17

A Self Navigation Technique using Stereovision Analysis

Sergio Nogueira[1], Yassine Ruichek[2] and François Charpillet[3]
[1,2] *Systems and Transportation laboratory, University of Technology of Belfort Montbéliard,*
[3] *LORIA Laboratory INRIA Lorraine*
France

1. Introduction

One of the most common issues in developing intelligent vehicle concepts is self localization and autonomous navigation. In this chapter, we are particularly interested in global localization of urban autonomous vehicles. To get a global position of a vehicle navigating in urban areas, localization systems must be adapted to different kinds of environments in presence of dynamics objects. In order to achieve global localization, there are two approaches. The first one is based on sensors data fusion. The second approach uses knowledge on the environment of navigation.

In most data fusion based methods, the environment is represented using a GIS (Geographic Information System). The environment representation is augmented by introducing structured elements from sensors. In [1], Chen proposes a method that estimates the camera movements using edge matching. The initial position is given by fusing data coming from a D-GPS (Differential Global Positioning System), a GIS and other sensors. Kais and al. [2] propose to fuse vertical features recorded into a GIS with images provided by an embedded camera. The principle is to construct regions around vertical features in order to compute the position and the orientation of the camera. The localization process is achieved by determining correspondences between virtual and real features.

The environment knowledge based approach is interesting when sensors unreliability is considered. In spite of GPS sensors can be accurate (for example by using RTK-GPS systems) GPS information is not adapted into dense urban areas. Indeed, because of the difficulty to detect satellites (due to the presence of buildings) and reflection of GPS signals, the GPS system may lose in his accuracy and may even provide false positions. This situation is known by the urban cayonning problem [3].

In order to discard this problem, the environment knowledge approach consists in creating an image key database. The camera position and orientation are computed for each image referenced during the learning phase. In [4], Katsura et al. propose a method, which adds a region analysis to distinguish different region types. The aim is to process specifically each region that evolves differently through the time. Based also on image key learning sequence, the method proposed by Royer and al. [5] computes 3D specific points. Bundle adjustments [6] are used in order to increase the model precision. For each movement of the camera, the position is determined using the closest 3D features.

Source: Stereo Vision, Book edited by: Dr. Asim Bhatti,
ISBN 978-953-7619-22-0, pp. 372, November 2008, I-Tech, Vienna, Austria

Between these two main approaches, there are some hybrid localization methods. For vehicle navigation in urban areas, Gerogiev and Allen propose a technique, which consists in fusing data coming from several sensors [7]. When the sensors information is unavailable, a camera makes a visual servoing between the images provided by the camera and the images coming from a geo-referenced image database.

In this chapter, the presented approach is based on image key learning sequence. The aim is to achieve self localization using only stereoscopic information. As Royer and al. in [5], the proposed stereovision based method constructs a 3D model. In the approach, the 3D model is built from 3D features reconstructed using SIFT based stereo matching and tracked using temporal matching.

This chapter is organized as follows. Section 2 presents the concept and fundament to construct a 3D model based on image SIFT features. Section 3 explains how to localize a camera using a database containing 3D points, reconstructed from the matching of SIFT features. Before concluding, results in various conditions are presented in section 4.

2. Model construction

In order to compute the global position of a camera, one needs to construct a 3D model composed by features that can be matched under image changes like translation, rotation and scaling. In external conditions like in urban environments, it is important that image features are partially invariant to illumination changes. Most of the existing approaches use Harris corner detector [8], which is sensitive to the scale of images. As a consequence, building a map requires a lot of images and a considerable number of points extracted in each image. Moreover, in the localization process, the vehicle exploration must be close to the learned trajectory.

In the proposed method, Scale Invariant Feature Transform (SIFT) features are used. Introduced by Lowe [9], SIFT features have detailed characteristics that make them suitable landmarks for robust Self Localization And Mapping (SLAM), because when mobile robots are moving in an environment, landmarks are observed from different angles, distances and under different illumination changes. Each reconstruction or localization step is based on the matching process of SIFT features.

Figure 1 shows an example of SIFT features extracted from the left and right images taken by a virtual simulated stereoscopic sensor with two cameras. The resolution of the images is 640x480. In this example, eight levels of scale are used to extract SIFT features. There are about 2000 features in each image. This quantity of features is generally sufficient for the considered task, but if desired, the number can be increased by increasing the scales and the image resolution.

At each reconstruction step-frame, the SIFT features are extracted from the two rectified stereo images and are then matched. The images are rectified using the Zhang method for stereo camera calibration [10]. Matched SIFT features are stable and serve as better landmarks for detecting and tracking the objects observed in the environment. The stereo matching of the SIFT features provides 3D points that serve to construct the 3D model.

2.1 Stereo matching

Considering the stereovision sensor, the right camera serves as the reference camera. The two cameras are separated by a distance $E=12cm$ and have the same focal length f (see Figure 2). The stereo matching of the SIFT features uses the following constraints:

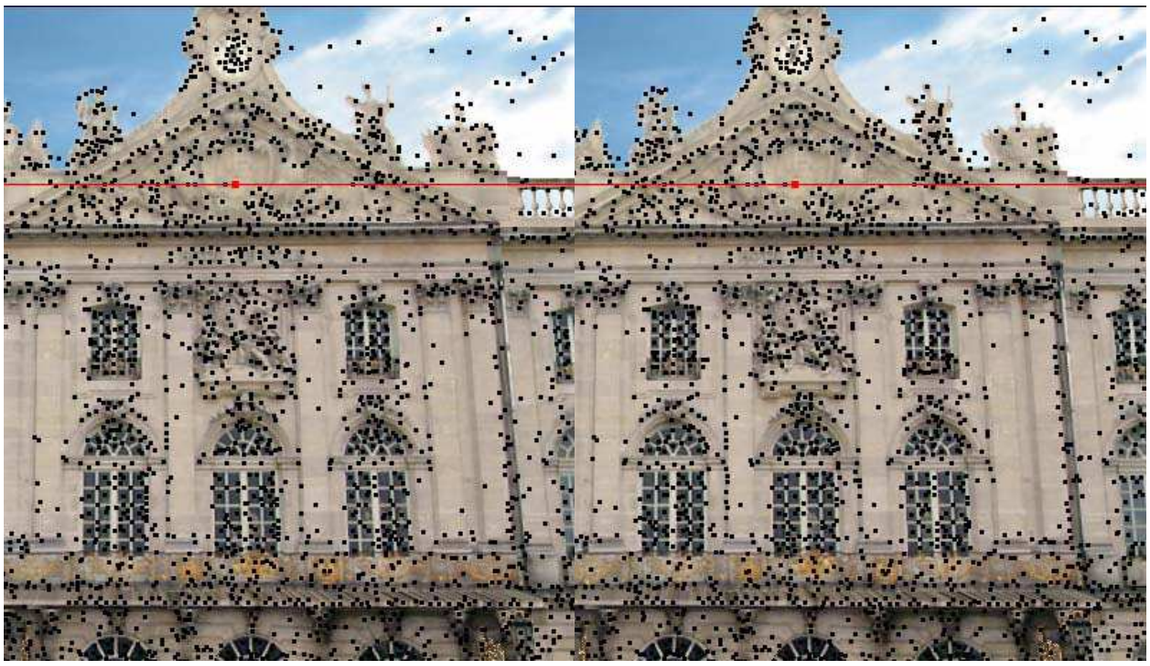


Fig. 1. SIFT features extraction

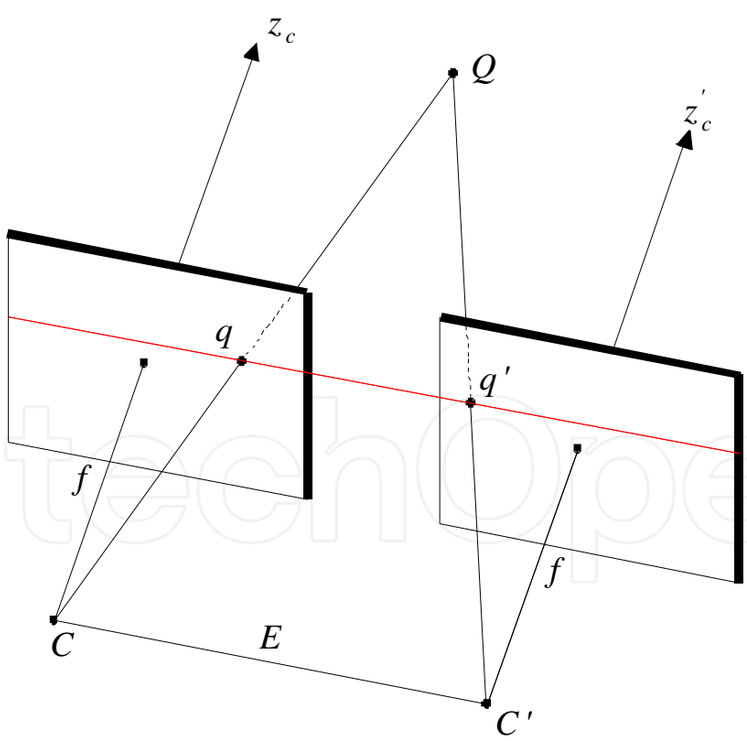


Fig. 2. Stereoscopic configuration

Disparity constraint. Considering a candidate match composed with a left feature and a right feature, the difference between the horizontal image coordinates of the features must be within a predefined disparity range.

Epipolar constraint. The vertical image coordinates of the left and right features must be within 1 pixel of each other, as the images are aligned and rectified.
Orientation constraint. The difference of the orientations of the left and right features must be within a predefined range.
Scale constraint. One scale must be at most one level higher or lower than the other. Adjacent scales differ by a factor of 1.5 in the proposed SIFT extraction procedure.
Uniqueness constraint. If a feature has more than one match satisfying the above constraints, the matches are considered as ambiguous and are discarded so that the resulting matches are more consistent and reliable.
After matching the SIFT features, a list of matched couples is obtained. For each couple, an image horizontal disparity is computed and then a 3D point is reconstructed by considering the intrinsic and extrinsic parameters of the cameras. The 3D point has its coordinates in a reference associated to the stereovision sensor. All the reconstructed 3D points are added to the database after computing their global position (see section 3). In order to use them as landmarks, the orientations and scales of the corresponding SIFT features are set respectively to the average of the orientations and the scales computed in the left and right images.

2.2 Model computation

Let's define an axis-aligned stereoscope *S* where the left and the right cameras are defined respectively by their optical centers *C* and C′, with the same focal length *f*, and separated with a distance *E* (cf. figure 2). Let $W = \{W_x, W_y, W_z\}$ be a 3D point from the world coordinate system, $q = \{q_x, q_y\}$ and $q' = \{q'_x, q'_y\}$ are the projections of the point *W* into the left and right cameras, respectively. Using the stereo triangulation technique, the point *W* can be computed given the image projections *q* and *q′* and knowing the intrinsic and extrinsic parameters:

$$W_x = \frac{W_z * q'_x}{f}, W_y = \frac{W_z * q_y}{f}, W_z = \frac{f * E}{|q_x - q'_x|} \tag{1}$$

Figure 3 shows the projection on the left and right images of the reconstructed 3D points after the stereo matching process.

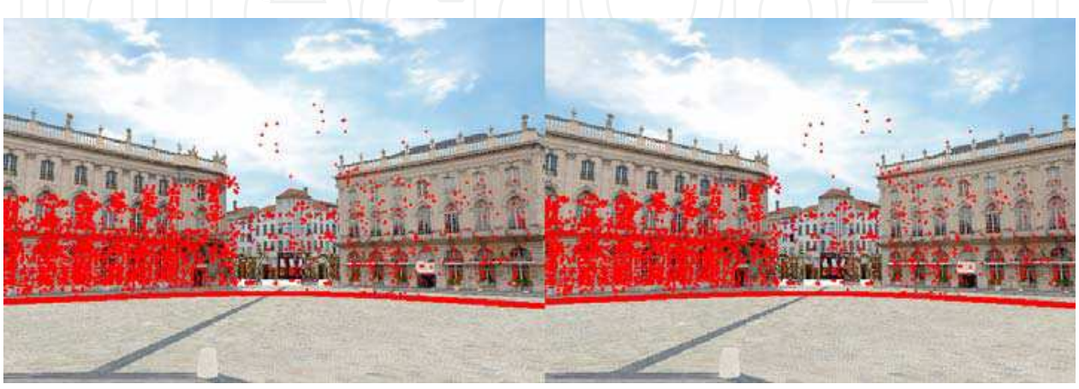


Fig. 3. Projection of the reconstructed 3D points on the left and right images

2.3 3D model construction

The global 3D construction is based on an incremental method. At the beginning, the first stereoscopic acquisition is used to initialize the 3D reconstruction process.

Let's $C_1, C_2 \ldots C_N$, be the camera positions computed respectively from the pose 1 to the pose $N\left(N \geq 3\right)$. The goal is to determine the position C_{N+1} of the camera, associated to the pose *N+1*. After image acquisition I_{N+1} from the unknown position C_{N+1}, the SIFT features are extracted. Let q_{N+1}^i be the ith extracted SIFT feature from the image I_{N+1}. The extracted points are then matched with the SIFT points extracted from the image I_N. A list of couples of matched points $\left(q_N^i, q_{N+1}^i\right)$ is finally obtained. Note that the point q_{N+1}^i may have any corresponding point in the image I_{N-1}. This means that from one image to the other, identified points may disappear and new points may appear. The 3D global position of the new points is determined by using a robust construction method described in the next section.

The main problem of the path reconstruction with temporal matching is that the estimation of each point position is based on the previous computed ones. Consequently, the computation errors increase throughout the reconstruction process. The bundle adjustment process [6] limits the error computation. This process increases the construction precision by using multiple views. It consists on a minimization process based on the Levenberg-Marquardt algorithm [11]. The function $f\left(C_E^1, \ldots, C_E^N, Q^1, \ldots Q^M\right)$ to be minimized is defined from the extrinsic cameras parameters C_E^i and the 3D points $\{Q^j\}$ extracted from stereo 3D reconstruction using multiple views. This function is expressed as follows:

$$f\left(C_E^1, \ldots, C_E^N, Q^1, \ldots Q^M\right) = \sum_{i=1}^{N} \sum_{j=1}^{M} \left\| q_i^j - \pi\left(P_i Q^j\right) \right\|^2 \tag{2}$$

Where $\left\| q_i^j - \pi\left(P_i Q^j\right) \right\|^2$ is the square Euclidian distance between the SIFT point q_i^j and the point $\pi\left(P_i Q^j\right)$, which is the projection of the 3D point Q^j on the camera at the ith position, called also the retro-projection of the point Q^j. P_i is the projection matrix obtained from the extrinsic $\left(C_E^i\right)$ and intrinsic parameters of the camera at the ith position.

In order to make the algorithm more robust, false matches are removed. The minimization process begins by keeping the points satisfying the condition $\left\| q_i^j - \pi\left(P_i Q^j\right) \right\|^2 < \varepsilon$, where ε is a constant fixed empirically to 9. The minimization algorithm converges when the number of the retro-projection points becomes stable.

Figure 4 shows an example of an environment 3D model construction. This figure represents a model view in two different render types. In the middle of the right sub-image, the circular shape represents the vehicle trajectory. The other shapes around the trajectory represent the 3D reconstructed points, corresponding to the environment 3D model.

3. Localization

After completing the 3D model construction of the environment of navigation, the camera position can be computed by matching the SIFT features extracted from the acquired images

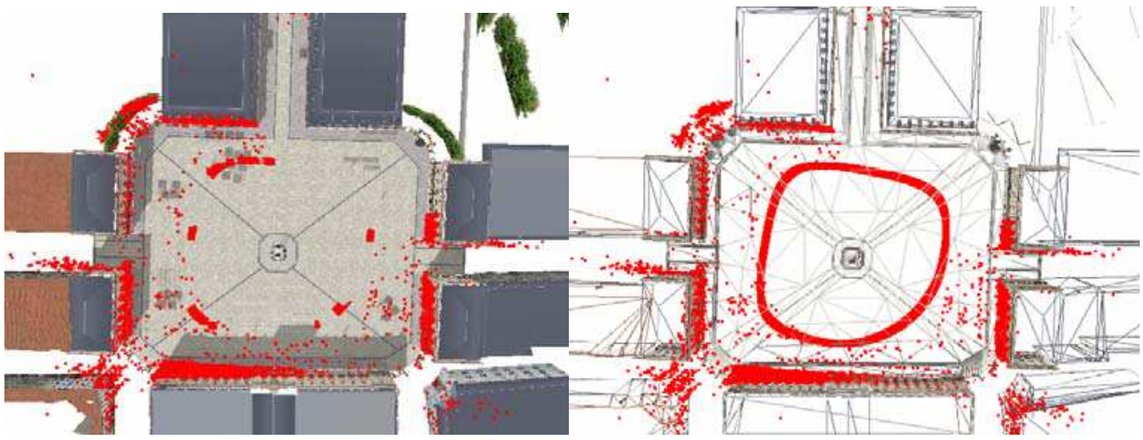


Fig. 4. Environment 3D model construction

with those stored in the database, corresponding to the learned trajectory. When the initial camera position is unknown, the extracted SIFT features are matched with all features from the database. This starting procedure necessitates a considerable computation time. However, when started, the localization process uses the last computed position in order to compute the current one. This allows filtering the features from the database in order to consider only those corresponding to the current view. This filtering stage reduces the complexity computation by ignoring more than 90% of the features during the matching procedure, allowing thus real time localization.

Let $\{Q^i\}$, $i = 0 \ldots n$, be the 3D points extracted from the stereoscopic sensor defined as C (see figure 5). The points Q^i are expressed into the stereoscopic coordinate system. In order to compute the global position of the stereoscopic sensor, the SIFT matching process is used to associate to each point Q^i a point W^i of the environment 3D model.

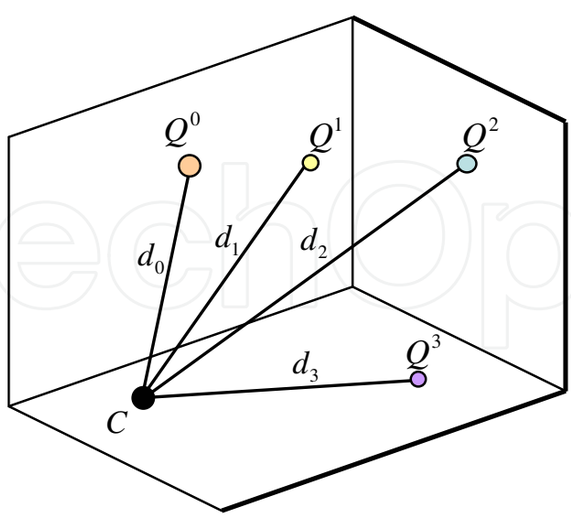


Fig. 5. Localization scheme

The Euclidian distances between the stereoscopic sensor and the points Q^i are computed and defined as d_j. The relation giving the camera position into the global coordinate system is expressed as follows:

$$\begin{cases} \left(C_x - Q_x^0\right)^2 + \left(C_y - Q_y^0\right)^2 + \left(C_z - Q_z^0\right)^2 = d_0^2 \\ \vdots \\ \left(C_x - Q_x^i\right)^2 + \left(C_y - Q_y^i\right)^2 + \left(C_z - Q_z^i\right)^2 = d_i^2 \\ \vdots \\ \left(C_x - Q_x^n\right)^2 + \left(C_y - Q_y^n\right)^2 + \left(C_z - Q_z^n\right)^2 = d_n^2 \end{cases} \tag{3}$$

In order to resolve the equations system (3), it is necessary to minimize the equation (4) using the Newton-Gauss algorithm.

$$S(\beta) = \sum_{i=0}^{n} r_i^2 \tag{4}$$

where $r_i = d_i^2 - f_i(\beta)$, $f_i(\beta) = (C_x - Q_x^i)^2 + (C_y - Q_y^i)^2 + (C_z - Q_z^i)^2$ and $\beta = (C_x, C_y, C_z)$.

The Newton-Gauss algorithm is particularly interesting when it starts with an initial value of β closed to the desired solution. This initialization step can be achieved using different techniques like Kalman filtering, fusion from several sensors like odometer or simply using the previous position. The minimization process can be expressed as follows:

$$\beta^{k+1} = \beta^k + \left(\mathrm{J}^{\mathrm{T}}\,\mathrm{J}\right)^{-1}\mathrm{J}^{\mathrm{T}}\,r \tag{5}$$

where J is the Jacobian matrix of $f(\beta)$:

$$\mathrm{J} = 2\begin{pmatrix} C_x - Q_0^x & C_y - Q_0^x & C_z - Q_0^z \\ \vdots & \vdots & \vdots \\ C_x - Q_n^x & C_y - Q_n^x & C_z - Q_n^z \end{pmatrix} \tag{6}$$

r and $f(\beta)$ are respectively the vectors composed with r_i and $f_i(\beta)$ ($i=1\ldots n$)

In addition, it is possible to increase the localization robustness by using the RANSAC technique [12]. This technique consists first in choosing randomly three points Q^0, Q^1 and Q^2 from the 3D reconstructed points, provided by the stereo matching of the SIFT features. Using these chosen points, the global position is then computed. These two steps are repeated until convergence, i.e., when the current calculation result is close to the previous one. Taking into account the RANSAC scheme, equation (3) (with n=4) can be rewritten as:

$$t^2 \cdot H_1 + t \cdot H_2 + H_3 = 0 \tag{7}$$

This equation has two solutions t_1, t_2:

$$t_1 = \frac{-H_2 - \sqrt{H_2^2 - 4H_1 \cdot H_3}}{2H_1} \text{ or } t_2 = \frac{-H_2 + \sqrt{H_2^2 - 4H_1 \cdot H_3}}{2H_1} \tag{8}$$

$$\text{where}: \begin{cases} H_1 = N_4^2 + N_2^2 + 1 \\ H_2 = 2\left(N_4 M_1 + N_2 M_2 - Q_z^0\right), \\ H_3 = M_1^2 + M_2^2 + \left(Q_z^0\right)^2 - d_0^2 \end{cases} \begin{cases} M_1 = N_3 - Q_x^0 \\ M_2 = N_1 - Q_y^0 \end{cases},$$

$$\text{with}: \begin{cases} N_1 = \dfrac{A_1 D_2}{B_1 A_2} - \dfrac{D_1}{B_1} \\ N_2 = \dfrac{A_1 C_2}{B_1 A_2} - \dfrac{C_1}{B_1} \end{cases}, \begin{cases} N_3 = -\dfrac{D_1 + B_1 P_1}{A_1} \\ N_4 = -\dfrac{B_1 P_2 + C_1}{A_1} \end{cases}$$

$$\text{and}: \begin{cases} A_1 = 2\left(Q_x^1 - Q_x^0\right); A_2 = 2\left(Q_x^2 - Q_x^0\right) \\ B_1 = 2\left(Q_y^1 - Q_y^0\right); B_2 = 2\left(Q_y^2 - Q_y^0\right) \\ C_1 = 2\left(Q_z^1 - Q_z^0\right); C_2 = 2\left(Q_z^2 - Q_z^0\right) \\ D_1 = \left(Q_x^0\right)^2 - \left(Q_x^1\right)^2 + \left(Q_y^0\right)^2 - \left(Q_y^1\right)^2 + \left(Q_z^0\right)^2 - \left(Q_z^1\right)^2 - \left(d_0\right)^2 + \left(d_1\right)^2 \\ D_2 = \left(Q_x^0\right)^2 - \left(Q_x^2\right)^2 + \left(Q_y^0\right)^2 - \left(Q_y^2\right)^2 + \left(Q_z^0\right)^2 - \left(Q_z^2\right)^2 - \left(d_0\right)^2 + \left(d_2\right)^2 \end{cases}$$

The camera (left or right) position estimated from equation (8) is considered to compute the retro-projection (see section 2.3) of each point Q^i. A retro-projection error can be thus computed for each point Q^i. If the error is less than 2 pixels, then the point Q^i associated to this error is considered as an inlier point with respect to the RANSAC procedure. The final estimation of the camera position is then computed using equation (3) and by considering only the inliers.

4. Experimental results

The proposed global localization method is tested and evaluated considering different environment conditions (cf. figure 6). The tests are achieved using a specific software simulation platform, which allows to generate 3D virtual environments and to construct mobile vehicles equipped with different video sensors. The evaluation consists to compare the real trajectory and the estimated one, (during the self global localization process). This comparison can be achieved by computing the deviation between the two trajectories. Recall that the estimated trajectory is based on the learned one, which corresponds to the constructed environment 3D model.

To construct the environment 3D model, the learning process is realized using a stereo sequence of 80 couples. The constructed model is composed with 9452 3D points, which represents a circular trajectory of about 142.42 meters (see figure 4).

Graphs 1–4 represent the evolution of the error (in meters) between the real and estimated trajectories in different environment conditions. The error represents the distance between the estimated camera position and the closest point of the real trajectory.

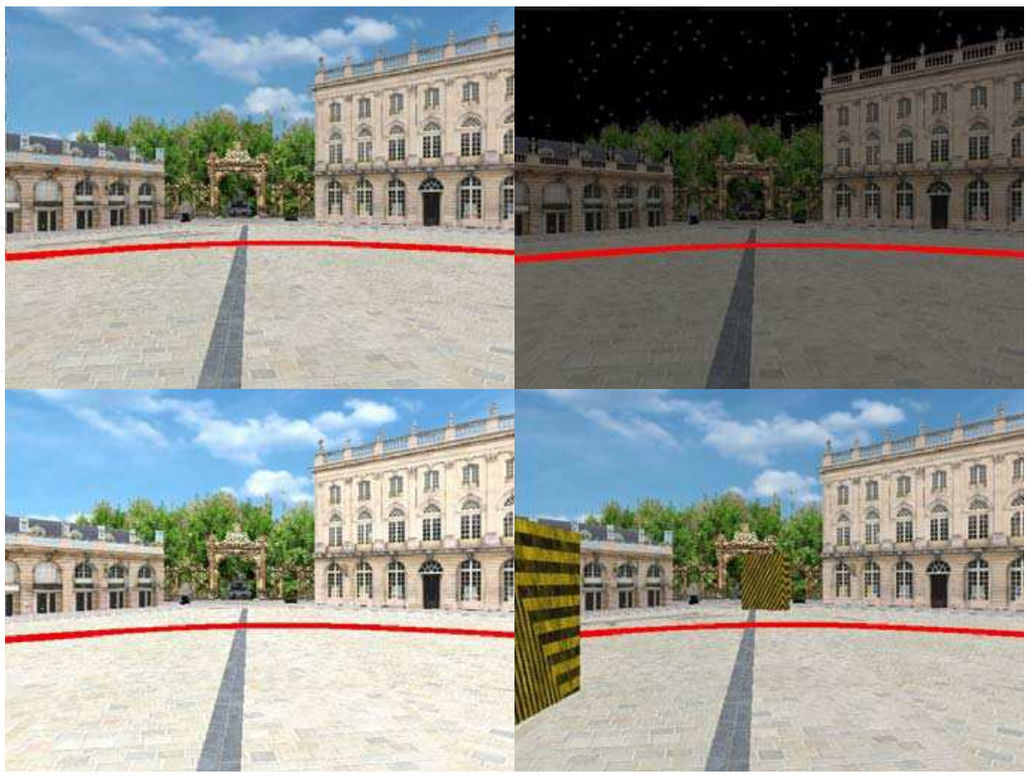
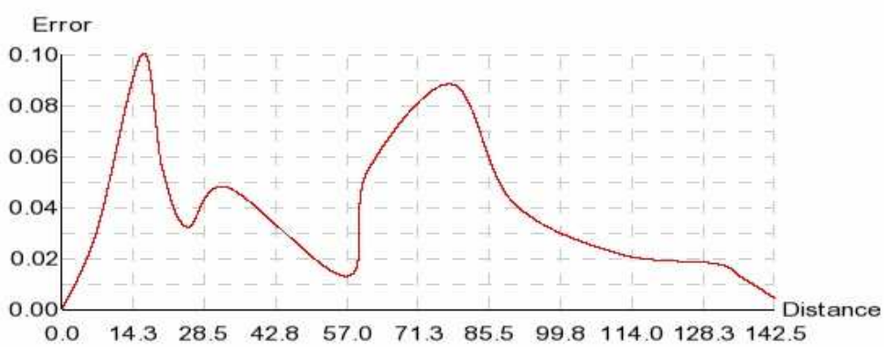


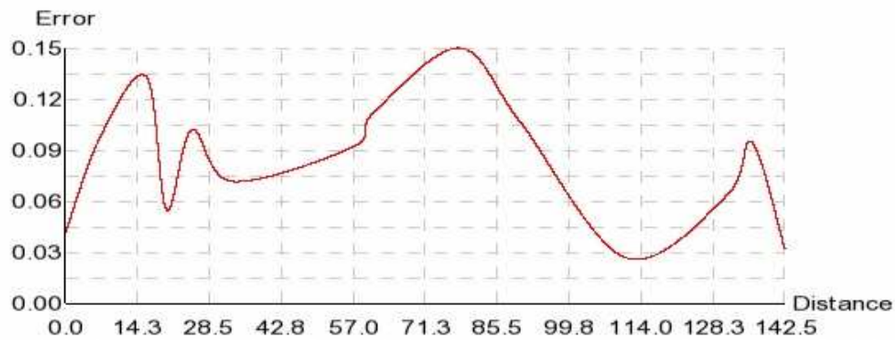
Fig. 6. upper left image: with normal illumination condition; upper right image: with dark illumination; lower left image: with high illumination; lower right image: with presence of objects in the environment



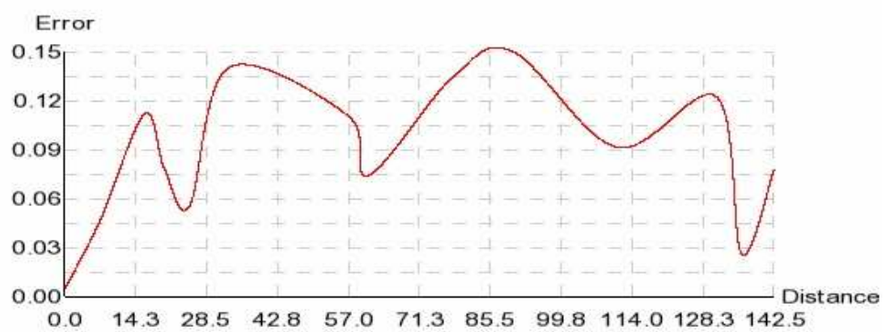
Graph 1. Error (in meter) in normal illumination conditions

The graph 1 shows that the error is less than 10 cm. Compared to classic GPS systems, the proposed localization process is more reliable.

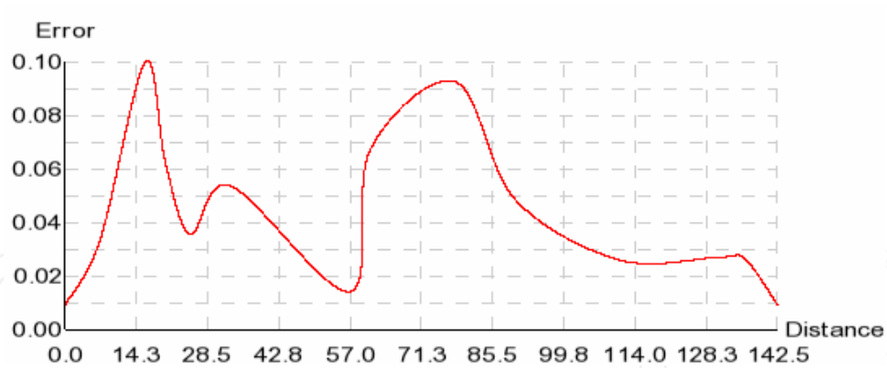
Most global localization techniques based on feature extraction are often sensitive to the environment illumination problem [13]. Graphs 2 and 3 show respectively the error obtained by the proposed method in dark and high illumination conditions. The localization results remain accurate. The pose of the camera is estimated with a maximum error of 15 cm. This performance is due principally to the using of the SIFT feature extractor, which is not considerably influenced by the illumination changes, and, as a consequence, the SIFT matching process provide enough correct matched points to get an efficient result.



Graph 2. Error (in meter) in dark illumination conditions



Graph 3. Error (in meter) in high illumination conditions



Graph 4. Error (in meter) in the case of presence of objects near to the learned trajectory

The last test consists in changing statically the learned environment by adding objects near to the trajectory, with normal illumination conditions. The error evolution shown in the graph 4 is close to the one associated to the case without presence of objects near to the learned trajectory (see graph 1). The presence of objects in the navigation environment decreases the number of matched SIFT features. This does not influence the precision of localization, while the illumination conditions remain unchanged. Indeed, when graphs 1 and 4 are compared, one can see that, globally, the precision is approximately identical.

5. Conclusion

This chapter proposes a robust method to achieve global localization. This method is based on a learning process, which consists to construct an environment 3D model from spatial and temporal matching of the SIFT features. Having the environment 3D model, the localization step is performed by searching correspondences between 3D reconstructed points and the 3D points belonging to the 3D model throughout the utilization of the SIFT features. The reliability of the proposed method is improved by using the RANSAC technique.

The proposed method is tested and evaluated in different environment conditions, using a software simulation platform. The tests show that the method is robust and provides reliable results. In deed, the error between the estimated and the real trajectories is less than 10 cm in normal illumination conditions. Thanks to the SIFT extractor, the maximum error does not exceed 15 cm, when considering illumination changes.

In terms of computation time, the localization process runs at 1.2 Hz for images with a resolution of 640x480 images, using a PC machine with a Core Duo running at 2.2 GHz. The SIFT extraction procedure consumes about 95% of the processing time. This is due to the high number of scales considered during the SIFT extraction. More tests are in progress in order to reduce the number of the SIFT points without loss of precision. Indeed, a relation can be established between the number of SIFT points, the processing time and the desired global precision.

Thanks to the using of stereovision and SIFT extractor, the proposed localization method is more interesting in terms of computation time and reliability. The work is in progress to compare the method with other ones using simulation and evaluation in real conditions through an experimental automated vehicle platform.

6. References

T. Chen. "*Development of a vision-based positioning system for high density area*". In Asian Conference on Remote Sensing (ACRS'99), Hong Kong, China, Nov 22-25 1999

M. Kais, S. Morin, A. de la Fortelle, and C. Laugier. "*Geometrical model to drive vision systems with error propagation*". In 8th International Conference on Control, Automation, Robotics and Vision (ICARCV'04), Kunming, China, Dec. 3-9 2004.

Cui, Sheyhi. "*Autonomous vehicle positioning with GPS in urban canyon environment*". IEEE Transaction on Robotics and Automation, 2003. Monocular Vision for Mobile Robot Localization and Autonomous Navigation,

H. Katsura, J. Miura, M. Hild, and Y. Shirai. "*A view-based outdoor navigation using object recognition robust to changes of weather and seasons*". In IEEE RSJ/International conference on Intelligent Robot and System (IROS'03), pages 2974-2979, Las Vegas, Nev., USA, Oct. 27-31 2003.

E Royer, M Lhuilier, M. Dhome, and T. Chateau. "*Towards an alternative gps sensor in dense urban environment from visual memory*". In 15th British Machine Vision Conference (BMVC'04), London, U.K., Sept. 7-9 2004.

Bill Triggs Philip F.Lauchian Richard I. Hartley and Andrew W. Fitzgibbon, "*Bundle Adjustment a Modern Synthesis*", "*Vision Algorithms: Theory and Practice*" vol. 1883, pp. 298—372. Springer Verlag, LNCS (2000)

A. Georgiev and P.K. Allen. "*Localization methods for a mobile robot in urban environments*". In IEEE Transactions on Robotics and Automation (ICRA'04), 2004.

C. Harris, and M.J. Stephens, "*A combined corner and edge detector*". In Alvey Vision Conference, pages 147-152, 1988.

David G. Lowe, "*Distinctive image features from scale-invariant keypoints*", International Journal of Computer Vision, 60, 2 (2004), pp. 91-110

Z. Zhang. "*A flexible new technique for camera calibration*". IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11):1330-1334, 2000.

W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. Numerical Recipies in C. Cambridge University Press, 1988. pp. 681-689.

Martin A. Fischler and Robert C. Bolles. "*Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartograph*y". Comm. of the ACM 24: 381–395

K. Mikolajczyk and C. Schmid. "*A Performance Evaluation of Local Descriptors*". IEEE Trans. on Pattern Analysis And Machine Intelligence, Vol. 27, No. 10, Oct. 2005

Stereo Vision

Edited by Asim Bhatti

ISBN 978-953-7619-22-0

Hard cover, 372 pages

Publisher InTech

Published online 01, November, 2008

Published in print edition November, 2008

The book comprehensively covers almost all aspects of stereo vision. In addition reader can find topics from defining knowledge gaps to the state of the art algorithms as well as current application trends of stereo vision to the development of intelligent hardware modules and smart cameras. It would not be an exaggeration if this book is considered to be one of the most comprehensive books published in reference to the current research in the field of stereo vision. Research topics covered in this book makes it equally essential and important for students and early career researchers as well as senior academics linked with computer vision.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Sergio Nogueira, Yassine Ruichek and François Charpillet (2008). A Self Navigation Technique Using Stereovision Analysis, Stereo Vision, Asim Bhatti (Ed.), ISBN: 978-953-7619-22-0, InTech, Available from: http://www.intechopen.com/books/stereo_vision/a_self_navigation_technique_using_stereovision_analysis

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821