2010 IEEE Intelligent Vehicles Symposium
University of California, San Diego, CA, USA
June 21-24, 2010

Detection of Independently Moving Objects Through Stereo Vision and Ego-Motion Extraction

Adrien Bak, Samia Bouchafa
UniverSud, Universite Paris XI
Institut d'Electronique Fondamentale
Orsay, France
Email : firstname.name@ief.u-psud.fr

Didier Aubert
UniverSud, LIVIC
INRETS, LCPC
Versailles, France
Email : firstname.name@inrets.fr

Abstract— **Vision-based autonomous vehicles must face numerous challenges in order to be effective in practical areas. Among these lies the detection and localization of independent-moving objects, so as to track or avoid them. In this paper a method that address this particular issue is presented. Information from stereo and motion is used to extract the ego-motion of the vehicle. Known defects of this estimation are exploited to detect independent-moving obstacles. This method allows an early and reliable detection, even for objects partially occluded. Besides, it highlights the errors in the disparity map, which can be used, in future works, to correct depth-estimation, through motion-estimation.**

I. Introduction

In order to develop an independent, mobile robot, one must first take a glance at the obstacle detection. The work described in this paper addresses such an issue. To that end, only visual information is to be used. Modern cars can be equipped with a variety of sensors (like GPS, proprioceptive sensors, collision detectors, *etc.*) but those sensors are.On the contrary, vision provides a much richer data and can serve several purposes, such as (but not limited to) localization, recognition and pathfinding. Anthropological and psycho-cognitive evidence [1] shows us the importance of visual information in human motivity and development. As such, computer vision, applied to intelligent vehicles is a highly active research topic, one can refer to [2] for a more extensive overview of the topic.

One can basically distinguish between monocular and binocular approaches. Monocular approaches, such as [3], [4] rely on image motion estimation, through the computation of optical flow [5]. Some authors, such as [6], estimate the motion directly from the image, but they still have to rely on the *brightness constraint equation*. However, the ill-posed nature of the optical flow computation problem makes the use of regularization or heavy smoothing constraints [7], [8] necessary. Such constraints can deteriorate the useful information in image region such as occlusion regions, or depth-discontinuities. Besides, monocular methods lack the exact knowledge of objects depth and can only determine the exact position of a given object up to a scale factor. On the other hand stereovision based methods provide, through calibration, an absolute measurement of a 3D space. Disparity information can be used in order to detect, without any other input, potential obstacles [9], [10].

The information provided by both cues is complementary, thus a current trend is to make those collaborate, in order to exploit motion analysis and scene structure. For instance, the past decade has seen many attempts to achieve a useful collaboration in the domain of obstacle detection [11], [12] or in the field of ego-motion recovery (odometry) and pathfinding [13]–[15].

Some authors, such as [16] have tried to estimate the ego-motion of a stereo-rig and then compute a 3D-displacement field due to this ego-motion, in order to identify dynamic objects. However their method differs from the one described here on several points. First, they use the predicted displacement field only to discriminate between static and dynamic objects, stereo-vision is then used to extract the different targets. This could lead to detection errors, for instance, two distinct objects with different motions but with the same disparity would be merged. Moreover, their method relies on a thresholding, whereas the proposed algorithm stems from a much more robust error analysis of the most important part of such an algorithm : the ego-motion extraction. Finally, the proposed method relies on robust feature points, which are insensitive to the aperture problem and allow for greater displacement than the correlation-based optical flow proposed in [16].

The method presented here stems from [13] work on d-motion estimation. According to the error model developed in section IV, the consistency of every point with the extracted ego-motion is checked through the use of a robust correlation technique. The proposed method does not only perform a static/dynamic detection, it allows a fine segmentation with respect to the obstacle's own ego-motion. As shown in section V, an early detection can be achieved even in hard cases (occlusion for instance).

In the first section, the motion model and the hypothesis will be described. In the second section, an ego-motion extraction algorithm will be described and tested. In section IV, a method to detect independent motion in a sequence of image pair is presented. The results of this method and their discussion is to be found in section V. Finally, section VI concludes and presents some improvement that can be brought to the current method.

978-1-4244-7868-2/10/$26.00 ©2010 IEEE

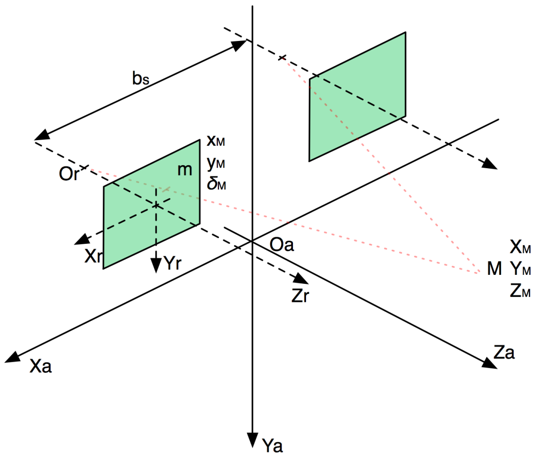

Fig. 1. Coordinate system to be used, at time $(t + \partial t)$

II. Model and Hypothesis

A mobile vehicle (*e.g.* a car), moving in a world constituted of either static or dynamic objects, is considered. The world is described by a set of two frames of reference. The first one, labelled $\mathcal{R}_a$ is absolute while the second one $\mathcal{R}_r$ is bound to the vehicle, with its origin located at the optical center of the right-hand-side vision sensor.

This vehicle is equipped with a rectified stereo rig, looking forward. The two image sensors are modeled by the pinhole camera model. Both focal lengths are identical and are noted f, the stereo rig's baseline is b_s, pixels are considered to be squared, with their dimension equal to t_p. Disparity is measured with respect to the right-hand-side coordinates. The choice of the algorithm used to recover sparse or dense disparity maps is left to the reader's discretion. One can refer to [17] for an extensive study of existing algorithms. The method that will be used is fully described in [10].

According to the standard pinhole model, a static world point $M = \left| \begin{array}{c} X_M(t) \\ Y_M(t) \\ Z_M(t) \end{array} \right|_{\mathcal{R}_a}$ is imaged by our system as :

$$m = \left| \begin{array}{c} x_M(t) = f\frac{X_M(t) - b_s/2}{Z_M(t)} \\ y_M(t) = f\frac{Y_M(t)}{Z_M(t)} \\ \delta_M(t) = f\frac{b_s}{Z_M(t)} \end{array} \right| \qquad (1)$$

Where δ is the disparity between right and left images, and m belongs to the disparity space. One can easily show that the disparity space is a projective space as there exists a projective transformation between the homogeneous coordinates of a point in the 3D euclidean space and the homogeneous coordinates of its image by the stereo rig. This work will be conducted in the disparity space, because of the isotropic nature of the associated discretization noise as shown in [13].

Static objects are assumed to be dominant in the images. This assumption will later allow to extract the vehicle ego-motion. Without any loss of generality and unless explicit notice, only two different times, labelled t and $t + \partial t$ are considered. At t, the two coordinates systems are coincident. The motion between $\mathcal{R}_a$ and $\mathcal{R}_r$, that occurs between t and $t + \partial t$ is decomposed in its translational and rotational components.

$$\overrightarrow{T}(t) = \left| \begin{array}{c} T_X(t) \\ T_Y(t) \\ T_Z(t) \end{array} \right|_{\mathcal{R}_a}$$

$$\overrightarrow{\Omega}(t) = \left| \begin{array}{c} \omega_X(t) \\ \omega_Y(t) \\ \omega_Z(t) \end{array} \right|_{\mathcal{R}_a}$$

The angles $\omega_X(t)$, $\omega_Y(t)$ and $\omega_Z(t)$ are considered to be small enough to linearize trigonometric lines. This assumption is valid in standard driving conditions. After the motion $\left\{\overrightarrow{T}, \overrightarrow{\Omega}\right\}$, the static world point M can be expressed as:

$$M = \left| \left| \begin{array}{c} X_M(t) \\ Y_M(t) \\ Z_M(t) \end{array} \right| + \overrightarrow{\Omega}(t) \wedge \left| \begin{array}{c} X_M(t) \\ Y_M(t) \\ Z_M(t) \end{array} \right| - \overrightarrow{T}(t) \right|_{\mathcal{R}_r}$$

$$M = \left| \begin{array}{c} X_M(t) - \omega_Y(t).Z_M(t) + \omega_Z(t).Y_M(t) - T_X(t) \\ Y_M(t) + \omega_X(t).Z_M(t) - \omega_Z(t).X_M(t) - T_Y(t) \\ Z_M(t) + \omega_Y(t).X_M(t) - \omega_X(t).Y_M(t) - T_Z(t) \end{array} \right|_{\mathcal{R}_r}$$

In order to model motion in the disparity space, the so-called d-motion formalism is used. Assuming that $m(t) \left| \begin{array}{c} x(t) \\ y(t) \\ \delta(t) \end{array} \right|$ is the image of a static world point, it will be mapped as $m(t + \partial t)$. As of now, $variable(t)$ will simply be noted $variable$, and $variable(t+\partial t)$ will be noted $variable'$. The coordinates of m' can be expressed through the projections of the motion components equations:

$$m' = \left| \begin{array}{c} x' \\ y' \\ \delta' \end{array} \right| = \left| \begin{array}{c} \frac{x + \omega_Z.y - \frac{T_X.\delta}{b_s}.f - f.\omega_Y}{\frac{\omega_Y}{f}.x - \frac{\omega_X}{f}.y - \frac{T_Z.\delta}{b_s} + 1} \\ \frac{y - \omega_Z.x - \frac{T_Y.\delta}{b_s}.f + f.\omega_X}{\frac{\omega_Y}{f}.x - \frac{\omega_X}{f}.y - \frac{T_Z}{f.b_s} + 1} \\ \frac{\delta}{\frac{\omega_Y}{f}.x - \frac{\omega_X}{f}.y - \frac{T_Z}{f.b_s} + 1} \end{array} \right| \qquad (2)$$

In the following the following will be used :

$$m' = P_{\left(\overrightarrow{\Omega}, \overrightarrow{T}\right)}(m)$$

III. Ego-Motion Extraction

A. Method

In order to identify moving obstacles, one must first evaluate its own ego-motion. For that we consider a set of N point correspondences $\{m_i, i \in [1, N]\} \rightarrow \{m'_i \in [1, N]\}$. Those correspondence can be provide by an optical flow method, but we'd rather rely on more robust, feature points extraction methods, such as Harris corner detector [18], the SURF detector [19], or level-lines junction [20]. Because of its trade-off between robustness and computational cost, a SURF detector will be used later on.

Equation (2) provides a linear system in $(\overrightarrow{\Omega}, \overrightarrow{T})$:

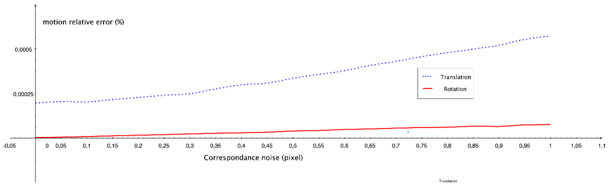

Fig. 2. Ego-motion estimation relative error Vs. correspondences noise. Used data consists of 500 points, 20% of which were outliers. The estimator used to evaluate relative error is the median of 10 000 realizations.

$$\begin{array}{c}\overrightarrow{\Omega}\\ \overrightarrow{T}\end{array} \cdot \begin{pmatrix} -\frac{x'_1 \cdot y_1}{f} & \frac{x_1 \cdot x'_1}{f} + f & -y & \frac{\delta_1}{b_s} & 0 & \frac{\delta_1 \cdot x'_1}{b_s} \\ f - \frac{y_1 \cdot y'_1}{f} & \frac{x_1 \cdot y'_1}{f} & -x & 0 & \frac{\delta_1}{b_s} & \frac{\delta_1 \cdot y'_1}{b_s} \\ \frac{\delta'_1 \cdot y_1}{f} & -\frac{\delta'_1 \cdot x_1}{f} & 0 & 0 & 0 & \frac{\delta_1 \cdot \delta'_1}{b_s} \\ \cdots & & & & & \\ -\frac{x'_N \cdot y_N}{f} & \frac{x_N \cdot x'_N}{f} + f & -y & \frac{\delta_N}{b_s} & 0 & \frac{\delta_N \cdot x'_N}{b_s} \\ f - \frac{y_N \cdot y'_N}{f} & \frac{x_N \cdot y'_N}{f} & -x & 0 & \frac{\delta_N}{b_s} & \frac{\delta_N \cdot y'_N}{b_s} \\ \frac{\delta'_N \cdot y_N}{f} & -\frac{\delta'_N \cdot x_N}{f} & 0 & 0 & 0 & \frac{\delta_N \cdot \delta'_N}{b_s} \end{pmatrix} = \begin{pmatrix} x'_1 - x_1 \\ y'_1 - y_1 \\ \delta'_1 - \delta_1 \\ \cdots \\ x'_N - x_N \\ y'_N - y_N \\ \delta'_N - \delta_N \end{pmatrix}$$

The purpose is to find a couple $\left(\overrightarrow{\tilde{\Omega}}, \overrightarrow{\tilde{T}}\right)$[1] that minimizes the following :

$$\epsilon = \sum_{i=1}^{i=N} dist\left(m'_i, P_{\left(\overrightarrow{\tilde{\Omega}}, \overrightarrow{\tilde{T}}\right)}(m_i)\right)$$

where $dist\,(a, b)$ can be any of the usual topological distances, extended to the disparity space. All experiments were conducted with the euclidean distance. That minimization is performed, using a RANSAC [21] approach in order to reject outliers, along with singular value decomposition to solve the linear system at every step of the process. RANSAC parameters are set assuming a minimal proportion of inliers of one third of the extracted feature points and to ensure a 5% probability of false rejections [22].

B. Results

In order to evaluate the precision of the ego-motion extraction, we proceed by different means, first synthetic data is used. Such data is constituted by a set of static points and animated by an arbitrary motion, with perfect correspondences (or with a perfectly known noise) feeding the ego-motion extraction algorithm. Those results can be found in Fig. 2.

The Sivic simulator [23] was also used. This presents the advantage of providing pseudo-realistic image-sequences and a perfect knowledge of the vehicle ego-motion. Fig. 3 shows an example of such pseudo-realistic images. Fig. 4 presents results for a test sequence from Sivic. This sequence presents urban landscape, with moderate traffic. After 600 frames, and about 250 meters, the positioning error is 2.5 meters. The average instantaneous error is less than 2%.

Errors in the ego-motion estimation stems from various sources:

- First, one can note the fact that RANSAC process isn't always optimal. For practical applications, we are bound to set a maximum number of iterations, and we do not have the certainty that the final result is the best we can obtain, depending on our input data.
- Second, there is the discretization of the disparity space, and more generally, the positioning of the feature points used to inverse the motion model.

From these two error sources, only the later can be quantified. However, through the use of the Sivic simulator, an upper bound to the relative error was estimated around 15%.

IV. Detection

A. Ego-Motion Error Propagation

The objective of this section is to detect image points that don't validate the motion model found in section III. For that, from (2), one can write :

$$\begin{cases} \mu = \frac{xy}{f}\omega_X - \left(f + \frac{x^2}{f}\right)\omega_Y + y\omega_Z - \delta f\frac{T_X}{b_s} + \frac{x\delta T_Z}{b_s} \\ \nu = \left(f + \frac{y^2}{f}\right)\omega_X - \frac{xy}{f}\omega_Y - x\omega_Z - \delta f\frac{T_Y}{b_s} + \frac{y\delta T_Z}{b_s} \\ \xi = \delta\frac{y\omega_X - x\omega_Y + \frac{T_Z}{b_s}}{x\omega_Y - y\omega_X - \frac{T_Z}{b_s} + 1} \end{cases} \tag{3}$$

Where $\mu = x' - x$, $\nu = y' - y$ and $\xi = \delta' - \delta$. The following notation will be used :

$$\left|\begin{array}{c} \mu \\ \nu \\ \xi \end{array}\right| = \Pi_{\left(\overrightarrow{\Omega}, \overrightarrow{T}\right)}\begin{pmatrix} x \\ y \\ \delta \end{pmatrix}$$

and $\Pi_{\left(\overrightarrow{\Omega}, \overrightarrow{T}\right)}$ can be called *displacement field*.

For every point m in the disparity-space at the time t, the coordinates of the point $m' = m + \Pi_{\left(\overrightarrow{\tilde{\Omega}}, \overrightarrow{\tilde{T}}\right)}(m)$ are computed. If the point m is the image of a static world point,

[1] the tilda symbol denotes an estimate

Fig. 3. Image extracted from a test sequence generated with the Sivic simulator

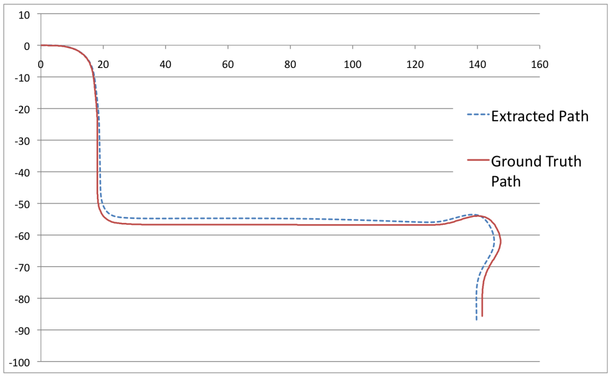

Fig. 4. Extracted and Ground Truth paths for a sequence of 600 stereo pairs.

m' should be its correspondent at the time $t+\partial t$, on the other hand, if m is the image of a world point, belonging to a dynamic object, m' shouldn't. However, as seen previously, the estimation of the ego-motion, isn't perfectly accurate. Through the propagation of uncertainty with respect to $\overrightarrow{\Omega}$ and $\overrightarrow{T}$, and with (3) :

$$\left\{ \begin{array}{c} \partial\mu = \frac{xy}{f}\partial\omega_X + \left(f+\frac{x^2}{f}\right)\partial\omega_Y + y\partial\omega_Z - \frac{f\delta}{b_s}\partial T_X + \frac{x\delta}{b_s}\partial T_Z \\ \partial\nu = \left(f+\frac{y^2}{f}\right)\partial\omega_X - \frac{xy}{f}\partial\omega_Y - x\partial\omega_Z - \frac{f\delta}{b_s}\partial T_Y + \frac{y\delta}{b_s}\partial T_Z \\ \partial\xi = \frac{\delta}{(1-a)^2}.\left(y\partial\omega_X - x\partial\omega_Y + \frac{1}{b_s}\partial T_Z\right) \end{array} \right. \tag{4}$$

where $a = y\omega_X - x\omega_Y + \frac{T_Z}{b_s}$

By replacing in (4) every motion-related variable by an estimate of its upper bound, the following can be defined:

$$\left\{ \begin{array}{c} \Delta\mu = \frac{xy}{f}\Delta\omega_X + \left(f+\frac{x^2}{f}\right)\Delta\omega_Y + y\Delta\omega_Z - \frac{f\delta}{b_s}\Delta T_X + \frac{x\delta}{b_s}\Delta T_Z = P_1\left(x,y,\delta\right) \\ \Delta\nu = \left(f+\frac{y^2}{f}\right)\Delta\omega_X - \frac{xy}{f}\Delta\omega_Y - x\Delta\omega_Z - \frac{f\delta}{b_s}\Delta T_Y + \frac{y\delta}{b_s}\Delta T_Z = P_2\left(x,y,\delta\right) \\ \Delta\xi = \frac{\delta}{(1-a)^2}.\left(y\Delta\omega_X - x\Delta\omega_Y + \frac{1}{b_s}\Delta T_Z\right) = P_3\left(x,y,\delta\right) \end{array} \right.$$

where P_1, P_2 and P_3 are polynomials. For every point, given its position in the disparity space and the current estimate of the ego-motion, the upper bounds $\Delta\mu, \Delta\nu$ and $\Delta\xi$ of the uncertainty of the displacement field can now be estimated.

With this result, for each point m of the disparity space, a 3D interval, at the time $t+\partial t$, can be defined by:

$$\begin{array}{rl} \mathcal{W}^m_{\left(\tilde{\overrightarrow{\Omega}},\tilde{\overrightarrow{T}}\right)} & = [x+\mu-\Delta\mu; x+\mu+\Delta\mu] \\ & \times [y+\nu-\Delta\nu; y+\nu+\Delta\nu] \\ & \times [\delta+\xi-\Delta\xi; \delta+\xi+\Delta\xi] \end{array}$$

where "$\times$" stands for the cartesian products.

So, for every point m in the disparity space that is the image of a static world point, it is known that :

$$\exists\ m' \in \mathcal{W}^m_{\left(\tilde{\overrightarrow{\Omega}},\tilde{\overrightarrow{T}}\right)} \text{ such as}$$
$$m' = m + \Pi_{\left(\overrightarrow{\Omega},\overrightarrow{T}\right)}(m)$$

By searching $\mathcal{W}^m_{\left(\tilde{\overrightarrow{\Omega}},\tilde{\overrightarrow{T}}\right)}$ for the correspondent of m, it can be determined whether m is a static point (*i.e.* there is actually a correspondent to m in $\mathcal{W}^m_{\left(\tilde{\overrightarrow{\Omega}},\tilde{\overrightarrow{T}}\right)}$), or if m moves independently (*i.e.* no correspondent can be found within the boundaries of $\mathcal{W}^m_{\left(\tilde{\overrightarrow{\Omega}},\tilde{\overrightarrow{T}}\right)}$). The only significant limitation at this point is the case of a dynamic point d, whose motion $\left(\overrightarrow{\Omega_d},\overrightarrow{T_d}\right)$ satisfies the following :

$$\begin{array}{c} \overrightarrow{\Omega_d} \in \left[\tilde{\overrightarrow{\Omega}} - \partial\overrightarrow{\Omega}\ ,\ \tilde{\overrightarrow{\Omega}} + \partial\overrightarrow{\Omega}\right] \\ \overrightarrow{T_d} \in \left[\tilde{\overrightarrow{T}} - \partial\overrightarrow{T}\ ,\ \tilde{\overrightarrow{T}} + \partial\overrightarrow{T}\right] \end{array}$$

such a point would have its correspondent within the boundaries of $\mathcal{W}^d_{\left(\tilde{\overrightarrow{\Omega}},\tilde{\overrightarrow{T}}\right)}$, because of the definition of $\mathcal{W}^d_{\left(\tilde{\overrightarrow{\Omega}},\tilde{\overrightarrow{T}}\right)}$. In other words, a static point and a dynamic one can not be discriminated if the motion of the later differs from the ego-motion of a quantity smaller than the extraction noise.

Yet, the presence or absence of a correspondent to m in $\mathcal{W}^m_{\left(\tilde{\overrightarrow{\Omega}},\tilde{\overrightarrow{T}}\right)}$, only allows to distinguish between static and dynamic points.

B. Independent Motion Detection

In order to refine a subsequent image segmentation, the 3D interval of the disparity space, at $t+\partial t$ is defined :

$$\begin{array}{rl} \mathcal{R}^m_{\left(\tilde{\overrightarrow{\Omega}},\tilde{\overrightarrow{T}}\right)} & = [x+\mu-\kappa\Delta\mu; x+\mu+\kappa\Delta\mu] \\ & \times [y+\nu-\kappa\Delta\nu; y+\nu+\kappa\Delta\nu] \\ & \times [\delta+\xi-\kappa\Delta\xi; \delta+\xi+\kappa\Delta\xi] \end{array}$$

The correspondent of m will be searched in $\mathcal{R}^m_{\left(\tilde{\overrightarrow{\Omega}},\tilde{\overrightarrow{T}}\right)}$. That way, one will be able to differentiate between static and dynamic objects, but also between two different dynamic objects.

This correspondence search is performed by calculating the Zero-mean Sum of Absolute Differences (ZSAD) over a small neighborhood, between m and $c\ \forall\ c \in \mathcal{R}^m_{\left(\overrightarrow{\Omega},\overrightarrow{T}\right)}$.

However if a basic correlation is well adapted to stereo pairing problems (due to the epipolar constraint), in a more general case, periodic (or continuous ones) objects can provide a single point with multiple good correspondence candidates. In order to avoid possible ambiguities, a standard *Winner Takes All* (WTA) approach is not used. Instead, every candidate c that has a ZSAD score satisfying the following conditions is considered:

$$\left\{ \begin{array}{c} ZSAD_c < Tr \\ \frac{ZSAD_c - min\left(ZSAD_{c'} | c' \in \mathcal{R}^m_{\left(\overrightarrow{\Omega},\overrightarrow{T}\right)}\right)}{ZSAD_c} < \alpha \end{array} \right.$$

Where Tr is a set threshold, usually set around $5 \times Nb$, where Nb is the number of pixels in the neighborhood used to calculate the ZSAD score, and α is an arbitrary tolerance threshold, usually set around 0.1 to 0.3.

The candidate m'_b that is the closest to $m + \Pi_{\left(\tilde{\vec{\Omega}}, \tilde{\vec{T}}\right)}(m)$ among the valid ones is retained. If there is no candidate satisfying the previously mentioned conditions, the point m is flagged with a special value as such a situation can mean several things:

- m lies within an occlusion region
- m is the image of a world point that presents a motion larger than the one imaged by $\mathcal{R}^m_{\left(\tilde{\vec{\Omega}}, \tilde{\vec{T}}\right)}$
- the disparity at point m is false.

Through this correlation process, every point m in the disparity space is associated with a vector:

$$\overrightarrow{A} = m'_b - m - \Pi_{\left(\tilde{\vec{\Omega}}, \tilde{\vec{T}}\right)}(m) = \begin{cases} \mu_A \\ \nu_A \\ \xi_A \end{cases}$$

Representations of this vector field can be found as Fig. 7 and Fig. 9 (see section V for a description of the used color encoding).

C. Mobile Objects Segmentation

Due to the semi-dense nature of the used disparity maps, one target will often be split into several blobs. A hierarchical clustering process [24] is used to group different parts of an object together. The distance used in order to complete such a process is based on motion and disparity. It can be expressed as :

$$\begin{aligned} Dist_{clustering}(a,b) = & \left|\arg\left(\overrightarrow{A}(a)\right) - \arg\left(\overrightarrow{A}(b)\right)\right| \\ & + W_1 \sqrt{\left\|\overrightarrow{A}(a)\right\|^2 - \left\|\overrightarrow{A}(b)\right\|^2} \\ & + W_2 \left|\delta(a) - \delta(b)\right| \\ & + W_3 Dist_{Image}(a,b) \end{aligned}$$

Where W_1 ,W_2 and W_3 are empirically determined relative weights between the different components of the distance, arg is the angle between the vector and the horizontal axis and $Dist_{Image}(a,b)$ is the Euclidean distance, in the image space.

Both these pieces of information are needed, to lift some uncertainties. For instance, if a rigid object is independently moving in a given direction, some part of this object can present a low or zero local contrast (*e.g.* some part of the body of a car), so those parts will appear with no independent motion. Disparity information must be used to associate those parts with a wider, independently moving, set.

V. Results & Discussion

For representation purposes, the projection of the $\overrightarrow{A}$ vector field upon the image space (*i.e.* without its disparity component, which is, anyway, the smallest one) will be displayed, using the color encoding illustrated in Fig. 5.

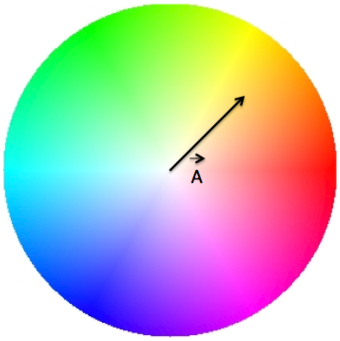

Fig. 5. Color representation used. Hue and Saturation are determined by the length and angle of the vector. Value is 0 if $m'_b \in \mathcal{W}^m_{(\Omega,T)}$, 0.6 otherwise.

Tests were conduced with 7 stereo sequences from the french LoVE (*Logiciel d'Observations des Vulnerables*) project. The stereo sensor was composed by two identical 640x480 CCD-cameras, equipped with 6 mm lenses, the baseline of this sensor was 58 cm. All those sequences were shot in urban areas and present low to medium traffic.

Fig. 8 shows the results of the presented algorithm for the image sequence illustrated in Fig. 6 & 7. The two targets are well defined. Even if the pedestrian on the right hand side is partially occluded, she is well detected, as every visible part of her body is labelled with the same tone. On the other hand, some body parts of the moving car are identified as static. This is due to its constant and untextured nature, and a way to circumvent this has already been exposed.

It is important to note the road post on the right side of the road. One can see in Fig. 7, that its upper part was attributed a wrong disparity value[2]. This disparity error yields to a motion prediction error. One can also notice some smaller errors, due to false correspondences in the correlation process.

Fig. 9 shows the result of our extraction process. Both targets are accurately located. Besides the road post, some false positives are presents. These are due to false correlation. At this time, there is no filtering process going on in order to eliminate such false positives. Figs. 10 & 11 illustrate the independent motion representation and target extraction for another part of the same sequence. In this part of the sequence, both targets (foreground and background pedestrians) are also well extracted. Figs. 12, 13, 14 & 15 present the extraction results, as well as independent motion representation for two other different sequences. Figs. 12 & 13 present a sequence in which the rotational component of the ego-motion is dominant with respect to the translational one. Figs. 14 & 15 present a much more complex scene. The images contain six different targets, with different motions.

VI. Conclusions & Future Works

A. Conclusions

A stereo-motion based algorithm that works on two consecutive disparity maps[3] and a set of feature correspondences

[2] this is due to the periodic nature of the post pattern.

[3] The frequency of acquisition of those disparity map can range from 2 Hz to 25 Hz, in other ways, some detection can still be achieved with highly time-separated events.

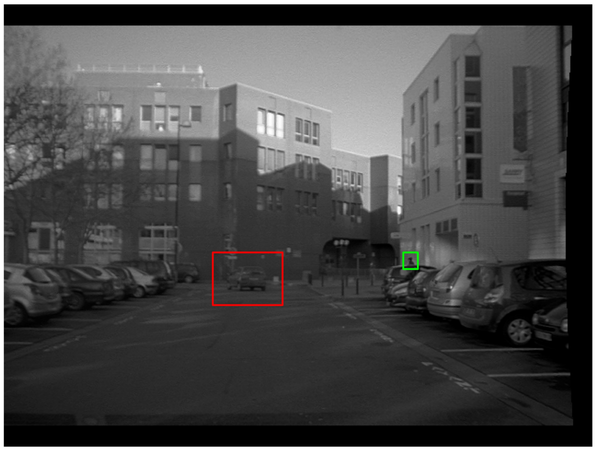

Fig. 6. Image extracted from a real test sequence. The vehicle within the red bounding box is driving forward and initiating a left bend. Within the green box lies the head of a partially occluded pedestrian walking from the right hand side of the field, toward the left-hand side.

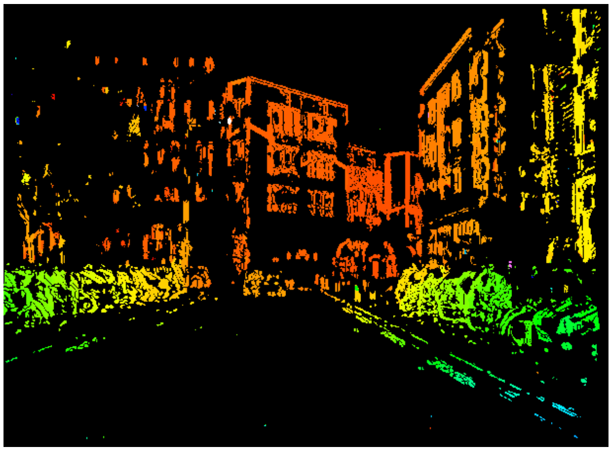

Fig. 7. Disparity Map computed with the method described in [10], from the image pair illustrated by Fig. 6. If one takes a close look, one will notice that a post on the right side of the road presents a disparity error. This is due to the periodic nature of the post pattern, *i.e.*, the upper part of the post is matched with the wrong post in the left image.

is presented here. First the ego-motion of the stereo-rig is estimated, in order to synthesize the corresponding displacement field in the disparity-space (*i.e.* optical flow extended to disparity variations). The precision of this displacement field evaluation for every point in the disparity space is estimated through error propagation. Through a robust correlation method, the points that don't fit in the displacement field can be detected. The work presented here uses this detection in order to achieve localization of independently moving objects. Though, it's worth noting that those points can also belong to disparity map errors.

B. Future Works

As seen previously, the presented method allows the detection of image-based regions that don't validate a majority motion model. These regions can be either an independent moving object, or a disparity error. Thus, this can be used to obtain a correction of the disparity map. Instead of the WTA approach used, several stereo-candidates can be stored for every point. That way, if a point does not comply with the motion estimation for a certain disparity value, but complies with it for another likely disparity-value, disparity can be re-estimated, assuming that the point is static. Errors like the one illustrated in Fig. 7 could be avoided. Such an algorithm is currently being studied.

Besides, the first drawback of this method is its computational cost. As a matter of example, on a standard laptop, a non-optimized C++ implementation runs at 5 frames (640x480) per second. However, all the involved processes present high data parallelism. So, besides using algorithmic means (*e.g.* pyramidal approach), the use of SIMD-optimized coding, or massively parallel computer architectures like GPGPU, or the CELL processor are considered. All those possibilities are currently studied.

Moreover, the method presented here relies only upon the analysis of two consecutive frames. Using some integration over time could improve the elimination of false detections and the identification of small motions. It could be realized by tracking all supposed obstacles through time and disparity space, for instance with a Kalman filter or a mean-shift approach. Such a tracking could also allow to circumvent the dominant motion hypothesis. For instance, if a large moving vehicle occupies a major part of the camera field, its motion will be interpreted as the ego-motion. Tracking this object would allow to label certain parts of the image as dynamic, feature points from within these parts would no longer be used for the ego-motion extraction. Moreover, ego-motion estimation can also be improved through the use of time integration. For instance, a Kalman filter can be used to predict ego-motion, in order to avoid potential aberrations. Such an aberration could be due to the sudden arrival of a dominant moving object in the camera field or to a low-contrast landscape, which can lead to a reduced number of extracted feature points.

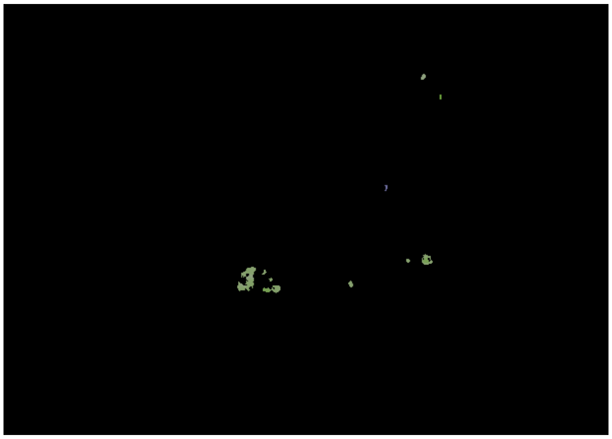

Fig. 8. The car and the pedestrian's head are well defined. Please note that a median blur is applied to the resulting image, in order to eliminate some noise points.

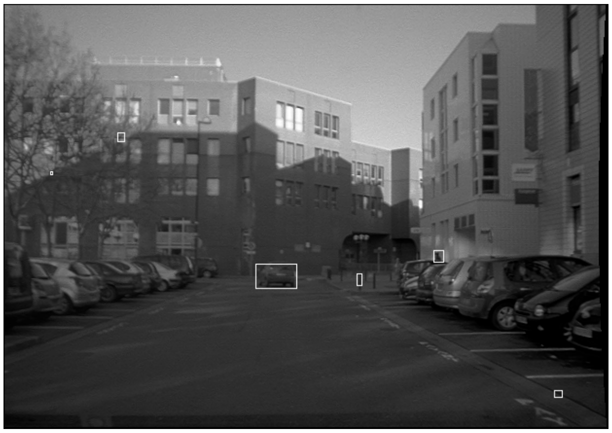

Fig. 9. Final Segmentation Realized. One can note the extraction of the road post, as well as two small false positives. Those false positives are due to false correlations. The targets, however are well defined.

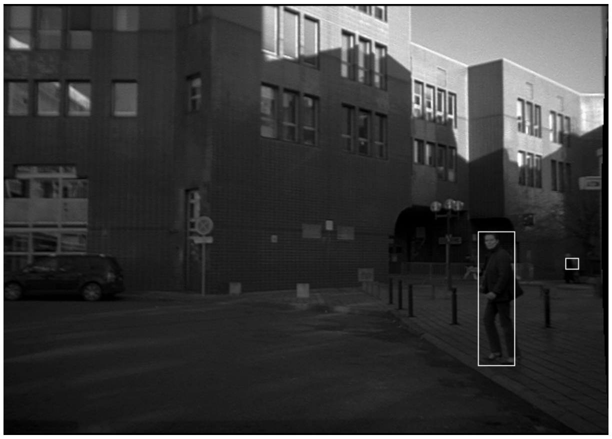

Fig. 10. Image extracted from the same sequence as Fig. 6. The two pedestrians are well extracted, even the one in the background.

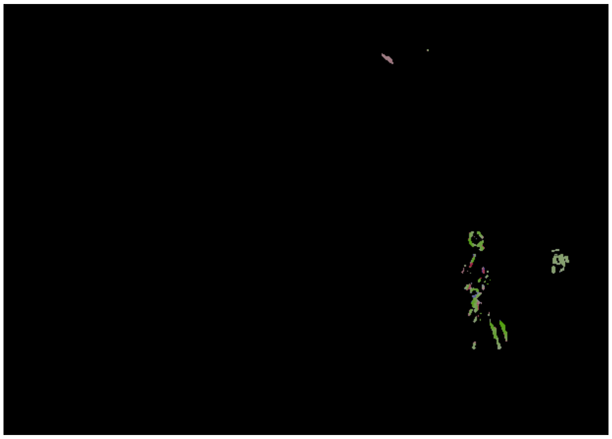

Fig. 11. Independent motion representation. One can note the pedestrian in the background.

VII. Acknowledgements

The authors would like to thanks Dr David Demirdjian for his precious help regarding the extraction of the ego-motion, as well as for his support. The authors would also like to thank Digiteo for its support to their projects.

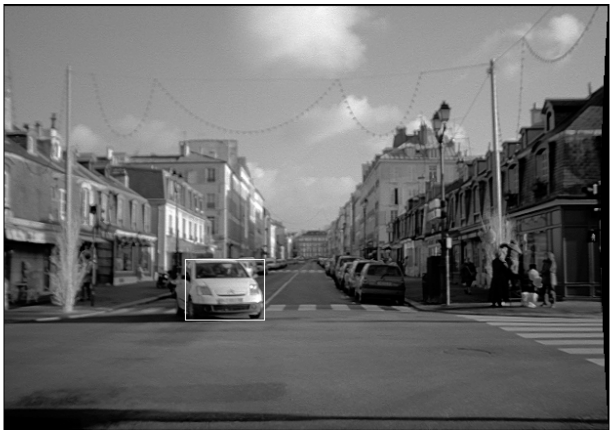

Fig. 12. This image is extracted from a different sequence, the stereo sensor is moving forward and turning to the left. The white car is well defined and detected. On the other hand the pedestrians on the right are standing still. A stereo-only detection algorithm would have detected them as potential obstacles in spite of they are on the sidewalk.

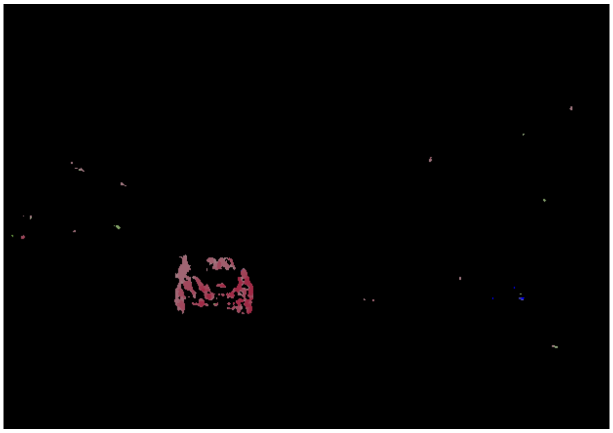

Fig. 13. Independent motion representation relative to the image in Fig. 12. One can note that the front of the vehicle appears with a wider (*i.e.* a more saturated tone) movement than its back. This is consistent with the motion of the car.

References

[1] A.D. Milner and M.A. Goodale, *The Visual Brain in Action* in Oxford University Press, 1996

[2] E. Dickmanns, *The Development of machine vision for road vehicles in the last decade*, IEEE Intelligent Vehicles Symposium, pp 268-281, vol.1, 2002

[3] M.Irani, B. Rousso and S. Peleg, *Recovery of Ego-Motion Using Region Alignment*, IEEE Transactions on Pattern Analysis and Machine Vision, pp 268-272, vol 19, n3, March 1997

[4] Y. Dumortier, I. Herlin and A. Ducrot, *4D-Tensor Voting Motion Segmentation For Obstacle Detection in Autonomous Guided Vehicle*, IEEE Intelligent Vehicles Symposium, pp 379-384, june 2008

[5] S. Beauchemin and J. Barron, *The Computation of Optical Flow*, Association for Computing Machinery Computing Surveys, pp 433-466, vol. 27, issue 3, september 1995

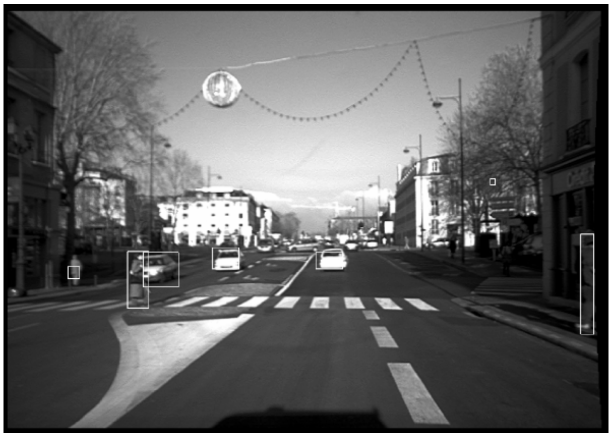

Fig. 14. This scene is more complex than the previous ones, as there are many targets, moving in different directions. However, our method extract them all. The vehicles in the background aren't detected because a minimum disparity under which data is not processed is set. This was done in this sequence only, because of the many moving vehicles in the background.

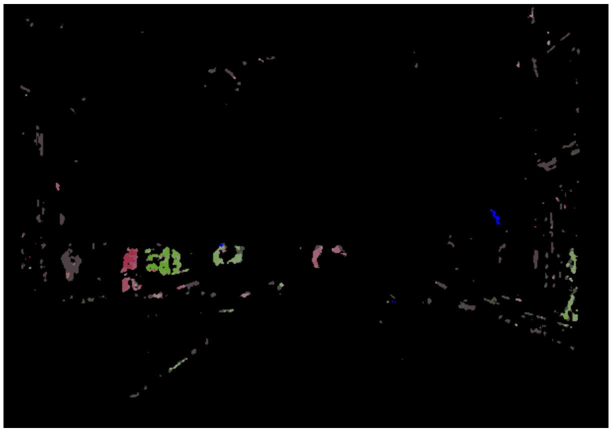

Fig. 15. Independent motion representation relative to Fig. 14

[6] G. Stein, O. Mano and A. Shashua, *A Robust Method for Computing Vehicle Ego-Motion*, IEEE Intelligent Vehicles Symposium, pp 362-368, , 2000

[7] B. Horn and B. Schunck, *Determining Optical Flow*, Artificial Intelligence, pp 185-203, vol. 17, 1981

[8] B. Lucas and T. Kanade, *An Iterative Image Registration Technique With an Application to Stereo-Vision*, Image Understanding Workshop, pp 121-130, 1981

[9] T.A. Williamson, *A High Performance Stereo Vision System For Obstacle Detection*, PhD Dissertation, Robotics Institute Carnegie Mellon University, Pittsburg, 1998

[10] N. Hautiere, R. Labayrade, M. Perrollaz and D. Aubert, *Road Scene Analysis by Stereovision : a Robust and Quasi-Dense Approach*, International Conference on Control, Autmation, Robotics and Vision, pp 1-6, 2006

[11] S. Heinrich, *Fast Obstacle Detection Using Flow/Depth Constraint*, IEEE Intelligent Vehicles Symposium, pp 658-665, vol. 2, June 2002

[12] U. Franke, C. Rabe, H. Badino and S. Gehrig, *6D-Vision : Fusion of Stereo And Motion For Robust Environment Perception*, Lecture Notes in Computer Science - Pattern Recognition, pp 216-223, vol. 3663, 2005

[13] D. Demirdjian and T. Darrell, *Motion Estimation From Disparity Images*, IEEE International Conference on Computer Vision, pp 213-218, vol. 1, July 2001

[14] A. Howard, *Real-Time Stereo Visual Odometry for Autonomous Vehicles*, IEEE Intelligent Robots and Systems, pp 3946-3952, September 2008

[15] H. Badino, *A Robust Approach For Ego-Motion Estimation Using A Mobile Stereo Platform*, Lecture Notes in Computer Science - Complex Motion, pp 198-208, vol. 3417, 2007

[16] A. Taludker and L. Matthies, *Real-Time Detection of Moving Objects from Moving Vehicles Using Dense Stereo and Optical Flow*, IEEE International Conference on Intelligent Robots and Systems, pp 315-320, Sendai, Japan, Sept. 2004

[17] D. Scharstein and R. Szeliski, *A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms*, International Journal of Computer Vision, 47 (1/2/3):7-42, April-June 2002

[18] C. Harris and M. Stephens, *A combined Corner and Edge Detector*, Alvey Vision Conference, pp 147-151, 1988

[19] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool, *SURF : Speeded-Up Robust Features*, Computer Vision and Image Understanding, pp 346-359, vol. 110, n3, 2008

[20] N. Suvonvorn, S. Bouchafa and B. Zavidovique, *Marrying level lines for stereo or motion*, Internationational Conference on Image Analysis and Recognition, Toronto, Canada, sept 28-30 2005, Proceedings, Lecture Notes in Computer Science 3656 Springer 2005.

[21] M.A. Fischler and R.C. Bolles, *Random Sample Consensus : A Paradigm For Model Fitting with Applications to Image Analysis and Automated Cartography*, Communication of the Association for Computing Machinery, pp 381-395, vol 24, june 1981

[22] R. Hartley and A. Zisserman *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2003

[23] D. Gruyer, C. Royere, N. du Lac, G. Michel and J-M. Blosseville, *SiVIC and RTMaps, interconnected platforms for the conception and the evaluation of driving assistance systems*, IEEE Conference on Intelligent Transportation Systems, 2006

[24] S.C. Johnson, *Hierarchical Clustering Schemes*, Psychometrika, pp 241-254, vol. 32, n 3, 1967