Localization of objects in automotive scenes with spatial and temporal information

Capucine Legrand, Vincent Fremont, Frédéric Large

Localization of objects in automotive scenes with spatial and temporal information

Capucine LEGRAND [1,2], Vincent FREMONT [2] and Frédéric LARGE [1]

Abstract—In the context of automotive driving assistance, this paper describes a generic (*i.e.* applicable to both vehicle interior and exterior scenes) vision based approach for scene content analysis. It makes use of temporal and spatial information from a stereoscopic sequence of images to localize objects and estimate their position and motion. The proposed method is divided into three steps. First, image features are selected, tracked and reconstructed in the 3D world space. Second, a clustering step is processed in the 5D space made of the positions and 2D motions parameters. The last step is devoted to clusters interpretation: it is out of the scope of the paper, however orientations are given to illustrate the capabilities of the proposed approach. The paper is organized as follows: first, the use of temporal and spatial information from a stereoscopic sequence is investigated. A state of the art of existing methods is presented. Then, a generic approach for object segmentation is proposed. Lastly, experimental results are presented.

I. INTRODUCTION

Throughout his drive, the driver observes and analyzes his environment mostly by vision. Vision gives him detection, localization and motion of the content of the scene. In the same way, driving assistance systems need to perform the same tasks. Considering a road scene characterized by the complexity of its geometric structure and by its dynamics, close and fast objects are susceptible to be the most dangerous ones. Two complementary vision techniques are well adapted to detect and segment such objects:

- Stereovision, which uses disparities, *i.e.* spatial differences between two simultaneous images taken from two different points of view. By analogy with a biological binocular vision system, the two sensors allow depth perception, with a better accuracy for close objects that can be more easily extracted from a disparity map.
- Apparent motion analysis, which uses the temporal differences between two images taken from the same point of view at different instants. Motion information is used by all biological vision systems to localize moving camouflaged objects. Objects with bigger relative motion, such as vehicle, can be more easily extracted from a motion field.

These two classes of approaches are usually investigated separately [1] [2] when real time constraints are needed. In the following sections, it is suggested to combine them in order to get a more efficient road scene analysis, in terms of reliability and robustness. Some key choices are proposed

capucine.legrand@mpsa.com, vfremont@hds.utc.fr, frederic.large@mpsa.com
[1] PSA Peugeot Citröen, route de Gizy, 78943 Vélizy-Villacoublay, France
[2] HEUDIASYC, UTC-CNRS, Centre de Recherches de Royallieu, 60205 Compiègne, France

so that it remains compatible with automotive application computation times.

The remainder of this paper is as follows: section II lists existing literature methods for automotive scene analysis, section III describes the proposed approach and section IV discusses the obtained experimental results.

II. RELATED WORK

Various methods using temporal, spatial information or both of them from a stereoscopic images sequence can be mentioned. Hereafter, it is proposed to briefly go through these approaches, those based on the motion, then those using only the stereoscopic information and finally those combining stereovision and motion.

A. Motion based methods

Temporal approaches for object segmentation in a road scene have to deal with various cases depending on whether the pair of cameras and the obstacles are static or mobile. Methods for the analysis of road scenes address the general case, where both the pair of cameras and the obstacles are moving.

A common approach consists in calculating the egomotion to cancel it and come down to a static camera case. To estimate egomotion, the motion of the road can be calculated by the use of a probabilistic function as in [3], or by wavelets in [2]. The optical flow [4] is also used, by Giachetti [5] with vehicle motion hypotheses, and by Enkelmann [6] with a planar world hypothesis. The obstacles are then identified by extracting the areas where the motion is different from the estimated global motion of the scene. Without calculating explicitly the egomotion, Torr [7] assumes that objects are far from the camera and characterizes the motion of the background as an affine transformation. Areas that do not fit this hypothesis are then identified as objects in motion.

Those constraints may not always be realistic. Moreover, the methods based only on motion estimation lack a robust detection of obstacles when their relative velocity is too small. That leads to the following observation: they cannot be considered as the best candidates to tackle the difficulties of 3D localization of road scene obstacles.

B. Stereoscopic based methods

Two classes of stereoscopic approaches are proposed in the literature. The first aims at reconstructing the 3D scene [1], or part of it (the road in [8]) from the perceived elements. The road plan can thus be rebuilt using the disparity map, making the elements located above this plan easier

to identify as objects. Criteria on position, orientation [9], neighborhood, or disparity similarity [10] are used to gather or to separate detected objects. The second class deals with specific representation of the disparities that may be better adapted to some applications. In [11] a representation called "v-disparities" makes the road appear as a slanted segment and obstacles as vertical segments. Bertozzi [12] rectifies the images in order to match, by projection, pixels associated with the road in both images, and thus rejects the objects that do not belong to the road.

These approaches, often very specific, do not take use of the motion of the obstacles in the scene, and are not generic enough for most of the automotive applications.

C. Stereokinetic based methods

The simplest stereokinetic methods use independently a motion based method and a stereoscopic method. In [13] segmentation is processed separately on motion and disparity, and the results are fused together by comparison. A probabilistic fusion can also be used [14] for this segmentation step. Another approach is to use the stereoscopic information as a way to improve the results of a motion based method. In [15], 2D motion vectors (optical flow) are segmented by using 3D models of motion and hypotheses on the camera motion. The stereoscopic matching completes the monovision analysis by adding depth information. In [16], a rough stereoscopic matching of segmented areas extracted in both left and right images allows to eliminate discontinuities and occultations in the scene. Some other methods use the stereoscopic information to extract areas where temporal information brings added value. In [17], features not matching the planar world assumption are associated with interest areas. The motion of segments belonging to these areas is then estimated through the use of a Kalman filter, in order to bring them together into objects.

The last major approach consist in estimating and segmenting at the same time all the motion fields, *i.e.* to use simultaneously spatial and temporal information. This is done for example in [18] where disparity, segment fields and optical flow are estimated simultaneously.

The stereokinetic based methods have been proved to be efficient but are still more complex than previous ones.

D. Work orientations

The method proposed in this paper aims to maintain the effectiveness of stereokinetic methods that allows temporal and spatial information complementarities while reducing the complexity and specificity. A method without preconception or assumption, which operates in real time, is chosen. To increase speed and robustness, we choose a sparse processing approach, by working on image features, not on all pixels. The principle of the proposed method is to combine these features according to their position in the space as well as their instantaneous 2D displacement. Hence, points with the same projected motion between two instants belong generally to the same object if they are neighbours.

III. PROPOSED METHOD

Let us consider a 3D point: $\vec{p} = \begin{bmatrix} X & Y & Z \end{bmatrix}^T$ and its rigid motion $\vec{V} = \begin{bmatrix} V_x & V_y & V_z \end{bmatrix}^T$. With a calibrated stereoscopic system with rectified parallel cameras (pinhole model), four images are available at times $t-1$ and t: I_r^t, I_l^t, I_r^{t-1} and I_l^{t-1}. The following relation between the 3D point and its projection $(x_r(t), y_r(t))$ in right image at t is:

$$X(t) = \frac{x_r(t).b}{d(t)} \quad Y(t) = \frac{y_r(t).b}{d(t)} \text{ and } Z(t) = \frac{f.b}{d(t)} \quad (1)$$

with $d(t) = x_l(t) - x_r(t)$ the disparity, b the baseline and f the focal length. The projected motion on the image, also called optical flow, is $(u, v) = (x_r(t) - x_r(t-1), y_r(t) - y_r(t-1))$.

As Fig.1 shows, a three steps method is proposed: extraction of 3D features from the scene, segmentation of the features in blobs and interpretation of these blobs.
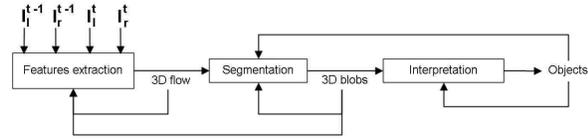

Fig. 1. Method diagram

- The extraction step aims at obtaining, for each feature, 3D localization $[X, Y, Z]^T$ and apparent 2D motion (u, v) characteristics as illustrated on Fig.2.
- The segmentation of the features in blobs allows to associate to a same object, the points belonging to the same spatial area and the ones having the same projected motion during a period of time. This clustering step is illustrated Fig.2.
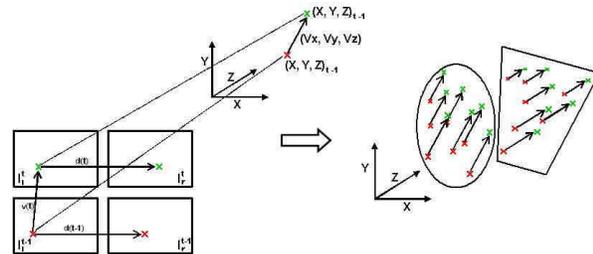

Fig. 2. Features extraction and segmentation

- Once obstacles are localized, dangerous ones are detected and objects are identified (vehicle, background, pedestrian, and others), in a last interpretation step.

A. Features extraction

The proposed method is based on features selection, their tracking and their reconstruction in the 3D space.

1) Features selection: here "Features" stands for segments, corners, points of interest or boundaries in the image. Working with features rather than with points is preferred because of temporal and spatial quality matching considerations. Moreover, easily identifiable features enhance the matching strength. The use of points of interest instead of working on all the pixels also reduces significantly the computation time.

As the mobile camera sees the scene from different viewpoints, the description of a point must be invariant to rotations, translations and illumination changes. Mozos shows [19] that the method of Harris [20] has a better repeatability and a good stability with regard to heavy computing time methods such as SIFT [21]. This method is based on the maximization of a self correlation function between a window and the same window shifted in several directions.

2) Features tracking: most of the methods used to track points are based on the hypothesis that each point keeps its luminance and its neighborhood. The tracking can be performed either by correlation, differential or frequency-based methods. According to [22], differential methods, based on the resolution of the optical flow equation, have two main advantages: direct subpixel motion estimation and low computation cost. Features tracking must be fast and precise for a good 3D reconstruction. As mentioned in the study of Barron [23], the differential method of Lucas and Kanade [24] fits these criteria. In this method, the optical flow is calculated by forming hypotheses of luminance conservation and weak motions between two consecutive images. This method is based on the optical flow equation stemming from a Taylor development and on the hypothesis of a locally constant flow on a neighborhood. Finally, the optical flow is found as the vector that matches the best with the equation in this neighborhood.

To improve the tracking, points maximizing four proposed confidence criteria are chosen:

- the quality criterion C_{ql} is defined during the selection of Harris points: the matrix based on autocorrelation introduced by Harris is used:

$$M(x,y) = e^{\frac{-(x^2+y^2)}{2\sigma^2}} \otimes \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \quad (2)$$

with $\otimes$ the convolution operator, σ^2 the variance, and I_x (resp. I_y) the first order derivative of image I in x (resp. y) direction. The confidence criteria C_{ql} is defined with the Harris function:

$$C_{ql}(p_n^t(x,y)) = \\ \mathrm{Det}(M(x,y)) - 0.04 \times (\mathrm{Trace}(M(x,y))^2 \quad (3)$$

with $p_n^{t-1}(x',y')$ the nth point detected at time $t-1$ with coordinates (x',y') tracked at time t in $p_n^t(x,y)$ with coordinates (x,y).

- the temporal criterion C_{tp} increases the trust in points

that are easy to track:

$$\begin{aligned} &\text{if } \exists\ p_n^{t-1}(x',y') \text{ and } C_{tp}(p_n^{t-1}(x',y')) < 1 \\ &\qquad C_{tp}(p_n^t(x,y)) = C_{tp}(p_n^{t-1}(x',y')) + 0.1 \\ &\text{else if } \exists\ p_n^{t-1}(x',y') \text{ and } C_{tp}(p_n^{t-1}(x',y')) = 1 \\ &\qquad C_{tp}(p_n^t(x,y)) = C_{tp}(p_n^{t-1}(x',y')) \\ &\text{else} \\ &\qquad C_{tp}(p_n^t(x,y)) = 0.1 \end{aligned}$$
(4)

- the similarity criterion C_{sim} discredits the small similar points on a neighbourhood (a $w \times w$ window) between two values of time with a correlation indicator:

$$C_{sim}(p_n^t(x,y)) = \\ \sum_{i=-\frac{w}{2}}^{\frac{w}{2}} \sum_{j=-\frac{w}{2}}^{\frac{w}{2}} \frac{\left| I^t(x+i,y+j) - I^{t-1}(x'+i,y'+j) \right|}{w^2} \quad (5)$$

- the motion criterion C_{mot} eliminates the points with inconsistent optical flow:

$$\begin{aligned} &\text{if } \sqrt{(x-x')^2 + (y-y')^2} < S_{mot} \\ &\qquad C_{mot}(p_n^t(x,y)) = 1 \\ &\text{else} \\ &\qquad C_{mot}(p_n^t(x,y)) = 0 \end{aligned}$$
(6)

with S_{mot} a threshold on the motion norm. This threshold is tuned considering the projection in the image of the maximal relative motion of an object in the scene.

The final confidence criterion is calculated for each point:

$$\begin{aligned} &\text{if } C_{mot}(p_n^t) = 0 \\ &\qquad C(p_n^t(x,y)) = 0 \\ &\text{else} \\ &\qquad C(p_n^t(x,y)) = \frac{C_{ql} + C_{tp} + (1 - C_{sim}) + C_{mot}}{4} \end{aligned}$$
(7)

An example of optical flow is presented Fig.3: the 2D motion is represent, between two consecutives moments, for points selected in the image.

3) 3D reconstruction: the 3D reconstruction of features is done through the two entry images at time t (left and right images). The sparse disparity map of the image is calculated. Then, the depth of the points of interest can be deduced, as well as their 3D positions.

Matching the points of interest between the left and the right images is based on correlation methods by comparing their neighborhoods. This matching research is made only along the horizontal axis because the images are supposed to be rectified. Furthermore, to make the results more robust, the correlations between the right and left images are crossed. The optimization approach of "WinnerTakes-All" [25] indicates that the best matching between the two correlation windows corresponds to the extremum of a cost function (SAD or ZNCC for example). The SAD (Sum of Absolute Differences) and ZNCC (Zero mean Normalized Cross Correlation) methods were evaluated: the obtained results confirm the conclusions of [25] and the ZNCC method has finally been preferred for its invariance in the uniform variations of luminance in the images, even if it increases the computing time. An example of disparity map obtained by ZNCC is illustrated Fig.3, the more the points are far from the camera, the darker they are.
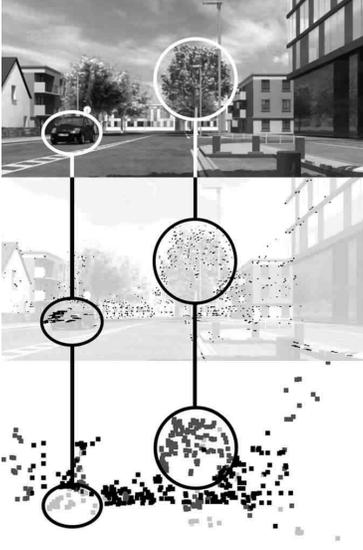
Fig. 3. Sparse optical flow and disparity map (light points are near to the camera and dark points are far) on a virtual sequence

Since stereoscopic system characteristics are known, the points of interest are then reconstructed in the 3D space at any given time, using their disparity $d(t)$ in equation (1).

B. Features segmentation

At this step, two types of information are available for each point: 3D position $[X, Y, Z]^T$ and projected motion (u, v). These five variables are used in a clustering procedure in order to bring together points that have a close 3D localization and a similar optical flow. It has to be noticed that the optical flow is preferred to 3D motion (available from 3D positions at two instants), because of the lower quality of the disparity map, that is not accurate enough to obtain an exploitable 3D motion.

Thus, a partition $P = \{C_1, C_2, \ldots, C_k\}$ is to be achieved from a set $J = \{p_1, p_2, \ldots, p_n\}$ of points of interest with: $C_1 \cup C_2 \ldots \cup C_k = P$ and $C_i \cap C_j = \phi$ with $i, j = 1, 2, \ldots, k$.

A choice of various unsupervised methods is available to proceed this clustering task [26]. The Hierarchical Ascending Classification (HAC) is chosen because of its easy way to use, for the possibility to deal with large data sets with few variables and because it is swift.

The HAC principle is to collect points according to a criterion of distance to classify in homogeneous groups the features points whose characteristics in different dimensions resemble most each other in a criterion of distance sense. This method determines among n individuals, the two individuals that look most alike with regard to the p specified variables ($p = 5$ in our case), and brings them together to form a cluster. At this level there are $n-1$ clusters, one being formed by two individuals, the $n-2$ others containing only a single individual. This process is iterated to determine which are the two clusters which look most alike, and by bringing

them together. This operation is repeated until a single cluster grouping all individuals is obtained. This process is based on the choice of a similarity criterion between the individuals and an aggregation criterion (dissimilarity between clusters). The interclasses distance used is the Euclidian distance. And, the aggregation criterion is the Ward criterion which consists in choosing at every stage the clusters that can be gathered with the minimal increase of intraclasses inertia. This criterion minimizes the variance within groups and maximizes the variance between groups and thus promotes the extraction of well separated clusters. The HAC leads in a stack of partitions that must be cut at a given threshold (tuned manually) for clustering.

To improve this clustering step, it is proposed to calculate the rigid motion of each cluster found to first eliminate the points with aberrant motion (compared to the motion of the cluster in which they are clustered) and second to group together clusters where the same motion is observed. The optical flow and depth constraints are expressed in the disparities space. Indeed, this space is projective and, for a parallel camera stereo rig, the noise is isotropic [27]. To estimate 3D motion, the linear system combining these two constraints is solved. For small rotation a linearization gives the 3D motion equation: $\vec{V} \approx \vec{T} - \vec{X}\vec{\Omega}$ with $\vec{T} = \begin{bmatrix} t_x & t_y & t_z \end{bmatrix}^T$ an instantaneous translation vector, and $\vec{\Omega} = \begin{bmatrix} \omega_x & \omega_y & \omega_z \end{bmatrix}^T$ an instantaneous rotation vector. In the disparities space, this constraint can be written:

$$\begin{bmatrix} \frac{d(t)}{d(t+1)}x(t+1) - x(t) \\ \frac{d(t)}{d(t+1)}y(t+1) - y(t) \\ \frac{d(t)}{d(t+1)}f - f \end{bmatrix} =$$

$$\begin{bmatrix} \frac{d(t)}{b} & 0 & 0 & 0 & f & -y(t) \\ 0 & \frac{d(t)}{b} & 0 & -f & 0 & x(t) \\ 0 & 0 & \frac{d(t)}{b} & y(t) & -x(t) & 0 \end{bmatrix} \begin{bmatrix} t_x \\ t_y \\ t_z \\ \omega_x \\ \omega_y \\ \omega_z \end{bmatrix} \quad (8)$$

The second constraint is the depth constraint of Harville [28].

$$Z(x, y, t) + V_z(x, y, t) = Z(x + v_x, y + v_y, t+1) \quad (9)$$

Equation (9) is also written in the disparities space as (the notation t is omitted for reading simplification):

$$-\frac{\partial d}{\partial t} = \frac{d}{fb} \begin{bmatrix} f\frac{\partial d}{\partial x} & f\frac{\partial d}{\partial y} & -(-d + x\frac{\partial d}{\partial x} + y\frac{\partial d}{\partial y}) \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & Z & -Y \\ 0 & 1 & 0 & -Z & 0 & X \\ 0 & 0 & 1 & Y & -X & 0 \end{bmatrix} \begin{bmatrix} t_x \\ t_y \\ t_z \\ \omega_x \\ \omega_y \\ \omega_z \end{bmatrix} \quad (10)$$

Equations (8) and (10) are stacked into a linear system solved by the Singular Values Decomposition method (SVD). Outliers are rejected by the method of M-estimators introduced by Huber [29]. The clusters with similar motion are combined.

C. Clusters interpretation

This very last stage depends a lot on the application and is not detailed in this paper devoted to generic part. Nevertheless, some orientations are proposed for the cases presented in the next section.

To interpret the clusters, it is necessary to distinguish the background from other objects. The cluster the most dispersed in the 3D space is identified as the background. To identify the other objects, it is possible to determine the category of these objects (pedestrian, car, truck) based on the knowledge about their real size, by using 3D information. Moreover, the localization of 3D obstacles allows to calculate the distance which separates the camera from these objects and to deduce, with motion information, an estimate of the time to collision. It can thus lead to accurate information on the clusters, even if this stage requires *a priori* knowledge.

IV. RESULTS

The first steps of the proposed method (features extraction and clustering) were experimented on both virtual and real sequences, representative of typical automotive scenes. Moreover, sequences of cockpit as well as outside frontal scenes have been used to validate the genericity of the approach. Only frontal sequences, more challenging, are presented here.

The computational efficiency of the method has been evaluated with several number of clusters and features. A highter number of clusters or features tend to improve the results, however it implies longer computation time. Hence, an arbitrary number of 5 clusers and 150 features has been choosen as the best observed compromise. Running on a 1.7GHz laptop computer using windows 2000, on a C/C++ implementation with no particular code optimization, the proposed approach averages 10 frames per second with 150 features per image.

The following figures are showing the raw output of the method before inetrpretation. Each feature point is assigned to a cluster that is represented by a specific shape and color. Fig.4 illustrates the results obtained on the virtual sequence. The cluster composed of squares (see (a) in Fig.4) is associated to the crossed vehicle. Outliers (see (b) in Fig.4) come from the way the scene has been built: repeated textures lead to locally inconsistencies in the disparities. The "circles" cluster corresponds to motionless distant points (for which the disparity is undetermined). The other clusters can be associated with other objects of the background. Similar results have been obtained with sequences acquired on real cameras as illustrated in Fig.5. Results allow to localise the vehicles over time. In Fig.5 frontal vehicles are well detected by the same color in each image. The added value of this stereokinetic approach with regard to a method using only the motion or only the stereoscopic information is illustrated Fig.6. To show this added value, the stereokinetic results are compared with results obtained with the same clustering method used only on motion data or only on 3D data. Motion based clustering does not separate objects with similar motion: see (a) and (b) in Fig.6 (top). A stereoscopic
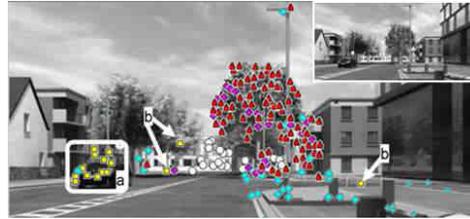

Fig. 4. Clustering results on a virtual sequence (Bounding boxes and arrows are manually added for reading simplification)
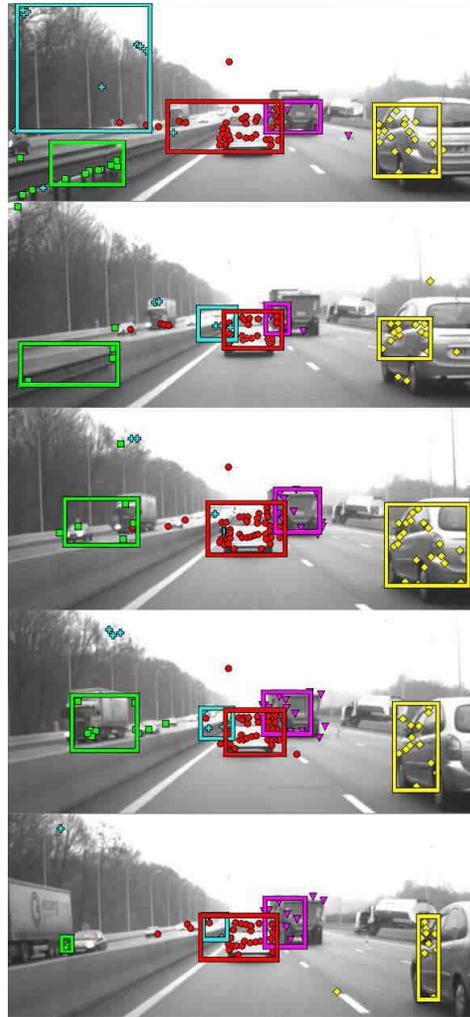

Fig. 5. Clustering results on a real sequence at different times

based approach with a clustering driven mostly by depth information also leads to misclassified objects, as shown in (c) and (d) in Fig.6 (middle). These problems are solved with the use of the mixed proposed method. Each cluster (see (e), (f) and (g) in Fig.6 (bottom)) corresponds to one object. A last cluster (triangles) does not verify this observation.

However it can be easily filtered by considering the density of the associated points.

First experimental results tend to prove that this algorithm

Fig. 6. Clustering results with motion based method (top), stereovision based method (middle), and stereokinetic based method (bottom)

is a good candidate for automotive scenes interpretation. Further more the clustering can be improved by tracking each cluster.

V. CONCLUSION AND ORIENTATIONS

In the automotive scope, the danger often comes from fast and/or close obstacles. Vision based methods exploiting both temporal and spatial information from a sequence of stereoscopic images are well suited to localize such obstacles.
In this paper, a generic approach is proposed. It consists in 3 steps: selection and tracking of feature points, clustering of the points according to their position and motion, and interpretation of the clusters. Experiments on both indoor vehicle applications such as occupant characterization (not presented here), and outside application such as frontal obstacle detection, allowed to validate the genericity of the approach. The obtained results turned out exploitable on all the tested sequences. However the classification can be furthermore improved with 3D motion estimation and tracking of the clusters. The interpretation step, specific to the targetted application, remains to be implemented.

REFERENCES

[1] V. LEMONDE, Stéréovision Embarquée sur Véhicule: de l'Auto-Calibrage à la Détection d'Obstacles. *Thèse de doctorat*, Institut National des Sciences Appliquées de Toulouse, 2005.

[2] C. DEMONCEAUX, A. POTELLE, D. KACHI, Obstacle detection in road scene based on motion analysis, *5th IFAC Intelligent autonomous vehicles*, 2004.

[3] G. STEIN, O. MANO, A. SHASHUA, A robust method for computing vehicle ego-motion, *Intelligent Vehicles*, pp. 362-368, 2000.

[4] B. HORN, B. SCHUNCK, Determining Optical Flow, *Artificial Intelligence*, Vol. 17, pp. 185-204, 1981.

[5] A. GIACHETTI, M. CAMAPANI, V. TORRE, The use of optical flow for road navigation, *IEEE transactions on robotics and automation*, Vol. 14(1), pp. 34-48, 1998.

[6] W. ENKELMANN, Obstacle detection by evaluation of. optical flow fields form image sequence, *Image and Vision Computing*, Vol. 9(3), 1991.

[7] P. TORR, D. Murray, Statistical detection of independent movement from a moving camera, *Image and vision computing*, Vol. 11(4), pp. 180-187, 1993.

[8] S. NEDEVSCHI, R. SCHMIDT, T. GRAF, R. DANESCU, D. FRENTIU, T. MARITA, F. ONIGA, C. POCOL, 3D lane detection system based on stereovision, *Intelligent Transportation Systems Conference*, pp. 161-166, 2004.

[9] T. WILLIAMSON, A High-Performance Stereo Vision System for Obstacle Detection, *Carnegie Mellon Technical Report*, CMU-RI-TR-98-24, 1998.

[10] R. ZHANG, R. WEISS, A. HANSON, Qualitative obstacle detection, *Tech. Report COMPSI*, TR94-20, 1994.

[11] R. LABAYRADE, D. AUBERT, J. TAREL, Real time obstacle detection in stereovision on non flat road geometry through v̈-disparitÿrepresentation, *Intelligent Vehicle*, Vol. 2, pp. 646-651,2002.

[12] M. BERTOZZI, A. BROGGI, GOLD: a Parallel Real-Time Stereo Vision System for Generic Obstacle and Lane Detection, *Image Processing*, Vol. 7(1), 1998.

[13] C.H.YANG, Joint disparity/motion estimation and segmentation for object-oriented stereoscopic image coding, *INRS Télécommunications Technical Report*, pp. 97-05, 1997.

[14] C. BRAILLON, K. USHER, C. PRADALIER, J. CROWLEY, C. LAUGIER, Fusion of stereo and optical flow data using occupancy grids, *Intelligent Transportation Systems Conference*, pp. 1240-1245, 2006.

[15] J. WANG, E. ADELSON, Spatio-temporal segmentation of video data, *SPIE: Image and Video Processing II*, vol. 2182, pp. 120-131, 1994.

[16] A. WAXMAN, Binocular image flows: steps toward stereo-motion fusion, *IEEE Trans. On pattern analysis and machine intelligence*, Vol. 8(6), pp. 7-15, 1986.

[17] Z. ZHANG, O. FAUGERAS, Three-dimensional motion computation and object segmentation in a long sequence of stereo frames, *International Journal of Computer Vision*, Vol. 7, No. 3, pp. 211-241, 1992.

[18] Y. ALTUNBASAK, A. TEKALP, G. BOZDAGI, Simultaneous motion-disparity estimation and segmentation from stereo, *ICIP*, pp. 73-77, 1994.

[19] O. MOZOS, A. GIL, M. BALLESTA, O. REINOSO, Interest Point Detectors for Visual SLAM, *Proceedings of the Conference of the Spanish Association for Artificial Intelligence*, 2007.

[20] C. HARRIS, M. STEPHENS, A combined corner and edge detector, *Alvey sision conference*, pp. 147-151, 1988.

[21] D. LOWE, Distinctive image features from scale-invariant keypoints, *Int. J. Computer Vision*, Vol. 60(2), pp. 91-110, 2004.

[22] E. BRUNO, De l'estimation locale à l'estimation globale de mouvement dans des séquences d'images *PhD thesis*, Université Joseph Fourier, Grenoble, 2001.

[23] L. BARRON, S. BEAUCHEMIN, D. FLEET, Performance of optical flow techniques, *In International Journal on Computer Vision*, vol. 12, pp. 43-77, 1994.

[24] B. LUCAS, T. KANADE, An iterative image registration technique with an application to stereo vision, *proc. DARPA IU workshop*, pp. 121-130, 1981.

[25] D. SCHARSTEIN, R. SZELISKI, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, *International Journal of Computer Vision*, Vol. 47(1), pp. 7-42, 2002.

[26] L. LEBART, A. MORINEAU, M. PIRON, Statistique exploratoire multidimensionnelle, *Editions Dunod*, 1997.

[27] D. DEMIRDJIAN, T. DARRELL, Motion estimation from disparity images, *Computer Vision*, pp. 213-218, 2001.

[28] M. HARVILLE, A. RAHIMI, T. DARRELL, G. GORDON, J. WOODFILL, 3D pose tracking with linear depth and brightness constraints, *Computer Vision*, 1999.

[29] P. HUBER, Robust Estimation of a Location Parameter, *Annals of Mathematical Statistics*, Vol. 35, pp. 73-101, 1964.