Homography from a Vanishing Point in Urban Scenes

Nicolas Simond, Patrick Rives
INRIA, 2004 route des Lucioles, 06 902 Sophia-Antipolis Cedex, France
E-mail : Firstname.Lastname@sophia.inria.fr

***Abstract*— In this paper, we address the problem of computing the egomotion of a vehicle in an urban environment using dynamic vision. We assume a planar piecewise world where the planes are mainly distributed along three principal directions corresponding to the axes of a reference frame linked to the ground plane with a vertical z-axis. We aim to estimate both the motion of the car and the principal planes in the scene corresponding to the road and the frontages of the building from a sequence of images provided by an on-board uncalibrated camera. In this paper, we present preliminary results concerning the robust segmentation of the road using projective properties of the scene. We develop a two-stage algorithm in order to increase robustness. The first stage detects the borders of the road using a contour-based approach and primarily allows us to estimate the Dominant Vanishing Point (DVP). The DVP and the borders of the road are then used to constrain the region where the points of interest, corresponding to the road lane markers, can be extracted. The second stage uses a robust technique based on projective invariant to match the lines and points between two consecutive images in the sequence. Finally, we compute the homography relating the points and lines lying on the road into the two images.**

***Index Terms*— Dominant vanishing point, cross-ratio, homography, road plane, urban environment**

I. Introduction and related work

A safe navigation of autonomous vehicles in an outdoor environment requires a robust localization process. During the last decade, the DGPS has become the most used technology for all outdoor environment applications. Nevertheless, the localization quality depends on the number of satellites the antenna can receive. High buildings and trees decrease the signal/noise ratio by obstructing the clear view and multiple paths corrupt the data. Chen [1] indicates 95% of the Tokyo urban area does not allow a GPS-based location. Furthermore, the resolution available with such a system is about one meter in the best case. That is not sufficient for the localization and guidance of a driverless vehicle which requires about ten centimeters accuracy.

Concurrently, due to the increasing power of computers, it is now realistic to use vision sensor(s) as a major part in the localization process in association with a DGPS-based system. Indeed, the structured environment allows to compensate for the loss of clear view. The urban scenes contain generally sets of parallel lines, the majority of them are aligned with the principal orthogonal directions of the world coordinate frame. Hence, a vision-based system sounds like a natural complement to the GPS information in the highly-urbanized area.

Nevertheless, the architectural characteristics of urban areas vary according to the brightness and the shadow conditions and to the type of scenes, from ”open” environments like large boulevards and main streets to ”closed” ones like the downtown old cities. In these different cases, the determination of consistent structures like streets and buildings in the image is one of the aims of segmentation we have to achieve.

All vision-based localization methods in urban areas are developed on two common assumptions:

- the ground is locally assumed to be plane,
- the man-made environment contains sets of ortho-parallel lines.

Parallel lines in the 3D scene converge in the image to a peculiar point, so called Vanishing Point (VP), when they are viewed under perspective projection. The understanding and interpretation of man-made environments can then be greatly simplified by the detection of such vanishing points.

Some authors only focus on improving the VP coordinate precision to better reconstruct the 3D scene. They generally present post-processing methods which enhance the edge detection quality. Rother [2] subsets the detected edges into 3 mutual orthogonal directions with respect to orthogonality, camera and vanishing line criteria. In the same way, Kosecka and Zhang [3] combine efficient image processing techniques and expectation maximization algorithm to partially calibrate a camera and estimate its relative orientation with respect to the scene.

The same geometrical environment assumptions and VP methods are also be used by authors to deal with mobile robot navigation in indoor environments for mobile robot navigation. Guerrero and Sagues [4] have developed such a vision-based navigation algorithm. They succeed in determining a qualitative free space ahead by compensating the rotation motion with a monocular uncalibrated camera. Lebegue and Aggarwal [5] describe an algorithm to automatically reconstruct environments like hallways.

Snaith et al. [6] work on a prototype vision system for the guidance of visually impaired people through urban environments. They detect doorways and vertical edges to facilitate center path travel. The dominant vanishing point (DVP), intersection of the majority horizontal vanishing

lines (VLs) in image, is computed as the highest accumulator score of a Hough Transform with the detected edges. Antone and Teller [7] combine Hough Transform and expectation maximization to finally decouple the rotation and the translation motions between the successive camera poses in an urban scene.

Otherwise, the geometrical and photometrical properties of roads allow some specific methods to segment them in the image plane. Some authors detect the road frontier edges assuming non-occluded parallel road lane markers. They generally track them in a video sequence with the introduction of a region of interest. Wang et al. [8] detect and track the natural edges of a uniform textured road. Sotelo et al. [9] succeed in isolating the ground plane in the image assuming a model of the road and a HSV decomposition.

On the other hand, Okutomi et al. [10] locate the ground plane by computing the image projective invariants using a calibrated stereovision system. Hu and Uchimura [11] propose a new model of multi-lane structured road, assuming road boundaries can be modeled by clothoids in order to simplify the matching between the 3D scene and the projective image.

The next section gives an overview of the proposed method highlighting the underlying assumptions on the scene and on the camera model. Section 3 describes how to estimate a camera displacement between two views of an urban scene and section 4 shows some experimental results. The last section concludes this article and introduces our future work.

II. Approach

A. Problem statement

Assuming the urban scene contains planar structures, it becomes possible to fully characterizes the camera displacement in the projective space by computing the homographies between two consecutive images of the scene. In this paper, we focus on the computation of the homography from the points and lines lying on the road.

Let us assume that:

- the urban road scene images contain sets of 3D orthoparallel segments,
- the road is locally planar with parallel boundaries,
- the camera model used is a pinhole camera model,
- the video sequence is recorded at a high frame rate, which means that the DVP coordinates move slowly between two consecutive images.

Nevertheless, all these restrictive assumptions can not be verified in an urban environment. The video sequences contain such an amount of dynamic objects (cars, pedestrians) occluding a significative part of the road. Fortunately, the urban environment contains an abundance of 3D lines. These lines are primarily the edges of the building frontages, the limits between two depth planes and the boundaries between two different chromatic regions. Roughly, the model of an urban street can hence be regarded as a hallway. We then particularly focus on all the left, bottom and right image foreground lines due to their high probability to belong to a VL.

Although the discrimination between the static and the dynamic part of an urban scene is difficult due to the complexity of the environment, the ground plane appears as the object of the static scene that has a fixed location in the image. We can use the assumption that the image contains a road plane bounded by two 3D curbs to constrain both the static environment and the road plane projection on the image.

The road generally appears as a large area whose borders are highlighted either by a kerb or by road markers, sometimes both. These limits are easily detected by their color variations (see Figure 1). Whatever the type of road (straight or curved), at the foreground of the image, the edge limits can be modeled by some segments. These segments are located in the image at each bottom corner with two opposite orientations and they converge to the DVP located in an area in the center of image.

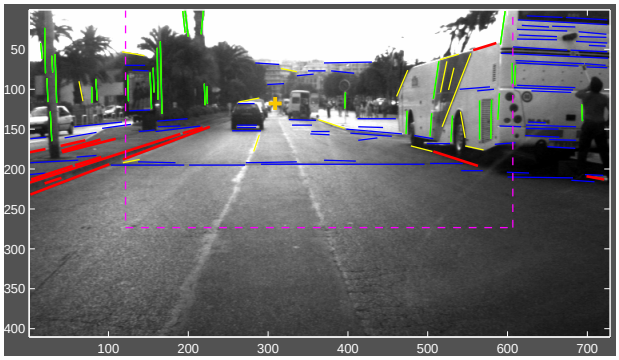

Fig. 1. Edges classification. We represent in blue (nearly) horizontal and in green (nearly) vertical segments which naturally intersect all other segments in yellow. The highlighted red segments are close enough to the precedent DVP location (orange '+') to be considered as potentially belonging to a future VL. The fixed confidence area, limited by the dashed magenta lines, represent the image foreground.

However, dynamic objects like cars and pedestrians occlude most of the static scene. Furthermore, they generally possess many edges which could potentially confuse a DVP detection program. We thus try to detect not only the road boundaries, but all the sets of 3D parallel lines which converge to the DVP in the image in order to robustify the DVP estimation process. We do not also address (nearly) horizontal segments because they rarely represent parts of horizon VL, but they naturally intersect the other image lines and could therefore induce DVP misdetections. Two cases are distinguished with the (nearly) vertical segments. If they are located outside the two road boundaries, they represent a 3D vertical direction, which can be used in a further reconstruction process. Otherwise, they could represent either a road lane marker or a vertical edge of an obstacle.

III. Tracking and matching features in an urban scene image

A. Identifying vanishing lines

Classicaly, a Canny operator is used to detect the edges and after polygonal approximation, to obtain the segments of lines. Due to the bad quality of the images, the edge detection data are particularly inaccurate. According to our experiments, the midpoint location seems to be the only reliable characteristic of each segment. The length and the orientation of extracted segments are noisy. However, we note that the longer the segment is, the more correctly located it is.

We thus come to the conclusion that it is better to reconstruct partially occluded VLs rather than trying to obtain an estimation of the DVP using corrupted segments. We first fit each segment to a line parameterized by an angle θ and a distance ρ from the origin:

$$\rho = cos\theta.u + sin\theta.v$$

where (u, v) are an image pixel coordinates.

The DVP tracking method assumes that the distance between two camera poses remains small. The DVP location in the $(n-1)$ image can hence be considered as a prediction of the current estimation at time (n). In the same way, the VL motions in image are minor.

We compute the distance between each segment line and the DVP position estimated on the previous image. We consider only the segments which have a distance smaller than a threshold dmax that we experimentally fix at 20 pixels. We then clusterize the selected segment lines in VL candidates by grouping them according to common characteristics. Besides, the segment lines are rated by their single segment length order. The longest is considered as a reference. We also search other segment lines that verify the following criteria in order, until they all belong to a VL candidate:

1) the contrast direction across the edge,
2) the segment orientation θ,
3) the image location under or above the horizon line,
4) the compatibility criterion. We expect that the segments which formed the same VL have endpoints location whose coordinates respect the appearance order (refer to Figure 2),
5) the distance between the segments and the reference segment midpoint is reduced.

The result of the grouping process is a list of K candidates of VLs. The DVP estimate (u_X, v_X) is then obtained by solving iteratively the following linear weighted least squares minimization problem: $min_X \sum_{k=1}^{K} (w_k.\mathbf{l_k^t}.\mathbf{X})^2$ where w_k is the sum of the segment lengths, $\mathbf{l_k^t} = [cos(\theta_k), sin(\theta_k), -\rho_k]$ and $\mathbf{X^t} = [u_X, v_X, 1]$.

The distance between the $(n-1)$ and the (n) DVP position has to be smaller than the dmax threshold. We compute for each VL candidate the distance to the DVP and the angle difference with a "true" VL which is the line defined by the DVP and the most remote segment midpoint. Only the best candidates are selected to perform another computation loop. The algorithm stops when all the remaining candidates are selected.

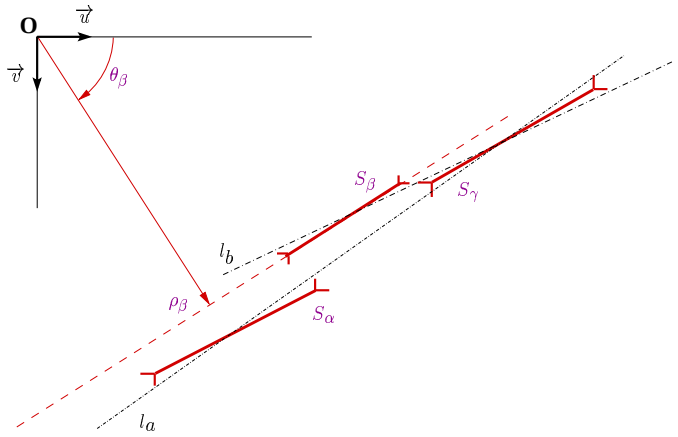

Fig. 2. Constraints on segments colinearity. The 3 segments $(S_\alpha, S_\beta, S_\gamma)$ have similar (ρ_k, θ_k) characteristics under a polar representation and the same contrast direction. They can belong to the same VL only if the segments S_α and S_β do not have any common u-ordinate. The two possible candidates are then the segment lines l_a and l_b.

B. Tracking and matching pencil of vanishing 2D lines

Let us assume two images I_1 and I_2 corresponding to two different positions of the camera. We detect in the two images two sets of VLs, converging to their respective DVP.

We can detect some VL mismatches by respecting an order constraint. For a positive θ, when θ increases, ρ has to decrease. It is the opposite behavior with negative θ. Hence, we can slightly modify an incorrect characteristic by assuming that the DVP estimation is correct.

The information on VL length can also be used to identify a new edge detection, in the particular case of two VLs which have close characteristics. We first consider that the sum of the segment lengths, which form a VL, does not swing in a large scale between two images. Second, a VL whose single segment length is smaller than 50 pixels is certainly an outlier.

Consider now that we succeed in matching two pencils of VLs from image I_1 and image I_2. The cross-ratio and the incidence are actually the only properties left invariant by a projective transformation. Hence, the projections of the 3D lines lying on the road plane onto I_1 and I_2 are related by a projective transformation characterized by the unicity of the cross-ratio (see Figure 3). We use this property for eliminating the VLs which do not belong to the road plane by checking the consistency of the cross ratio :

$$CR(l_a, l_b, l_c, l_d) = \frac{\sin(\alpha_1)}{\sin(\alpha_2)} / \frac{\sin(\alpha_3)}{\sin(\alpha_4)}$$

where α_i with i∈[1,2,3,4] are oriented, signed angles.

The VLs lying on the ground plane project in the camera retinal plane with a particular transformation, called planar homography (induced by a plane). A homography is

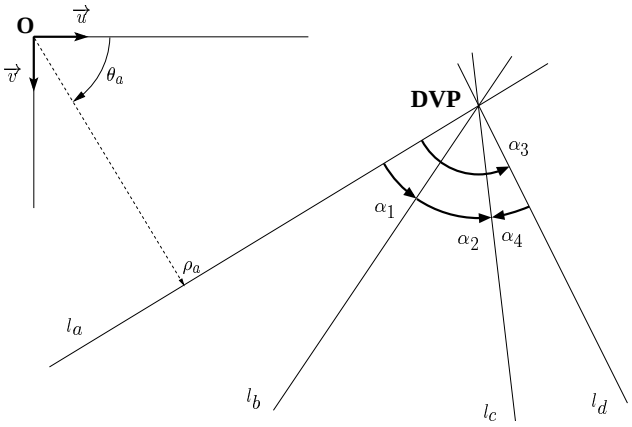

Fig. 3. Cross-ratio of 4 vanishing lines (l_a, l_b, l_c, l_d) lying on the road plane and converging to the DVP.

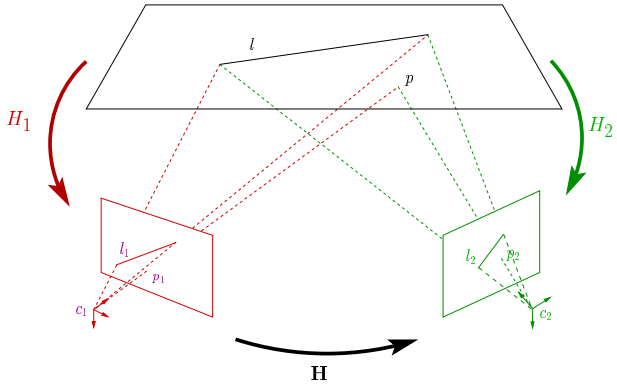

Fig. 4. The planar homography $\mathbf{H}$ between two views is obtained by composition of single view homographies $\mathbf{H_1}$ and $\mathbf{H_2}$. The lines $\mathbf{l_1}$, $\mathbf{l_2}$ and the points $\mathbf{p_1}$, $\mathbf{p_2}$ respectively represent the projections of the line $\mathbf{l}$ and the point $\mathbf{p}$, lying on the ground plane onto the camera retinal planes whose center are $\mathbf{C_1}$ and $\mathbf{C_2}$.

described by a 3*3 matrix $\mathbf{H}$, which has 8 entries: 9-1 of scale factor. Therefore, $\mathbf{H}$ is determined uniquely by solving a linear system of equations containing at least 4 correspondences. The homography contains the translation and rotation motions, up to a scale factor, between the two camera frames $\mathbf{C_1}$ and $\mathbf{C_2}$. We can also compute the displacement between I_1 and I_2 by introducing their respective planar homography H_1 and H_2:

$$\mathbf{l} \propto \mathbf{H_1^t}.\mathbf{l_1},\ \mathbf{l} \propto \mathbf{H_2^t}.\mathbf{l_2} \Rightarrow \mathbf{l_1} \propto [\mathbf{H_1^{-t}}.\mathbf{H_2^t}].\mathbf{l_2}$$

$$\mathbf{l_1} \propto \mathbf{H^t}.\mathbf{l_2}\ with\ \mathbf{H} \propto \mathbf{H_2}.\mathbf{H_1^{-1}}$$

However, the computation of the homography is not possible using only a pencil of lines due to the linear dependency between the lines. All lines of the pencil can in fact be parameterized by only a couple of them: e.g. $\mathbf{l}_\lambda = \mathbf{l_a} + \lambda.\mathbf{l_b}$. We have to match and track two new features which belong to the ground plane. That can be done by using a Harris detector to extract the points of interest $\mathbf{p}$ which belong to the road plane. These points also verify the previous homographies (see Figure 4):

$$\mathbf{p_1} \propto \mathbf{H_1}.\mathbf{p}, \mathbf{p_2} \propto \mathbf{H_2}.\mathbf{p} \Rightarrow \mathbf{p_2} \propto [\mathbf{H_2}.\mathbf{H_1^{-1}}].\mathbf{p_1}$$

$$\mathbf{p_2} \propto \mathbf{H}.\mathbf{p_1}$$

Presently, The VL matching is manually initialized by an operator who selects 4 matched VLs lying on the ground plane, in the first two images. The cross-ratio computed from these 4 VLs is considered as principal. All the remaining VLs of the two images are computed by keeping 3 of 4 principal VLs. The selected VL replaces the closest principal VL. Hence, the cross-ratio computation respects the order constraint. We next match remaining VLs of the two images by comparing their location between the two nearest principal VLs and their cross-ratio result.

If the algorithm detects some principal VL misdetections in a new frame, it replaces them by setting a second group matched VL to the principal group. This is clearly possible only if there are enough VLs in the second group to compensate the loss of some principal VL(s).

C. Tracking and matching ground plane feature points

Considering that we succeed in determining the road boundaries, we can segment the region of the image corresponding to the ground plane. We use the same image derivative operations of the Canny edge detection to compute a Harris feature point detection, according to [12]. Only the points which have an Harris score higher than 90% of the best score are considered correctly detected. The matching method we use is based on singular value decomposition of an appropriate correspondence strength matrix [13].

The image search area is reduced with the road boundaries. We select the correct matching points lying on the ground plane by introducing two image location criteria. We first keep count of the feature points which are very close to or on a VL lying on the ground plane. As we experimented, very few points per image verify this strict condition. Secondly, we eliminate all the detected points located less than 4*dmax from the DVP because these points have a high probability of appearing close to a VL without lying on the ground plane.

The resulting points are generally corners of road lane markers. To be considered as ground plane feature detections, the point coordinates have to be very close to a segment endpoint which belongs to a VL. Furthermore, the VLs have to be matched between the two images. This allows us to detect some mismatching points.

Hence, as soon as the homography computation is validated, we can project all the feature points of the first image into the second one. We then highlight some new feature points lying on the ground plane by computing the distance between the projected and matched coordinates of points. The non-colinearity condition of 3 coplanar points implies that the feature points have to be sufficiently far from each other to improve the homography constraints.

IV. Experiments

We validated our approach using video sequences recorded in the streets of the old city of Antibes and in

the harbor neighborhood. We use an uncalibrated stereo vision system. The speed was about 10 m/s and the frame rate was 25 Hz. The sequences contain more than 1000 black and white images of size 728*410 pixels.

Figure 5 shows a single road video sequence with a large curvature radius which leads to a fork intersection before the end of the curve. The DVP location (orange cross) is tracked with a selection of the best edges (red segments) which are clusterized into VL candidates (red dashed lines). Many outliers, like non-parallel road markers, are detected, but they introduce minor corruptions in the estimation.

Figure 6 presents the DVP(u_x, v_x) positions estimated along the sequence of images. We first note that the trajectories present only two sudden changes of level. The first one corresponds to the lack of kerb in the field of view. The second one is quite unavoidable, it corresponds to the road lane changing. The v-ordinate of the DVP seems stationary; this is expected due to the fixed horizon line in the image.

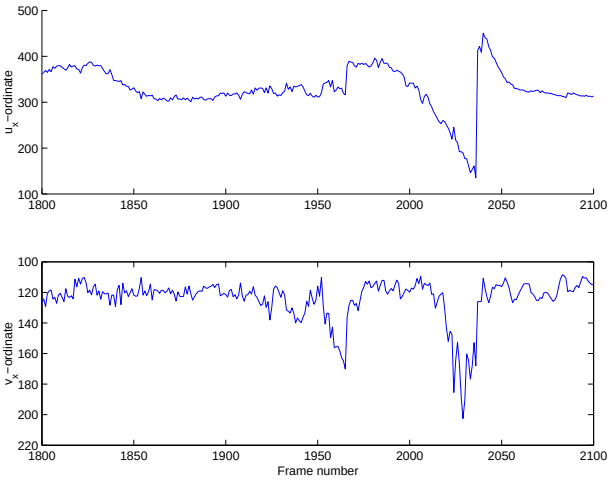

Fig. 6. Chronogram of the estimated DVP coordinates (u_X at the top, v_X at the bottom), along a video sequence recorded on a curving road. The discontinuity in frame 2036 comes from the algorithm detecting a new DVP when the vehicle leaves the main road.

Figure 7 presents the result of a homography estimation between the two images I_1 and I_2. We solve the linear system which has as entries the VLs and the feature points matched on the road plane. As expected, the projection of the VLs can be considered less efficient than the projection of feature points. This is in fact a vision representation bias. A comparison between the characteristics of the features proves that the distance between points is more significant than the distance between lines.

We emphasize that a satisfying homography estimation essentially depends on the conditioning of the system matrix. We first improve the quality of the result by normalizing the two subsystems related to VL and feature point solutions. According to this, we modify the system to obtain normalized singular values.

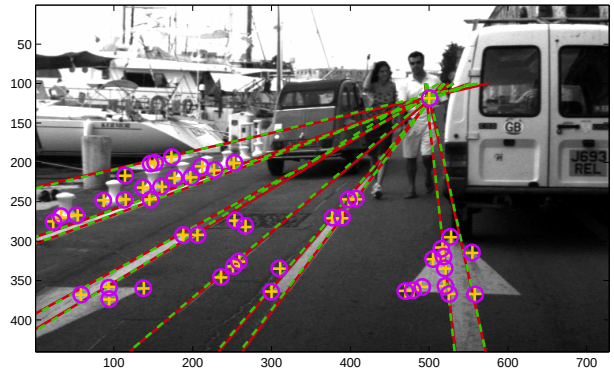

Fig. 7. Re-projection of the features of image I_1 into image I_2, resulting from the estimation of the homography. The original image contains red VL and orange '+' features. Their correspondent projections are green dashed VL and magenta 'o', respectively.

V. Conclusion and Future Work

We presented preliminary results for the estimation of the vehicle motion in an urban environment where a DGPS-based localization is unreliable. We proposed a robust method to detect and track the DVD along a sequence using realistic scene assumptions and projective geometry. Experimental results validate the robustness of the approach.

In the work currently in progress, we try to detect automatically the 4 principal VLs at the start of the sequence. In the same way, we want to improve the image feature detection by introducing dynamic selection criteria to take into account the DVP pose in the image. If need be, we will reduce the jittering of the DVP location with a filtering process.

VI. References

[1] T. Chen, *Development of a Vision-based Positioning System for High Density Area*, Proc. of Asian Conference on Remote Sensing, Hong Kong, China, Nov 22-25 1999.

[2] C. Rother, *A new approach to vanishing point detection in architectural environments*, Proc. of 11th British Machine Vision Conference, pp. 647-655, Bristol, UK, Sept. 11-14, 2000.

[3] J. Kosecka, W. Zhang, *Efficient Computation of Vanishing Points*, Proc. of IEEE ICRA'02, pp. 3321-3327, Washington DC, May 2002.

[4] J.J. Guerrero, C. Sagues, *Uncalibrated vision-based on lines for robot navigation*, Elsevier, Mechatronics 11, no. 6, pp. 759-777, Sept. 2001.

[5] X. Lebegue, J.K. Aggarwal, *Generation of Architectural CAD Models Using a mobile Robot*, Proc. of IEEE ICRA'94, pp. 711-717, San-Diego, CA, USA, 8-13 May 1994.

[6] M. Snaith, D. Lee, P. Probert, , Image and Vision Computing Journal, Volume 16, No 4, pp. 251-263, 1998.

Fig. 5. Top-left, bottom-right, every 20 frames: results of continuous tracking of the DVP (orange '+') over a video sequence. The DVP is the weighted least square best estimation of intersection of all the VL candidates (red dashed lines).

[7] M. E. Antone, S. Teller, *Automatic Recovery of Relative Camera Rotations for Urban Scenes*, Proc. of IEEE CVPR'00, pp. 282-289, Head Island, SC, USA,13-15 June 2000.

[8] R. Wang, Y. Xu, Y. Zhao, *A Vision-based Road Edge Detection Algorithm*, Proc. of of IEEE Workshop of Applications of Computer Vision, pp. 237-241, Orlando, Florida, USA, Dec. 3-4, 2002.

[9] M. A. Sotelo, F.J. Rodriguez, L. Magdalena, *vision-based Navigation System for Autonomous Urban Transport Vehicles in Outdoor environments*, Proc. of IV'02, Versailles, France, June 17-21, 2002.

[10] M. Okutomi, K. Nakano, J. Maruyama, T. Hara, *Robust Estimation of Planar Regions for Visual Navigation Using sequential Stereo Images*, Proc. of IEEE ICRA'02, pp. 3321-3327, Washington DC, May 2002.

[11] Z. Hu, K. Uchimura, *Dynamical Road Modeling and Matching for Direct Visual Navigation*, Proc. of IEEE Workshop of Applications of Computer Vision, pp. 237-241, Orlando, Florida, USA, Dec. 3-4, 2002.

[12] C. Schmidt, *Appariement d'images par invariants locaux de niveaux de gris*, PHD document, Grenoble, France, Jul.2nd, 1996.

[13] M. Pilu, *A direct method for stereo correspondence based on singular value decomposition*, Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 261-266, San Juan, Puerto Rico, June 17-19, 1997.