j) = \eta_2 \cdot P_{\mathcal{H}_c}(p, \kappa, \nu) = \eta_2 \sum_{i=0}^{\nu} \binom{\nu}{i} p^i (1 - p)^{\nu - i}.
\]

Classically (e.g. [9], [13]), the parameters \((p, \kappa, \nu)\) are computed considering the image domain, and the tested image subdomain. For instance following [9] approach, \( p \) would be the relative area of the window \( W_j \) tested (i.e. \( \frac{|W_j|}{|D|} \)), \( \nu \) the total number of ‘object’ pixels (i.e. \( |D \setminus \bar{D}| \)), and \( \kappa \) the number of ‘object’ pixels falling in the window (i.e. \( |W_j \cap D \setminus \bar{D}| \)). In [13], \( p \) is equal to the probability for a pixel to be ‘object’ labeled within the whole image: \( p = \frac{|D \setminus \bar{D}|}{|D|} \) and for window \( W_j \) tested, \( \kappa \) is the number of ‘object’ pixels within \( W_j \); \( \nu = |W_j \cap D| \), and \( \nu \) is the number of ‘known-label’ pixels within \( W_j \).

Here we propose a multitemporal extension of these measures. We call ‘monotemporal’ the case where a single pair of instants is considered, namely \((t-1, t)\), and ‘multitemporal’ the case where several pairs of instants are considered, e.g. \((t-1, t), (t-2, t-1), (t-3, t-2)\). Pixel-level detection according to \( NFA_1 \) is applied to each time couple and the result is projected in the geometry of the image acquired at instant \( t \) (so that the different detection results are superimposable). Then, rather than 2D windows, 3D windows are considered, i.e. the parameters \((p, \kappa, \nu)\) are computed considering 3D domains defined by the image two dimensions plus the time dimension: if \( N \) is the number of time couples considered, and \( t \) is a subscript is added to previous notations to refer to the time couple:

\[
p = \frac{\sum_{t=1}^{N} |D_t \setminus \bar{D}_t|}{\sum_{t=1}^{N} |D_t|}, \quad \kappa = \sum_{t=1}^{N} |W_{jt} \cap D_t \setminus \bar{D}_t|\quad \text{and} \quad \nu = \sum_{t=1}^{N} |W_{jt} \cap D_t|.
\]

Then, we define the \( W_{jt} \) consistency measure by Eqs.(4-5) where \( \eta_2 \) is a normalization term chosen here equal to \( \left\lfloor |W_j| \right\rfloor \cdot 2^{\left| W_j \right|} \), where \( \left| W_j \right| \) is the total number of considered windows.

Finally, the detected windows are those having a \( NFA_2 \) value greater than a minimum value, \( NFA_2^{\text{max}} \). Note that the set of windows may include overlapping windows, in particular when considering windows of different sizes, e.g. different spatial dimension or different number of time couples. Thanks to the expectation property of the NFA measure (the NFA corresponds to an expectation that can be interpreted in itself, conversely to a probability measure), we are able to compare the levels of consistency of each window (even having different sizes). Thus, when two (or more) significant windows overlap, we can select the more significant.

4 Results

In order to measure the performance of our approach, the previous algorithms have been tested first on simulated data and second on actual data. The images are typical from those acquired from a camera on-board a driving car. For simulated images, we use the CIVIC simulator initially developed by the LIVIC/IFFSTAR laboratory. The assumed cameras are frontal and acquires grey level visible images (i.e. in the range 0.4±0.75 μm). The actual data have been acquired in the framework of the LoVe (Localization of Vulnerables) project that aims at contributing to road safety and gathers several French laboratories and industrial partners.

The simulated scene is as follows. The camera is aboard a vehicle driving so that the displacement between two successive image acquisitions is roughly translational, about 0.3 m. In this scene, the road is bordered with a few homes on both sides. It is located near a crossroads with pedestrian crossings. There are four motionless pedestrians and two other cars than the car embedding the camera. The first car drives in the same lane as the camera and the second car is coming facing in the opposite lane. Both cars have approximately the same speed as the camera.

Figure 1: Comparison between ‘monotemporal’ (one time couple) and multitemporal detection. Detection is performed according to \( NFA_2 \) criterion using pixel-level detection based on \( NFA_1 \) criterion. To facilitate the qualitative performance evaluation, the results are superimposed on the image gray level even if the detection only exploits the disparity images.

First, we check the interest of the multitemporal approach. Figure 1 shows respectively the results obtained using one couple of data, namely \((t_0,t_1)\) or using three time couples, namely \((t_0,t_1), (t_1,t_2)\), and \((t_2,t_3)\). In this case, the detection is relative to the second scene model, i.e. the scene observed at previous time, \( t-1 \), and projected at time \( t \) assuming the 6D movement of the camera. We note that using only one couple leads to several false positives, in particular in the sky, which can be removed using the time consistency constraint. Concerning the false negatives, first remind that here only the two cars that have an own motion should be detected. Now the disparity values vary as \( fb/\text{Z}_d \), i.e. for an object, the
variations of values of the disparity of the pixels representing the inside the object are less than 2-3 pixels (maximum values reached in the case of a car nearby, e.g. 10 m distance, moving in the opposite direction to the camera at a speed below 50 km/h with a camera at 30 frames/s and a baseline of 0.5 m). Now, the disparity value range is about 250 pixels with a noise equal or greater than the computed range of disparity change inside objects (2-3 pixels). Thus, the changes in disparity values that are detected are the changes due to occultation or uncovering phenomena, and they mainly occur at the borders of the objects. This explains why only the border of the car coming facing is detected. It also explains why the car in front of the car embedding the camera is not detected: the size of the uncovering areas between two successive images is too small.

Figure 2: Comparison between the two kinds of objects detected according the used background model: either virgin scene only including the road (first row), or scene observed at previous time (second row). The first column presents the pixel level detection according to NFA1 criterion (grey: unknown label pixel, black: background pixel, other colors: object pixel detected between 1 up to 3 times). The second column presents the windows level detection according to NFA2, based on pixel level detections.

We now aim at comparing the results of the two models leading to the ‘background’ image (either virgin scene only including the road, or scene observed at previous time). Figure 2 shows the results obtained using the first model (first row) and the second model (second row), respectively. In this example, the car movement is only approximately known. The first column shows results obtained after NFA1 (at pixel level). The shown images are color compositions where a channel represents a time couple detection. Grey values correspond to pixels where the disparity value is absent and therefore no information is available for detection. Colored pixels (i.e. excepting black or grey pixels) correspond to object pixels inconsistent with the ‘background’ model, detected at least in one time couple (in the three time couples for the white pixels). All the images have been projected in the geometry of the last considered image. In the first column (in particular with the second model) we see the successive disoccultations due to the facing car move that are represented by different colors. We note especially the difference between the two models that allow detecting either the objects protruding from the road plane or the moving objects (restricted to, as explained earlier, the regions where the movement induces an occultation or uncovering of the background).

Now, since the two models are complementary, it seems natural to combine them to refine the information provided by each of them. Here, the second model has been used to label the objects having a proper motion among the objects protruding from the plane of the road. Figure 3-left presents the result of such a labeling: the car coming facing is detected using both background models and thus recognized as a moving object (moreover, we note that it is slightly better represented in terms of object windows due to the complementary of the two models).

Finally, for comparison, Figure 3-right presents the results obtained using the \( u  \)-disparity approach. Precisely, the used algorithm has two main steps. The first one deals with the road plane estimation, in the \( v \)-disparity image, \( v \Delta \), as explained at the end of the Section ‘Context model’. Just note that we use two additional filters to make the detection more robust: a first one on \( v \Delta \) values before Hough transform, and a second one corresponding to a tracking of the road (mainly in order to ensure the temporal consistency of the estimates).

The second step focuses on the salient objects detection and bounding. It is based on \( u \)-disparity information. The \( u \)-disparity image \( u \Delta \) is defined similarly to \( v \Delta \) except for the direction of \( \Delta \) histograms: for every column \( u \) of \( \Delta \), the disparity histogram of \( u \) is the \( u \) column of \( u \Delta \), i.e. in \( u \Delta \), the grey level of pixel \( s \), located at row \( j \) and column \( u \), equal to the number of pixels of column \( u \) of \( \Delta \) having a disparity value equal to \( j \). Now for an object likened to a vertical plane parallel to the image plane (which is roughly the case of the researched objects: pedestrians, cars, road signs, etc.), the disparity is approximately constant, and it corresponds to a horizontal line segment in the image \( u \Delta \). Due to the approximate feature of the previous assumption, the researched objects rather correspond to connected components horizontally elongated in \( u \Delta \). (Note that once more the advantage of considering \( u \Delta \) image rather than the disparity image is the increase of the robustness of the detection due to the cumulative feature of histograms). Then, a hysteresis threshold provides the \( u \Delta \) connected components. For each of them likely to represent an object, the bounding box width and location (in columns in the original image) is equal to the column width and location of the connected component. The bottom row of the bounding box (in the original image) is deduced from the location in row of the connected component in \( u \Delta \) (remind that in \( u \Delta \) the row value is a disparity value) and the road plane equation previously obtained, which boils down to assume that the object is touching the road plane. Finally a filter on the object...
estimated size and position in Z coordinate allows removing some false alarms.

We now aim at comparing the proposed approach based on the disparity values with the same approach based on the image grey level values (rather than the disparity values) as proposed in [13]. The same algorithm, namely the succession of $NFA_1$ and $NFA_2$ criterion, is applied except that the background image is now the grey level image projected from time $t-1$ to time $t$ knowing the grey level image at $t-1$, the camera 6D movement between $t-1$ and $t$ and the disparity image at $t-1$. Thus, note that, in the case of the grey level based approach, the disparity images are also required. Figure 4 presents the obtained result using the same disparity images and camera 6D movement values as previously. We note that there are several false positives. Indeed the multitemporal filter can only cope with uncorrelated noise (mainly in the sky). Here, the remaining false positives are located at the borders of the road markings (see pixel-level detection on Figure 4-left). They are due to the fact that the 6D camera movement assumed is only approximate (and thus induces a ‘correlated’ noise). Indeed when camera movement is optimized (as proposed in [13], i.e. based on $NFA_1$ criterion), most of the false positives are removed, as shown by Figure 5-left. Anyway such an optimization is time consuming so that one supplementary interest (besides the double detection of the protruding objects relative to the road and that with an own motion) is the robustness of the results relative to the knowledge of the camera 6D movement.

![Image](image.png)

Figure 3: left: Results of the combination of the 2 background models: the houses and the motionless pedestrians are well identified as road plane protruding objects; the car facing the embedded camera is well identified as an own motion object; only the car driving with the same motion as the camera incorrectly recognized as a static object; right: results obtained using v-disparity approach. We note that the salient objects are well detected except the houses that are too far (last filter on object geometric features and location).

Finally, Figure 6 shows an example of result obtained in the case of actual data. It represents a crossroads with two cars, one crossing it and another standing on the crossroads. We note that the driving car is effectively identified as a moving object protruding the road plane. Among other objects protruding that are detected, there are the advertising board, a man on a motorcycle on the sidewalk, and less clearly, the car stopped in the crossroad. The buildings are too far to be well detected. We also note that the subparts of the objects too close to the road are not detected, and that there are a few false positives on the road due to errors in disparity values.

5 Conclusions

Being related to the frontal distance between camera and objects, the disparity is relevant to detect and separate the objects. However, it is very noisy (and not dense) which led to favor cumulative approaches. Here we investigate some a-contrario approaches whose strength is their genericity and their robustness to different models about the researched objects and to changing contexts. We show that, using the same algorithm, but considering different ‘background’ model (model of the road or previous image acquisition) the distinguished objects are different and combinable.

Short term perspectives deal with the evaluation of the performance on actual data in various conditions, and the algorithm coding optimization and the acceleration of the calculations, in order to reduce the computing time.

6 References


Figure 4: Results of the proposed algorithm applied to grey level values rather than disparity values.

Figure 5: Results of the proposed algorithm applied to grey level values (*left*) or disparity values (*right*) after camera 6D movement optimization according to NFA criterion [13].

Figure 6: Actual data example: *left*: Results of the combination of the 2 background models: the car crossing is well identified as an own motion object; the stopped car, the advertising board and a motionless man are well identified as road plane protruding objects; there are a few false positives; *right*: results obtained using v-disparity approach: the detection of the protruding objects is very satisfying and robust but there is no distinction between motionless and own motion objects.