Monocular visual odometry using features in the Fourier domain

Gastón Araguás, Jorge Sánchez, Luis Canali
Centro de Investigación en Informática para la Ingeniería (CIII)
Universidad Tecnológica Nacional - Facultad Regional Córdoba
garaguas@scdt.frc.utn.edu.ar

Abstract—
This paper presents a method to obtain odometry information using a monocular camera. It is intended for a vehicle moving on a plane, smooth surface. Only one camera is used and odometric information is obtained processing image blocks in the Fourier domain. From different shifts among image blocks an homography that can translate the set of features associated with the vehicle pose of a given time interval to those of the vehicle's initial pose can be computed.

I. Introduction

Since long time video sensors intended for navigation of terrestrial vehicles have been of interest within the Robotics community. Notwithstanding several navigation schemes developed, those based in vision are still considered because their good accuracy/price ratio. Steady growth of on-board computer power is making video-based navigation ever more attractive, for terrestrial as well as aerial vehicles. Literature gives the name of "visual odometry" to the process of obtaining measurements of pose variations (i.e. position + orientation) of a given robot using monocular or stereo camera arrangements. This work introduces a method intended for visual odometry based on spectral features. It uses a technique of image registering known as phase correlation for calculations of homography between different poses of the system with respect to the initial.

Recent papers explore different possibilities of visual odometry measurement, using stereo or monocular systems [26] [8] [17] [19], considering six DOF movements (6-DOF) [1] [14] or 3-DOF systems [4] [24] [19], with cameras carried by humans [7] or in helicopters [1] [14], with cameras aimed to the ceiling [5], or to the floor [7] or looking at the front [24] [6], there is also a report using an optical mouse as odometric sensor [21]. In most of these works the methodology used is similar: it deals with finding features in two successive images to infer the camera motion (egomotion) or for building a 3D map where to localize the camera(s) (SLAM) [20] [23], using some statistical method as RANSAC [10] [18] or MLE [27] to improve fitting of the set of features found with the movement model.

Optical flow is also used as in [3], where position info is obtained by integration of translational velocity of the floor and orientation by integration of angular velocity of the image ceiling. Errors reported in this paper are about 3.3% in the most favorable case comprising indoor navigation on a carpet.

In a growing group of cases, visual odometry is used along with other sensors as global positioning system receivers (GPS) or inertial units (IMU), fusing information by means of Kalman or extended Kalman filters. If the three sensor systems are available as in [26] prevailing strategy is to increment the priority of data arriving from the better performing sensor at the time, and so IMU is preferred where accelerations are prevailing, visual odometry when velocity information is relevant and GPS whenever accelerations and speeds are quite small.

II. Egomotion estimation

For terrestrial vehicles moving on a single plane the problem is reduced to the estimation of each of the three intervening DOFs. Considering a camera rigidly attached to the robot body and knowing the rigid transformation between the camera and the robot coordinates system [13], pose changes of the robot can be calculated by estimating the pose change of the camera's reference frame.

We assume for the moment that the camera has its focal axis normal to the plane of movement. Let us represent a point P in the (F) coordinates system by P^F. In this case a set of points in the world coordinates system (WCS) P_i^W belonging to the navigation plane will have in the frame of an image captured at time A coordinates P_i^A. The same set of points in the image frame taken at time $B > A$ will have coordinates P_i^B. Because of the planar robot movement, the relation between the two coordinates systems is a rigid transformation [11]. Finally, if camera parameters are known [25] then it is possible to refer the set of points to the camera coordinates system (CCS) and therefore calculate the change of pose of the vehicle finding the rigid transformation existing among the camera's coordinates systems at time A and B.

Without loss of generality we assume the plane of movement is on $z = 1$ in the CCS, so that the set of point's coordinates in the CCS at time A are

$$P_i^A = \left(x_i^A, y_i^A, 1^A\right)^T \tag{1}$$

these are homogeneous coordinates of points in R^2. The transformation between CCS when the robot is moving in a plane is a planar Euclidean transformation (a composition of translations and rotations in R^2). Representing with $\mathcal{R}_{AB}$ the rotation matrix describing the frame (B) in the coordinates system (A) and with t^A the translation vector

between frames, a point P^B can be written in coordinates of the system (A) with

$$P^A = \mathcal{R}_{AB} P^B + t^A \tag{2}$$

In homogeneous coordinates this planar transformation is given by an unique matrix H called *homography matrix*, in this case

$$H_{AB} = \begin{pmatrix} \mathcal{R}_{AB} & t^A \\ \mathbf{0^T} & 1 \end{pmatrix} \tag{3}$$

with $\mathbf{0}$ a null vector.

When the robot moves the same set of points is captured from the camera at times $A < B < C < \cdots$ and represented in the moving CCS with coordinates P_i^A, P_i^B, P_i^C, etc. respectively. To obtain the coordinates of the points in the frame (A) knowing them in frame (C) we can first obtain the coordinates of the set in the frame (B) and then maps it into the frame (A)

$$P_i^A = H_{AB}\, P_i^B = H_{AB}\left(H_{BC}\, P_i^C\right) = H_{AB} H_{BC}\, P_i^C \tag{4}$$

or directly apply the complete homography from (C) to (A)

$$P_i^A = H_{AC} P_i^C \tag{5}$$

from (4) and (5) we have $H_{AC} = H_{AB} H_{BC}$.

The homography matrix that links the coordinates of the set in the actual frame (that is at the actual time) with the first frame give us the change of pose of the camera relative to the origin, and therefore the change of pose of the robot. According to (3) and (5), the translation of the robot at time K in homogeneous coordinates is given by the third column of H_{AK} and the rotation by the rotation matrix $\mathcal{R}_{AK}$ forming this homography. At this point to estimate the homography between frames is equivalent to estimate the change of pose of the robot.

Fig. 1. Working platform

III. Image rectification and spectral features

Image features normally used for visual odometry are corners and/or edges. A Harris detector is classically used to find corners [12], and then movement estimation is obtained using Lukas-Kanade method [16]. Unfortunately this kind of features are not always present on an image, specially on those captured from floors with homogeneous textures where gradient is very small. Another difficulty arises from image features that are not uniformly distributed in the captured frame, which may cause a bad estimation of vehicle's odometry. When dealing with this problem, as an example, in [7], error introduced by the momentaneous loss of image useful features is made equivalent of the slipping of the wheels of a vehicle with odometry based in optical encoders attached to them.

An alternative to gradient-based image features is image registration using spectral features.

A. Phase correlation method

If the vehicle moves on a plane surface, pictures of the floor taken by its camera will show displacements and rotations from one another in accordance with vehicle movements. Computing successive shifts and rotations from the obtained images it is possible to figure out the vehicle's change of pose. The process of computing the variation between two images, namely the determination of displacement and rotation is known as *image registration*.

There are several methods and techniques for image registration capable to find and compensate for image perturbations such as shift, scaling, rotation, deformation, perspective effects, etc. [2]. Taking into account first only the displacement between images due to sensor movement the method known as phase correlation can be used for calculations [15].

The Phase Correlation Method (PCM) is based on the fact that the Fourier transforms of two identical functions shifted one from the other differ only in phase. This is also known as Fourier's shift theorem.

Let two images i_a and i_b differing only in a displacement (u, v), so that

$$i_a(x, y) = i_b(x + u, y + v) \tag{6}$$

then their Fourier transforms are related by

$$I_a(\omega_x, \omega_y) = e^{j(u\omega_x + v\omega_y)} I_b(\omega_x, \omega_y) \tag{7}$$

where I_a and I_b are the Fourier transforms of images i_a and i_b, and u and v are the amounts of the displacements of the images in each axis. That means that both transforms are equal in magnitude but they have a difference in phase that is directly related with the displacement between the images. This displacement can be computed using (7) calculating the cross power spectrum of Fourier transforms I_a and I_b.

The cross-power spectrum (CPS) between two complex functions is defined as

$$\frac{F(\omega_x, \omega_y) G^*(\omega_x, \omega_y)}{|F(\omega_x, \omega_y)||G^*(\omega_x, \omega_y)|} \tag{8}$$

The CPS of a function with itself is 1. Then, computing

the cross-power spectrum of I_a yields

$$\frac{I_a(\omega_x,\omega_y)I_a^*(\omega_x,\omega_y)}{|I_a(\omega_x,\omega_y)||I_a^*(\omega_x,\omega_y)|}=1 \tag{9}$$

where I_a^* is the conjugate of I_a. Using (7), (9) can be written as

$$Q(\omega_x,\omega_y)=\frac{I_a(\omega_x,\omega_y)I_b^*(\omega_x,\omega_y)}{|I_a(\omega_x,\omega_y)||I_b^*(\omega_x,\omega_y)|}=e^{j(u\omega_x+v\omega_y)} \tag{10}$$

obtaining the phase correlation matrix. This mean that the phase of the cross power spectrum among the two transforms is the difference of phase between them. If the inverse transform of the phase correlation matrix is computed an impulse centered exactly in the value of the displacement between the pictures is obtained

$$\mathcal{F}^{-1}[Q(\omega_x,\omega_y)]=q(x,y)=\delta(x-u,y-v) \tag{11}$$

Using discrete Fourier transforms the impulse changes into an unit pulse centered in (u,v).

B. Rotation and translation

Using PCM a good measure of displacement can be obtained if the images are not rotated with respect to each other. Although PCM can be extended to register images that are only rotated as described in [22], to detect rotations it is mandatory to change to polar coordinates, and so interpolations are needed for those values falling out of the grid. In the case of both translation and rotation are present, the coordinates transform results very expensive in terms of computer time, since it is necessary to find by iteration the rotation angle that best approximates the phase correlation matrix to an unit pulse. Instead, we compute rotations using the information of the displacements taking place in different regions of images. This is a good approach whenever rotations between successive images are small. This means that visual odometry can be performed using information from zones of the image and it is not necessary to work on the whole image. This zones of the image are the meaningful features for pose determination, and as they are in Fourier's domain they are called spectral features.

A spectral feature is the representation in Fourier's domain of a portion of the image of $2^n \times 2^n$, n being a positive integer which value is fixed from the maximum allowed displacement between two images; which in turn is a function of the vehicle's velocity. Size is selected taking into account the efficiency of fast Fourier's transform on data arrays whose dimensions are integer powers of 2.

Spectral features have certain advantages over those traditionally used (corners, edges, blobs), since they must not be looked after, and so no special interest operators (such as Harris) must be used as a first step of feature extraction. It must be remembered that a spectral feature is a region of the image with predefined size. They use spectral information of the image, which is better when navigating in regions with little gradient information (i.e, corners and edges) such as carpets, wooden and plastic floors, etc.

The number of spectral features to be used is also a tradeoff. A minimum of three is required for the homography calculation, but usually more than three features are used to overdetermine the system and thus perform a more robust estimation.

C. Projective Rectification

A monocular system such as the one described, having its focal axis perpendicular to the navigation plane can be used only for odometry measurements. Nevertheless an equivalent system can be configured from a camera having an oblique focal axis using a projective rectification of the image thus generated. This has the advantage that the same monocular device used for other tasks, such as obstacle avoidance, environment acquisition and maping, etc. can be used for the odometry measurements too.

In monocular system vehicles with the camera's focal axis oblique to the plane of navigation (as the working platform in figure 1), change of pose of the robot system can be described using the induced homography between successive rectified images of the navigation plane Π. The projective distortion introduced in the system can be corrected in order to obtain an equivalent, non-oblique system, as shown in figure 2, if we known the camera height h with respect to the floor and the angle Φ between the focal axis of the camera and the normal $\mathbf{n}$ to the navigation plane.

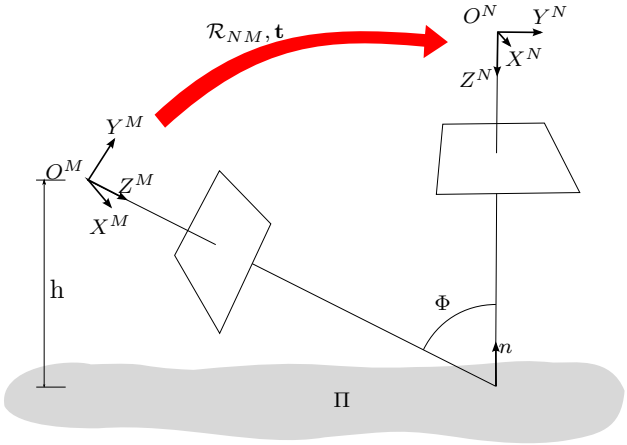

Fig. 2. Proyective rectification

Let us assume, for the sake of simplicity, that axis Y_c of the CCS is in the plane formed by the focal axis and the normal to the navigation plane Π. Following [9], let P^M be the coordinates in the CCS of a point on the navigation plane, the coordinates of the same point seen from a virtual camera with their focal axis normal to the plane will be given by

$$P^N=\left(\mathcal{R}_{NM}+\frac{\mathbf{t}\cdot\mathbf{n}^\mathbf{T}}{h}\right)P^M \tag{12}$$

where $\mathcal{R}_{NM}$ is a rotation matrix and $\mathbf{t}$ is a translation vector that represent the location and orientation of the

perpendicular camera with respect to the oblique one, and

$$\mathcal{R}_{NM} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\Phi & \sin\Phi \\ 0 & -\sin\Phi & \cos\Phi \end{bmatrix} \quad \mathbf{t} = \begin{bmatrix} 0 \\ t_y^N \\ t_z^N \end{bmatrix} \tag{13}$$

Adding the condition that $z^N = 1$ the coordinates $t_x^N = |\mathbf{t}|\cos\Psi$ and $t_y^N = |\mathbf{t}|\sin\Psi$ of the translation vector can be found trigonometrically, where Ψ is the angle between $\mathbf{t}$ and $\mathbf{n}$. By means of the law of cosines in the triangle formed by $|\mathbf{t}|$, $\frac{h}{\cos\Phi}$ and $z^N = 1$ we have

$$|\mathbf{t}|^2 = \frac{h^2}{\cos\Phi} + 1 - 2h \tag{14}$$

and by means of the law of sines in the same triangle

$$\sin\Psi = \frac{h}{|\mathbf{t}|}\tan\Phi \tag{15}$$

If K is the camera calibration matrix, the coordinates (in pixels) of points m_1 and m_2 on the original and rectified images are related through

$$m_2 = K\left(\mathbf{R} + \frac{\mathbf{t}\cdot\mathbf{n}^\mathbf{T}}{h}\right)K^{-1}m_1 \tag{16}$$

IV. Homography estimation

Now, the goal is to calculate the homography that links the actual pose with the initial one, that is to say that for a set of points captured at time K it is known H_{AK} so that we map the coordinates system (K) into the first one (A)

$$P_i^A = H_{AK}P_i^K \tag{17}$$

Points used for the calculation of the homography correspond to the spectral features of images captured from the floor. Displacement calculation using image registration is performed on this set of images.

If the images captured from the camera in two successive samples are slightly displaced and rotated from each other, meaning that sampling rate is high in relation with the vehicle velocity, movement of image spectral features can be modeled using pure translation. In this case, the parameters describing the general movement of the vehicle are obtained considering the whole set of individual displacements of those features, as shown in figure 3. That is, the transformation between each feature will be performed considering only pure translation, and by setting a feature arrange sparsely distributed in the whole image the set will perform a complete planar transformation including translation and rotation.

Process begins registering the features of the first and second captured images. In this way the two first sets of point's coordinates corresponding to each image frames are obtained and subsequently the homography between frames is computed. With each new image acquired the homography is recalculated, thus relating the new image frame with image frame zero. This yields the transformation of current pose to initial pose, saying (B) the current frame and (A) the intial one we have $P_i^A = H_{AB}P_i^B$.

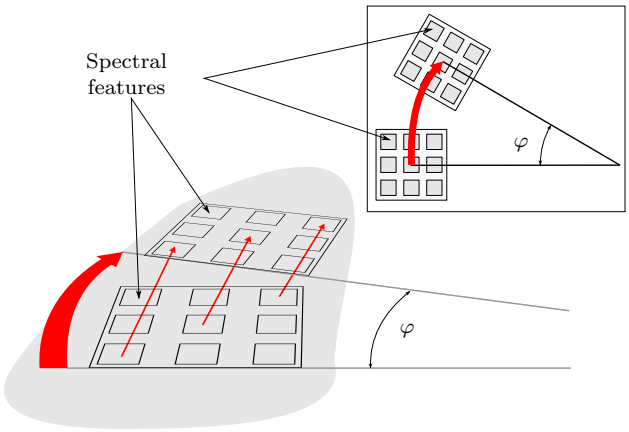

Fig. 3. Estimation of the Euclidian planar transformation using PCM method to register features

When pose change overpasses a threshold value (e.g. 5 pixels for translation or 0.05 radians for rotation between frames) the current image and its homography H_{AB} are stored as a process anchor value. From this point new images are registered with image zero and with this anchor image, obtaining the two homographies between both frames, that is H_{AC} and H_{BC} where (C) refers to the current frame. Now two ways exists to port current pose (or frame) to zero pose: one is direct H_{AC} and the other is composed $H_{AB}H_{BC}$. They may differ from one another due to errors in the registration process of noisy images. From this two ways the best one in terms of registering errors is chosen and stored. If the direct calculation (in this case H_{AC}) results with less error, the intermediate anchor is discarded and the current image takes its place (in this case (C) becomes (B)).

For each new frame the homography transformations between them and the two previous anchors are computed in the some way. The numbers of anchors will grow until the maximum is arrived. This is a system variable which must be chosen in terms of precision and computational cost. The number of anchors to be used is related to vehicle maximum speed. In general a number between 5 and 8 has been found to be adequate.

When the system goes far from the origin, tha calculation of the homography of actual pose with respect to pose zero is made by composition and not by direct registration, because images are no more overlapped. As the vehicle goes farther, new anchors are generated. When anchors number reaches chosen value N the older is discarded to store a new one, keeping the discarded anchor's homography.

The process keeps on with the acquisition of new images and new anchors. At any arbitrary time (K) we can compose the homography between the current frame (K) and the frame (A)

$$H_{AK} = H_{AB}H_{BC}H_{CD}\ldots H_{JK} \tag{18}$$

V. Conclusion

The method presented allows odometry calculation using a monocular camera system. It can be used aiming the camera directly to the floor or with an oblique focal axis. The first configuration yields better resolution in movement determination and the second allows the use of visual information obtained for other tasks such as obstacle avoidance, trajectory planning, maping, etc.

Spectral features have advantages over corner detection in zones with low gradient values. Displacement calculation using phase correlation lends itself well to be embedded in logical devices as FPGAs, allowing the construction of a stand alone visual odometric sensor with relative simplicity.

Tests performed on short trajectories (indoor paths of approximately $5m$ long) have shown error values of about 1% of the length of a known trajectory. Tests on longer paths have been not performed yet, but error values are expected to be the same. Main cause of error is the integration process needed to obtain position information because the system is a velocity measuring device. One of the interesting possibilities arising from this work is fusion of visual odometry with the information of a low-cost GPS receiver, and/or with that of the optical encoders on the vehicle´s wheels, for navigation in low structured environments.

References

[1] O. Amidi, T. Kanade, and K. Fujita. A visual odometer for autonomous helicopter flight. *Journal of Robotics and Autonomous Systems*, 28:185–193, 1999.

[2] L. G. Brown. A survey of image registration techniques. *ACM computing surveys*, 24:325–376, 1992.

[3] J. Campbell, R. Sukthankar, and I. Nourbakhsh. Techniques for evaluating optical flow for visual odometry in extreme terrain. In *Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, volume 4, pages 3704–3711 vol.4, 2004.

[4] J. Campbell, R. Sukthankar, I. Nourbakhsh, and A. Pahwa. A robust visual odometry and precipice detection system using consumer-grade monocular vision. In *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, pages 3421–3427, 2005.

[5] G. Caron. Homography-based monocular visual odometry. 2007.

[6] Nguyen Xuan Dao, Bum-Jae You, and Sang-Rok Oh. Visual navigation for indoor mobile robots using a single camera. In *Intelligent Robots and Systems, 2005. (IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 1992–1997, 2005.

[7] S. DiVerdi and T. Hollerer. Groundcam: A tracking modality for mobile mixed reality. In *Virtual Reality Conference, 2007. VR '07. IEEE*, pages 75–82, 2007.

[8] C. Dornhege and A. Kleiner. Visual odometry for tracked vehicles. *Proc. of the IEEE Int. Workshop on Safty, Security and Rescue Robotics (SSRR), Gaithersburg, Maryland, USA*, 2006.

[9] O. D. Faugeras. Motion and structure from motion in a piecewise planar environment. *INT. J. PATTERN RECOG. ARTIF. INTELL.*, 2:485–508, 1988.

[10] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24:381–395, 1981.

[11] D. A Forsyth and J. Ponce. *Computer vision: a modern approach*. Prentice Hall Professional Technical Reference, 2002.

[12] C. Harris and M. Stephens. A combined corner and edge detector. *Alvey Vision Conference*, 15:50, 1988.

[13] J. A Hesch, A. I Mourikis, and S. I Roumeliotis. Determining the camera to robot-body transformation from planar mirror reflections. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008. IROS 2008*, page 3865â3871, 2008.

[14] J. Kelly, S. Saripalli, and G. S. Sukhatme. Combined visual and inertial navigation for an unmanned aerial vehicle. *6th International Conference on Field and Service Robotics - FSR 2007 42 (2007)*, 2007.

[15] C. D. Kuglin and D. C. Hines. The phase correlation image alignment method. *Proc. Int. Conf. on Cybernetics and Society*, 4:163–165, 1975.

[16] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *International Joint Conference on Artificial Intelligence*, 81:674–679, 1981.

[17] Qiang Lv, Wenhui Zhou, and Jilin Liu. Realization of odometry system using monocular vision. In *Computational Intelligence and Security, 2006 International Conference on*, volume 2, pages 1841–1844, 2006.

[18] D. Nister. Preemptive ransac for live structure and motion estimation. *Machine Vision and Applications*, 16:321–329, 2005.

[19] D. Nister, O. Naroditsky, and J. Bergen. Visual odometry. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–652–I–659 Vol.1, 2004.

[20] D. Nister, O. Naroditsky, and J. Bergen. Visual odometry for ground vehicle applications. *Journal of Field Robotics*, 23:3–20, 2006.

[21] J. Palacin, I. Valganon, and R. Pernia. The optical mouse for indoor mobile robot odometry measurement. *Sensors and Actuators A: Physical*, 126:141–147, 2006.

[22] B.S. Reddy and B.N. Chatterji. An fft-based technique for translation, rotation, and scale-invariant image registration. *Image Processing, IEEE Transactions on*, 5:1266–1271, 1996.

[23] S. Se, D. Lowe, and J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. In *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, volume 2, pages 2051–2058 vol.2, 2001.

[24] Hui Wang, Kui Yuan, Wei Zou, and Qingrui Zhou. Visual odometry based on locally planar ground assumption. In *Information Acquisition, 2005 IEEE International Conference on*, page 6 pp., 2005.

[25] Z. Zhang. A flexible new technique for camera calibration. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22:1330–1334, 2000.

[26] Z. Zhu, T. Oskiper, O. Naroditsky, S. Samarasekera, H. S. Sawhney, and R. Kumar. An improved stereo-based visual odometry system. *USA, August*, 2006.

[27] M. Zucchelli, J. Santos-Victor, and H.I. Christensen. Maximum likelihood structure and motion estimation integrated over time. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 4, pages 260–263 vol.4, 2002.