MOBILE ROBOT NAVIGATION BY AN ACTIVE CONTROL OF THE VISION SYSTEM

Patrick Stelmaszyk*, Hiroshi Ishiguro, Saburo Tsuji****

*ITMI, Chemin des prés, ZIRST 38243 Meylan, France

**OSAKA UNIVERSITY, Dept of Control Engineering, Toyonaka, Osaka 560, Japan

ABSTRACT

In this paper, we argue that a mobile robot's environment can be determined by computing local maps surrounding feature points, called fixation points. These fixation points are obtained by searching the scene for points which present some interesting cue for robot navigation. This 3-D computation is based on a monocular active vision system composed of a camera, mounted on a rotating table accurately controlled by a computer, which gazes the fixation point as the robot moves. The system then computes the local map and updates it with each new observation in order to increase its accuracy and robustness. Real experimentation in a complex indoor scene illustrates that the 3-D scene coordinates can be obtained with a good accuracy by integrating several observations.

1 Introduction.

Robot navigation in a simple environment can be achieved through a two dimensional analysis in which a single camera is used to find free space and avoiding obstacles [Kuhnert 86]. Nevertheless, as soon the robot environment becomes too complex, such a system fails and the computation of 3-D structure seems necessary. Usually, the determination of this 3-D structure in mobile robotics, is based on stereo or similar techniques which reconstruct the robot's complete environment [Tsuji 86], [Ayache 87], [Brooks 88]. While such a 3-D representation can be extremely useful for some recognition or inspection tasks, the huge amount of features used for 3-D reconstruction requires complex and time consuming computations.

The idea the authors want to stress in this paper, is that the robot navigation in a complex environment can be efficiently performed by computing not the whole structure of the environment, but only the structure surrounding a few selected feature points. Navigation can then be considered as a goal-oriented task in which each goal represents the area surrounding the feature point detected in the scene. This paper, concerns the determination of the surrounding 3-D structure based on a new concept of robotics called active vision. Aloimonos [Aloimonos 87] has presented theoretical works for estimating structure from motion, shading and contour using a camera which gazes a feature point (called "fixation point") at its focal center. Ballard [Ballard 88, Ballard 89], has developed the animated vision system which controls the vergence of a pair of cameras in order to keep the projection of the fixation point at the center of both the image planes. The 3-D structure of surrounding points is then computed. Recently, Sandini [Sandini 90] also proposed a technique for determining 3-D information on the basis of constraints imposed by the active motion.

The concept of such an approach can be extended for estimating the local 3-D structure in the vicinity of fixation points for mobile robot navigation. Our active visual system is composed of a camera mounted on a rotating table accurately controlled by a computer which gazes a fixation point while the robot moves. During the robot motion, assumed to be composed of either a pure translation or pure rotation centered on a fixation point, the projection of the fixation point is kept on the camera optical center. The 3-D structure of features surrounded this point is then computed. This local map is updated with each new frame by combining new data to improve accuracy and robustness. As soon this analysis is achieved and no more information is expected, a new fixation point is processed. Figure 1.1 represents a robot navigation scenario in which the robot looks respectively at point A and B before rotating around point C and carrying on by looking at next fixation points of the scene.

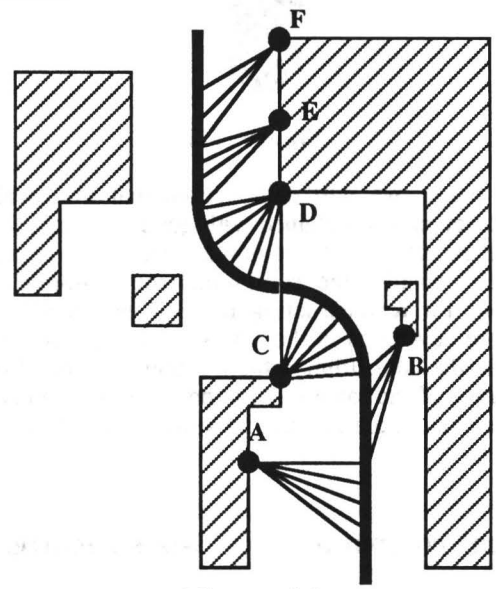

Figure 1.1

The selection of fixation points requires a lot of attention. Such a point must either provide a cue for robot navigation

or represent a potential obstacle. With respect to these 2 constraints, we consider that a fixation point has to be selected in the scene with respect to geometric and photometric properties. Geometrical properties consist in the detection of areas with respect to the robot's distance. Each such an area is classified for determining the processing order. The closest area will be processed first and, after a while, the second nearest one and so on. The photometric properties, based on gradient, texture and color analysis, consist in pointing out the fixation point, inside the area, which allows an efficient, robust and easy detection under different angle of view or different illumination.

For achieving the fixation point selection with respect to the above considerations, a static camera is added to our robot (see figure 1.2). Such a camera, equipped with a wide angle lens, allows the perception of all scene points in the image (the active camera's field of view is constrained to a small scene area by the short range lens). The 3-D structure of all these scene points (included the scale factor which corresponds to the distance to the fixation point) can be roughly estimated by a motion stereo technique.

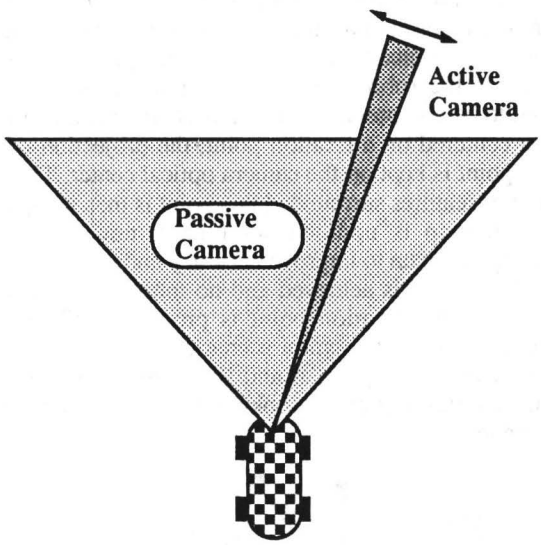

Figure 1.2

In this paper, we will demonstrate that the local map computation can be obtained with good accuracy. We will suppose the robot does not stop at each image acquisition and then consider the continuous robot motion. Basic equations for reconstructing the 3-D position of scene points are developed in the next section. Section 3 briefly presents the algorithm which integrates the different observation of the same fixation point from different robot position. Real experimentation validate the technique in section 4.

2 Local computation of 3-D structure.

Our method is based on the assumption that the robot moves either along linear or circular paths centered on feature points. An active camera is mounted on a rotating table, and the 3-D structure of scene points are computed up to a scale factor. The table controller uses visual feedback for keeping the fixation point at the image center while the robot moves. Only the 3-D structure of vertical edge lines is computed due to the lack of a three axis rotation controller in our experimentations. Such a limitation requires us to estimate scene points location in a 2-D top view representation. Although such information is sufficient in indoor scene environments composed of many vertical lines, an extension of the equations developed in this paper to the general case is straightforward.

Let us assume a coordinate system X, Y, Z attached with respect to the camera, with the Z axis pointing along the optical axis. Let $T = (U, V, W)^T$ be the translational component of the camera motion is and $R = (A, B, C)^T$ be the rotational component . Let $P = (X, Y, Z)^T$ be the instantaneous coordinates of some point in the environment. The perpendicular component u of the instantaneous velocity is expressed by [Bruss 83] :

$$u = \frac{fX'}{Z} - \frac{fXZ'}{Z^2}$$
$$= f\left(-\frac{U}{Z} - B + \frac{Cy}{f}\right) - x\left(-\frac{W}{Z} - \frac{Ay}{f} + \frac{x}{f}\right) \qquad (1)$$

in which f is the camera focal length and (x, y) is P's projection in the image plane.

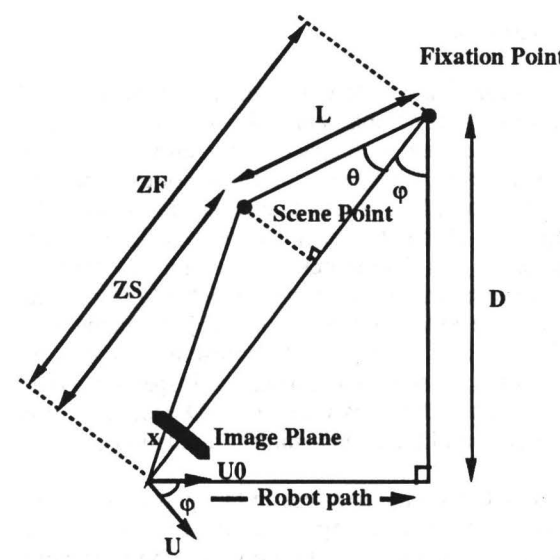

Figure 2.1

Let us assume that the robot moves linearly. By controlling the panning of the camera, the fixation point can be made to project always to the optical image center. Location of scene points relative to the fixation point are estimated while the robot moves. The 3-D structure computation is based on an intermediate cylindrical representation (θ and L) which simplifies the presentation. As indicated on figure 2.1, L is the distance between the fixation point and the object point and θ the angle formed by the center of projection, the fixation point and the object point. As explained before, we consider a top view and do not deal

with the height of scene points. Since the robot moves on a flat floor the camera motion parameters are expressed by the following vectors: $R = (0, \omega, 0)^T$ and $T = (U_0\cos\varphi, 0, U_0\sin\varphi)^T$ in which U_0 represents the linear velocity of the robot, ω the camera angular velocity and φ the angle between the camera axis and a perpendicular to the robot path. The general equation 1 can then be written as :

$$u = -\frac{U_0\,(x\sin\varphi - f\cos\varphi)}{Z} - \omega(f + \frac{x^2}{f}) \qquad (2)$$

Since the fixation point is fixed at the image center while the robot moves, both optical flow and position are equal to zero. From equation (2), we can compute the distance ZF.

$$ZF = -\frac{U_0\cos\varphi}{\omega}$$

The distance ZS, for some other point in the field of view of the camera, is derived from equation (2):

$$ZS = \frac{U_0\,(x\sin\varphi - f\cos\varphi)}{(u + f\omega + \omega\frac{x^2}{f})}$$

and the computation of the ratio ZS/ZF, noted β, allows the elimination of the robot velocity Uo :

$$\frac{ZS}{ZF} = \beta = \frac{f - x\,tg\varphi}{(\frac{u}{\omega} + f + \frac{x^2}{f})}$$

The scene points in cylindrical coordinates are obtained by basic geometric relation in which D, as explain later, represents the scale factor:

$$\theta = tg^{-1}\left(\frac{x}{f}\frac{\beta}{1-\beta}\right) \qquad (3)$$

$$L = \frac{1-\beta}{D\cos\theta}\cos\varphi \qquad (4)$$

As mentioned before, the 3-D structure of scene points is estimated relative to the fixation point. If some external sensors can be used for accurately determining the distance between the fixation point and the camera (Sandini 90), we prefer to consider that all scene point coordinates are expressed up to a scale factor represented by D in equation (4). Although such a scale factor limits the interest of our technique, it can be estimated with sufficient accuracy using motion stereo.

In the case of rotational motion, it has been demonstrated (Ishiguro 90) that we obtain the same equation for θ, while the angle φ is set to zero when computing L. But in both the cases, the equation (3) cannot be computed when $\beta = 1$. This situation occurs when the scene point is located along a line parallel to the image plane and passing through the fixation point. However, the determination of the distance L can be directly computed by using basic trigonometric relation ($x/f = L/ZF$) and the sign of the angle θ, which is equal in such a case to $\pi/2$, is provided directly by looking at the sign of x.

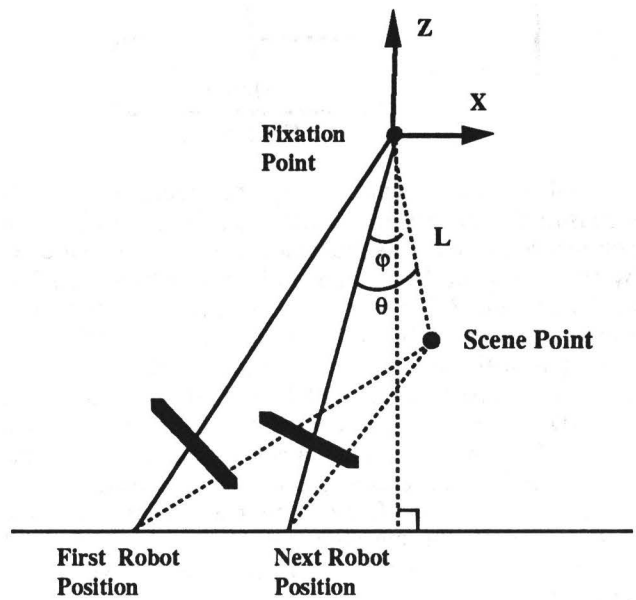

Figure 2.2

While moving, the 3-D structure of all visible scene points are expressed in the fixation point centered coordinate system. By looking at figure 2.2 which represents the robot in two consecutive positions, the Cartesian coordinates can be deduced by basic trigonometric relations :

$$X = L\cos(\theta - \varphi - \frac{\pi}{2}) \qquad Z = L\sin(\theta - \varphi - \frac{\pi}{2})$$

3 Integration of the 3-D observations.

Even if the accuracy is rather good in the fixation point vicinity, the error is large enough to justify the use of techniques for decreasing it. In this section, we will briefly present how the different observations are combined and merged.

The integration process is illustrated on figure 3.1. Tokens (vertical edges in our application) tracked in consecutive images are reconstructed in the fixation point coordinates system and considered to us as a 3-D observation. The first 3-D observation is assigned into to the model and then updated at each new observation. Each 3-D observation is composed of the 3-D coordinates of each scene point, as far

as its uncertainty, represented by a covariance matrix and its label number.

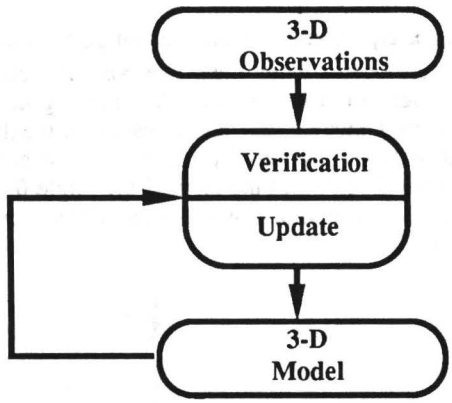

Figure 3.1

If a false match occurs during the tracking, the 3-D reconstruction process may produce a 3-D observation with non-realistic value. This problem is overcame, in our case, by computing the Mahanobolis distance between the 3-D model and the 3-D observation. A distance less than 1 means that the probability that the token corresponds to the same physical entity is greater than 52 % in the two-dimensional case [Ramparany 89]. The 3-D observation and the 3-D model are then merged based on an Extended Kalman Filter which provides a new estimation of both the position and the uncertainty. Each 3-D observation being expressed in the same frame coordinates (centered on the fixation point), the merging process can be performed without any geometrical transformation between two different observations.

Each 3-D element of the model also contains two factors. The first one is a confidence factor which represents the story of the sequence. This factor is incremented each time a match between the observation and the model is performed. A high value then indicates the element is present in the scene with a great confidence. The second factor is intended to deal with temporal perturbations (like vibrations, noise, shadow ...) which affect the computation of the 3-D element. If the matching is done, this factor is incremented. Otherwise, it is decrement and removed from the model if its value falls beneath a threshold and the confidence factor is not high enough to make sure that this element really exists in the scene. The final local map will be only composed with high confidence factor elements.

4 Results.

The following experimental assessments have been performed in the case of linear robot motion. A camera is mounted on a rotating platform which swivels with a step of 0.1 degree. During the robot displacement, the platform is controlled so that the camera tracks a fixation point. The first and last images of a sequence are shown in figure 4.1, the robot being successively located at the positions labeled 1 and 17. Between each observation, the robot moves 5 cms.

The equations developed in section 2, give the 3-D structure in the continuous case which assumes the robot doesn't stop between two acquisition. Such an assumption requires special hardware for processing data in real time which was not available for these experiments. Nevertheless, in the following, the continuous equations have been used for validating the approach although the robot displacement is expressed in term of distance instead of velocity.

All image edge segments are tracked in the consecutive images of the sequence. Each token contains a label number which allows the determination of the corresponding edge lines in two images separated by a large distance [Crowley 88, Stelmaszyk 89]. In our application, the distance correspond to 5 image frames (around 25 cms). For each pair of images (1&6, 2&7, ...n&n+5), we perform the 3D reconstruction given in section 2. Each 3-D reconstruction corresponds to the same physical scene viewed by a different point but represented in the same fixation point centered coordinates.

The figure 4.2 indicates the front view (left side) and the top view (right side) of the reconstructed scene when the robot has integrated twelve 3-D observations (images 1&6, 2&7 ... 12&17). The front view indicates the positions of segments which have been tracked at least 3 times in the different images (confidence factor = 3). We can check that the fixation point is always located on the image center. In the top view, the fixation point is represented by the intersection of the vertical and horizontal axes and the robot trajectory is given at the bottom. The lines joining the fixation point and the robot trajectory show the robot position and the camera angle during different acquisitions. Scene points are shown as crosses followed by their label number. Looking at such a label, one can find the corresponding segment in the front view, place it into the raw image 4.1 and check roughly the matching validity. Nevertheless, the front view must be seen by positioning ourselves at the robot position. For instance, the point labeled 4, located on the upper right part of the top view, is effectively located on the left of the fixation point if we look at it from the robot's position.

Linear Motion							
D = 104 cms, U = 5 cms/frame, f = 591 pixels, t = 5 frames							
Point (lab)	X meas (cms)	Z meas (cms)	L meas (cms)	Conf Facto r	X real (cms)	Z real (cms)	Error (cms)
54	95	-42	141	4	89.5	-42.6	5.5
4	0	137	137	5	14.4	140.5	2
44	95	13	96	3	93.5	12.6	1.5
5	0	95	95	9	0.7	95.7	1
48	68	-16	70	3	59.7	-22.5	10.5
43	-2	-40	40	11	-3.6	-39.7	1.62
18	0	27	27	5	0.8	25.3	1.8
33	26	0	26	5	26.3	1.4	1.4

Table 4.1

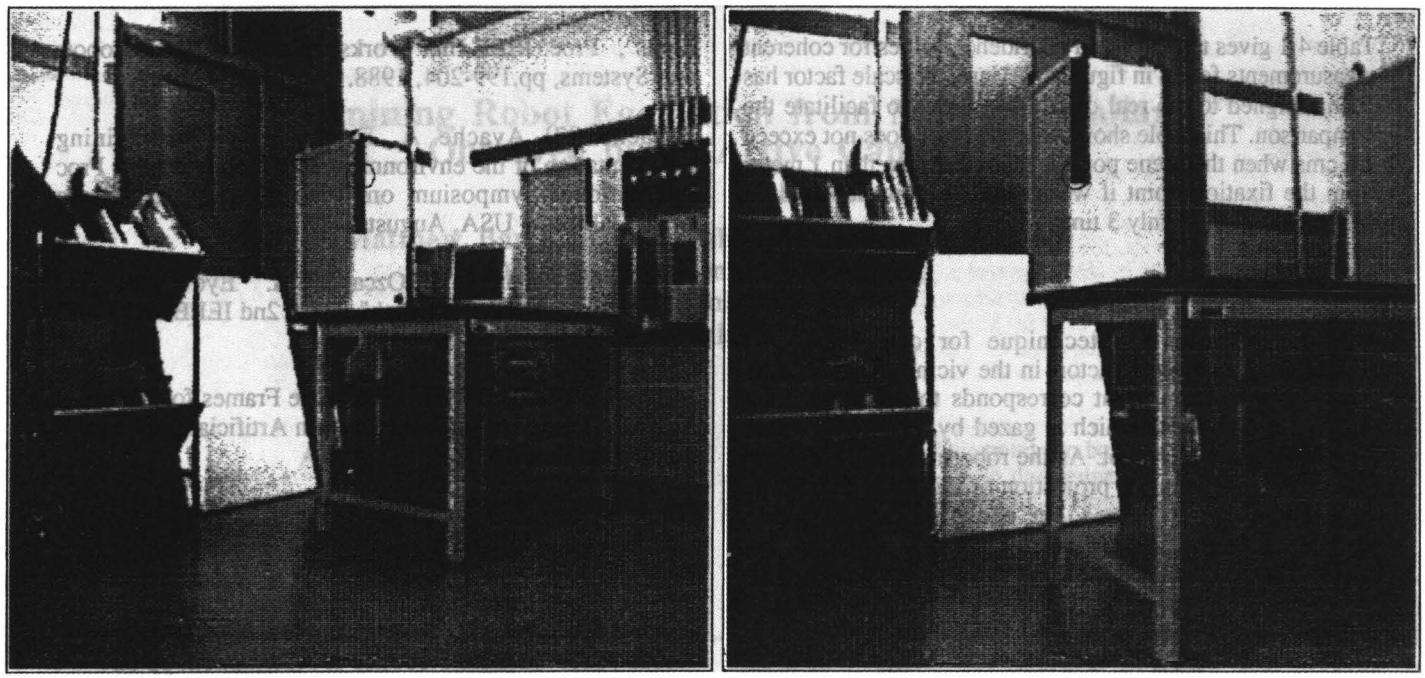

Figure 4.1

First and last raw images of the sequence. The fixation point is surrounded by the circle.

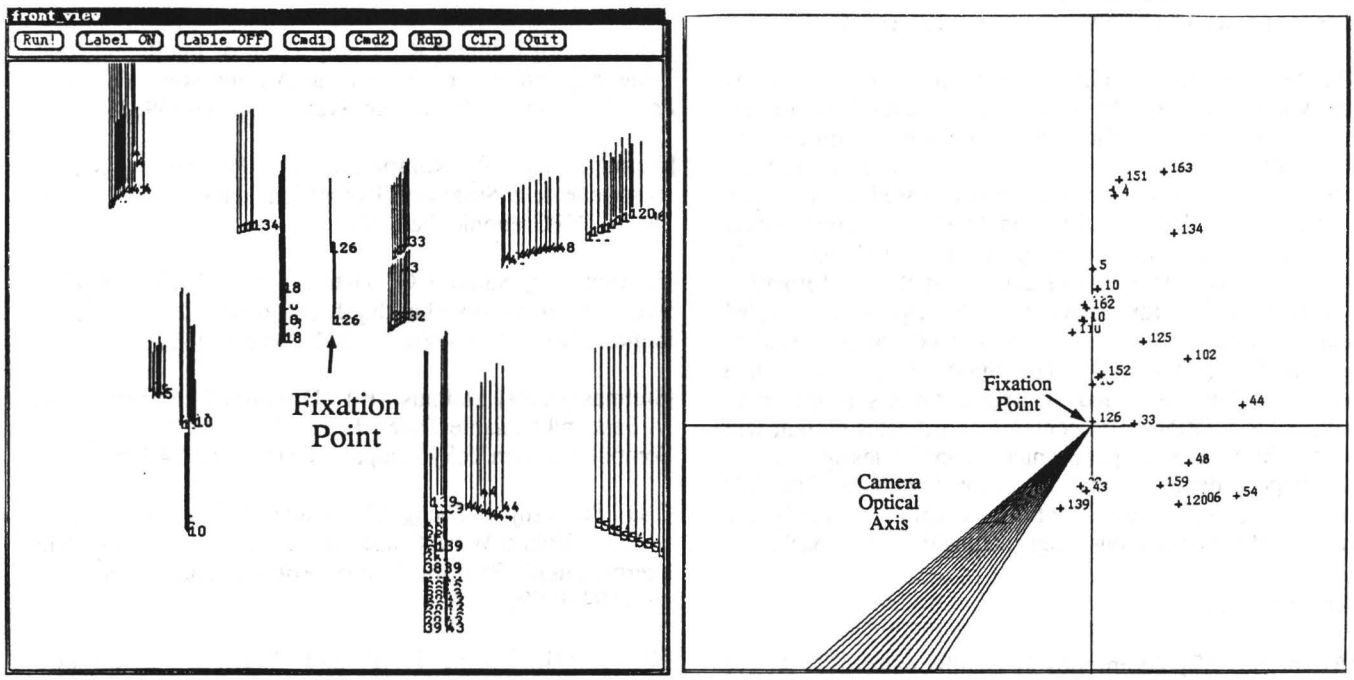

Figure 4.2

Left side: All vertical edges matched in the sequence.
Right side: Top scene view.

Table 4.1 gives the error and confidence values for coherent measurements found in figure 4.2. Here, the scale factor has been assigned to the real distance in order to facilitate the comparison. This table shows that the error does not exceed 1.8 cms when the scene point is located at less than 1 meter from the fixation point if we ignore the point labelled 48 which was merged only 3 times.

5 Conclusion.

This paper presents a technique for computing 3-D structure, up to a scale factor, in the vicinity of a fixation point. This fixation point corresponds to a feature point detected in the scene which is gazed by an active camera located on a mobile robot. As the robot moves, the camera rotates so as to keep the projection of this fixation point at the camera center. Instead of using the robot motion parameters in the computation of the 3-D structure, the authors have introduced the camera angular velocity. This information is measured accurately by a precise shaft encoder.

By taking into account the error attached to each measurement, we combine different observations so as to increase both accuracy and robustness. The integration process is facilitated by the use of active vision which expresses the 3-D coordinates in an object centered coordinates system. Such a representation is invariant with respect to the robot motion and merging two different robot position doesn't require any geometrical transformation.

Experimentations in a real and complex environment demonstrate that the error of measurements is less than 1.8 cms for points which have been merged at least five times.

This experimentation validates the technique and the authors are working now on the integration of several local maps into a global one. Such an integration requires the automatic selection of the fixation points as far as their positions. A motion stereo technique, based on the static camera mounted on the robot, should allow the computation of both the scale factor and the fixation point's position in the global map. Although inaccurate, such an information will be sufficient for allowing the description of a global map represented in a path centered coordinates system [Asada 88] [Zheng 90]. The main feature of such a representation it attached the origin directly to the robot path. Each fixation point centered coordinates system will be then inserted in the path centered coordinates system. We also hope to develop a real time implementation of the 3-D structure computation equations in order to verify the validity of the continuous image acquisition's approach.

References.

[Aloimonos 87] Aloimonos J, Bandyopadhyay. "Active Vision", Proc Image Understanding. Workshop, pp 552-573, 1987.

[Asada 89] Asada M, Fukui Y, Tsuji S. "Representing Global World of A Mobile Robot with Relational Local Maps", Proc. IEEE Int'l Workshop on Intelligent Robots and Systems, pp.199-204, 1988.

[Ayache 87] Ayache A, Faugeras O. "Maintaining representation of the environment of a mobile robot", Proc International Symposium on Robotics Research, Santa Cruz, California USA, August 1987.

[Ballard 88] Ballard D H, Ozcandari A. "Eye Fixation and Eearly Vision : Kinetic Depth", Proc. 2nd IEEE Int'l Conf. Computer Vision, pp. 524-531, 1988.

[Ballard 89] Ballard D H. "Reference Frames for Animated Vision", Proc 11th Int Joint Conf on Artificial Intelligence, pp1635-1641, 1989.

[Bruss 83] Bruss A R, Horn B K P. "Passive Navigation", Computer Vision Graphics and Image Processing, vol. 21, pp. 3-20, 1983.

[Brooks 88] Brooks R. "Visual Map Making for Mobile Robot", Proc Int Conf Robotics & Automation, pp 824-829, 1985.

[Crowley 88] Crowley J L, Stelmaszyk P, Discours C. "Measuring image flow by tracking edge-lines". Second Int Conf on Computer Vision (ICCV). Dec 1988, USA.

[Ishiguro 90] Ishiguro H, Stelmaszyk P, Tsuji S. "Acquiring 3-D Structure by Controlling Visual Attention of a Mobile Robot", IEEE Intern Conference on Robotics and Automation.May 13-18 1990, The Hyatt Regency Cincinnati, Cincinati Ohio, USA .

[Kuhnert 86] Kuhnert K. "Comparison of Intelligent Real Time Algorithm for Guiding an Autonomous Vehicle", Proc Autonomous Intelligent System, pp334-339, 1986.

[Ramaparany 89] Ramparany F. "Perception Multi-sensorielle de la Structure Geometrique d'une scene", PHD-Thesis, INPGrenoble, Feb 1989.

[Sandini 90] Sandini G, Tistarelli M. "Active tracking Strategy for Monocular depth Inference Over Multiple frames", IEEE PAMI Vol 12, N 1, January 1990.

[Stelmaszyk 88] Stelmaszyk P, Discours C, Chehikian A. "A Fast and Reliable Token Tracker", IAPR Workshop on Computer Vision.Tokyo (Japan), October 12-14 1988.

[Tsuji 86] Tsuji s, Zheng J.Y, Asada M. "Stereo Vision of a Mobile Robot: Word Constraints for Image Matching and Interpretation", Proc Int Conf on Robotics and Automation, pp: 1594-1599, 1986.

[Zheng 90] Zheng J. Y and Tsuji s. "panoramic Representation of Scenes for Route Understanding", Proc 10th Int'l Conf on Pattern recognition, 1990.