VISUAL ODOMETRY FOR AUTONOMOUS LOCALIZATION ON MARS

F. Souvannavong[1], C. Lemaréchal[1], E. Remetean[2], L. Rastel[2], and M. Maurette[2]

[1]*Magellium*
24, rue Hermès
31521 Ramonville Saint-Agne, France
contact : fabrice.souvannavong@magellium.fr

[2]*Centre National d'Etudes Spatiales*
18, avenue Edouard Belin
31401 Toulouse Cedex 9, France
contact : michel.maurette@cnes.fr

Abstract

We present a visual odometry system for short term self localization. It targets autonomous planetary exploration in an unknown rough environment. The singularity of the method is to estimate long elementary displacements at low frequency, 0.1Hz against more than 1Hz usually. Our algorithm is evaluated in a Mars-like visual environment, on short and long trajectories, as a whole and step by step. We provide preliminary characteristics of its performances that can be used to design an autonomous locomotion system.

Key words: Robotics Autonomy; Imaging Technologies; Rover Localization; Visual Odometry; ExoMars mission.

1. INTRODUCTION

To navigate reliably in autonomy, a mobile robot must know exactly where it is. Self localization in an unknown rough environment is then a key and challenging issue in mobile robotics targeting planetary exploration. We believe that visual odometry is a promising solution to the short term self localization problem in these conditions. This image processing technology computes elementary displacements between consecutive frames, that are next, integrated to localize the rover.

We present a visual motion estimation (VME) system, first initiated by the LAAS [1], that estimates elementary displacements of approximately 200mm~128px. The proposed method estimates rover motion between consecutive 3D point clouds, derived from consecutive stereoscopic views. The matching of 3D points is achieved such that geometrical constraints are preserved among selected matching points. Its performances are tightly linked to the 3D reconstruction and input rotational motion accuracies. Input ego-motion from wheel odometry is optionally used to reduce searching areas. The algorithm outputs either the complete 6D localization of the rover or only its 3D position completed by the orientation supplied by an external sensor.

Preliminary experiments target the evaluation of the system for an ExoMars like mission. It is important to give its step by step or as a whole characteristics. Externally measured localizations of each acquisition are essential for this purpose. Thus, we set up an high precision measurement method that provides the orientation of cameras with 0.05° accuracy and there position with 0.5mm accuracy. Preliminary experiments run in a Mars-like visual environment on the place of Serom at CNES, prove the efficiency of the method in various situations. On our data set, the system is able to localize the rover with a relative error inferior to 5%, from 122 stereoscopic views acquired on an 27m trajectory. We identify two major issues we will propose to solve. The first affects performances and arises from the possible presence of strong static or moving shadows. The second affects the execution time and arises from the high computational requirements of the matching.

The paper is organized as follows. We, first, present related work. Next, we describe the visual odometry problem and our implementation. We focus on landmark detection, matching of corners and motion estimation. We propose alternatives for each task, to improve system speed and accuracy. Next, we present the evaluation environment, including the measurement method of reference localizations, the description of the stereoscopic system and data sets. We continue by presenting the evaluation result. Finally, we conclude with some preliminary recommendations for an ExoMars like mission.

2. RELATED WORK

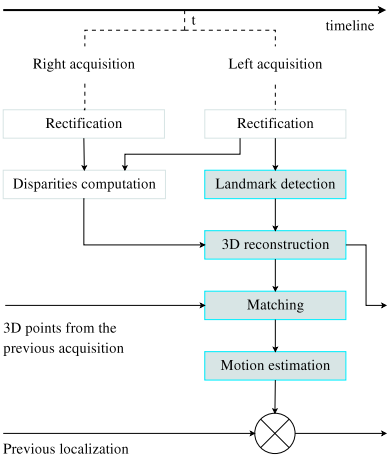
As far as we know, the idea of estimating camera motion from a sequence of images appeared in early '80s [2], and is now a very active field since camera quality and computation power have greatly improved. This technology is, in particular, very attractive to solve the problem of self localization of rovers targeting autonomous planetary exploration [3, 4]. Most solutions found in the literature assume small displacements in pixels between acquisitions, or relatively good initial estimates. It facilitates image feature tracking or matching in 2D or 3D, but imply high frequency processing [5]. In this way, our work differs by focusing on large displacements. The proposed method is related to the work presented in [6]. Yet, it mainly differs on two points. Firstly, they compute an initial set of matching landmarks using a visual similarity measure, whereas we use only geometrical constraints considering good attitude measurements are provided. Secondly, we introduce a different matching score that is more precise and flexible to select matching landmarks.

A recent challenge is now to improve long term localization with SLAM methods [7]. These systems have the ability to recognize visited places and consequently to adjust model parameters. We don't present such a system that requires more memory and computation time. Nevertheless, we experiment our method on long term trajectories to evaluate its limits, inherent to dead reckoning.

3. VISUAL ODOMETRY IMPLEMENTATION

Visual odometry principle is to integrate elementary ego-motion estimates of a set of cameras (usually one or two). Camera ego-motion estimation is achieved by detecting and tracking visual landmarks, which the 2D and 3D characterization allows to measure the displacement. The proposed visual odometry system relies on a calibrated stereoscopic bench. This configuration has the advantage of providing an accurate 3D euclidean reconstruction of landmarks for each acquisition pair. This simplifies the motion estimation problem since the 3D characterization of landmarks and the camera motion are not simultaneously estimated.

The ego-motion estimation is decomposed in four steps : image acquisition and rectification, landmark detection and characterization, feature matching and motion estimation. An overview of the method is presented in Fig. 1(a).



(a) Odometry Principle. The paper focuses on gray boxes.

(b) Acquisition system.

3.1. Landmark reconstruction

3.1.1. Acquisition and rectification

We assume that image pairs are acquired at time instant $(t-1)$ and (t) such that for a nominal displacement, half of images are overlapping. Images are first sub sampled to get a reasonable working size. Next, they are rectified to make the two view geometric constraint simple as depicted in Fig. 1(c). Applied correction maps are provided by the calibration, not described in this paper. We will retain that it computes dense correction maps to make the two view geometric constraint simple, i.e. pinhole projection model and rectilinear stereo rig. Projection and reconstruction formulas are, then, very simple on rectified images. Let $p_l(u, v)$ be the projection of $M(X, Y, Z)$ on the left sensor. Let d be the disparity such that $p_r(u + d, v)$ is the projection of M on the right sensor, $baseline$ the distance between optical centers and $wpix$ the pixel size and f the focal length. We have :

$$Z = \frac{baseline.f}{wpix.d} \quad X = \frac{wpix}{f}.Z.u \quad Y = \frac{wpix}{f}.Z.v \tag{1}$$

3.1.2. Landmark detection and reconstruction

Landmarks are extracted on left images at $(t-1)$ and (t). The type of available landmarks is mainly limited to corners in an unknown rough environment. Many corner detectors are proposed in the literature. The recent SIFT feature [8] has been proved very efficient but it is very time consuming and is therefore not appropriate. The well known Harris' detector [9] has, finally, retained our attention for its efficiency and simplicity.

We propose to modify the Harris' detector to get a better corner distribution. Indeed, areas with high contrasts tend to have a strong response. The risk is, then, to have many points gathered in the same area. To deal with this inconvenient, Harris' scores are normalized in their close neighborhood, insuring a good distribution while preserving a discriminative response. This is important, since the matching algorithm uses only best Harris' corners to reduce computation time, as we will see next. In the perspective to reduce computation time, landmarks are selected only in estimated overlapping areas. This approach requires, then, an initial estimation of the motion.

Once corners are extracted in previous and current left images, the stereoscopic configuration allows to compute their 3D coordinates in stereoscopic bench frames, respectively at time instant $(t-1)$ and (t). Without giving all details, the idea is to identify similar points in the left and right views by correlating image patches. A correlation score is very well adapted in this case, since we are working on rectified images.

3.2. Matching

The matching aims at identifying similar landmarks in left images at $(t-1)$ and (t). This step is crucial for motion estimation. Most solutions compare image features around landmarks. However, they assume small changes or heterogeneous environments which is not the case here. We propose to match points that are geometrically compatible, knowing the rotational motion R provided by an external sensor. Let $\mathcal{L}_p$ and $\mathcal{L}_c$ be the set of landmarks in respectively the previous left frame at $(t-1)$ and the current left frame at (t).

3.2.1. Principle

The algorithm relies on the observation that length and angles are preserved in the euclidean space. Thus, given two pairs of matching points $M^1(M_p^1, M_c^1) \in \mathcal{L}_p \times \mathcal{L}_c$ and $M^2(M_p^2, M_c^2) \in \mathcal{L}_p \times \mathcal{L}_c$, we have two constraints :

$$\|\overrightarrow{M_p^1 M_p^2}\| - \|\overrightarrow{M_c^1 M_c^2}\| = 0 \tag{2}$$

$$|\cos(\overrightarrow{M_p^1 M_p^2}, (R.\overrightarrow{M_c^1 M_c^2}))| = 1 \tag{3}$$

These constraints allow to compute a matching score as follows :

$$S_M(M_p, M_c) = \sum_{M_p', M_c'} f(\overrightarrow{M_p M_p'}, \overrightarrow{M_c M_c'}) \tag{4}$$

$$f_b(\overrightarrow{M_pM'_p}, \overrightarrow{M_cM'_c}) = \begin{cases} 1 & \text{if } (|||\overrightarrow{M^1_pM^2_p}|| - ||\overrightarrow{M^1_cM^2_c}||| < \epsilon_n \text{ and } \ldots \\ & |\cos(\overrightarrow{M^1_pM^2_p}, (R.\overrightarrow{M^1_cM^2_c}))| > (1-\epsilon_p)) \\ 0 & otherwise \end{cases} \quad (5)$$

where ϵ_n and ϵ_p are thresholds depending on the 3D reconstruction and the input rotational motion accuracies. A first iteration, very time consuming, gives a set of potential matches composed of points with the highest matching score. However, this step gives many mismatches and a consistency checking step is required. Its objective is to find a large subset of potential matches that are mutually consistent. This is achieved by iteratively removing matches with the smallest consistency score. The consistency score, $S_C(.)$, is similar to the matching score, where the sum is conducted on remaining matches. The algorithm stops when the consistency score of all remaining matches equals their number.

$$S_C(M) = \sum_{M'} f(\overrightarrow{M_pM'_p}, \overrightarrow{M_cM'_c}) \quad (6)$$

The algorithm is denoted *matching from binary scores* or abbreviated fb-matching.

3.2.2. Improvement

Previously proposed comparison function $f(.)$ is limited by its binary formulation. One direct limitation is the incapacity to differentiate many potential matches having exactly the same score. We, thus, introduce a different comparison function $f'(.)$ that provides a continuous matching score :

$$f_c(\overrightarrow{M_pM'_p}, \overrightarrow{M_cM'_c}) = \begin{cases} 0 \text{ if } (|\cos(\overrightarrow{M^1_pM^2_p}.(R.\overrightarrow{M^1_cM^2_c}))| < (1-\epsilon_p)) \\ \max(0, \frac{1-fmax}{\epsilon_n}|||\overrightarrow{M^1_pM^2_p}|| - ||\overrightarrow{M^1_cM^2_c}||| + fmax) \text{ otherwise} \end{cases} \quad (7)$$

This formulation has also the advantage of being more flexible. It is easy to add new constraints such as visual similarities or vector length weighting. This method is denoted *matching from continuous scores* or abbreviated fc-matching.

3.2.3. Optimization

The matching algorithm is particularly time consuming since the number of comparison explodes with the number of landmarks $O(n^4)$. Even with some filtering on minimal point distances, the complexity remains very high. However, only Harris' corners with the highest response are sufficient, in most cases, to get a good motion estimate. We propose a new approach that iteratively increases the number of landmarks, until enough points are matched.

Let order Harris' corners by decreasing detection score in the previous image. We, then, create the landmark subsets of equal size : $\mathcal{L}^1_p$ with first corners, $\mathcal{L}^2_p$, ..., $\mathcal{L}^n_p$ with last corners. This decomposition is also applied on corners in the current image. The algorithm, depicted in Fig. 1(d), iteratively runs following steps : add new corners, update matching scores of previous and new potential matches by comparing subsets, find a consistent subset of matches (not represented in the figure), continue if the subset of matches is too small. This method is denoted *fast matching*.

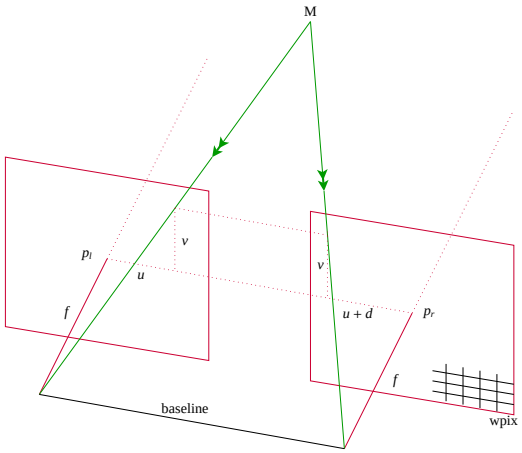
3.3. Motion estimation

We have now same landmarks identified and localized in previous and current images. Two options are offered at this stage : either the 6D ego-motion is estimated (translation and rotation) or only the translation that, then, completes the input rotation. We use the well known linear solution proposed in [10] to find the optimal 6D rigid transformation $(R|T)$, that minimizes the following energy :

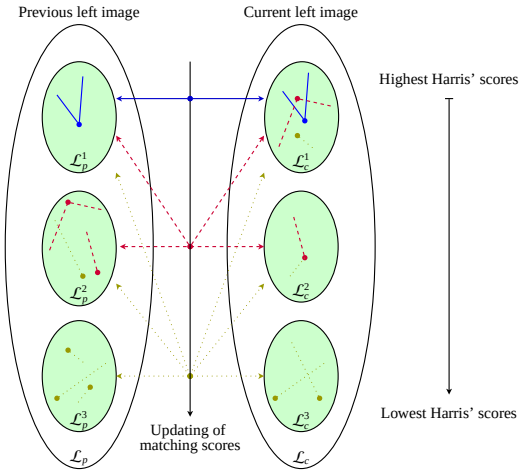
$$E(R|T) = \sum_M ||M_p - (R.M_c + T)||^2 \quad (8)$$

This method is called 6D-VME. In the case, we are confident in the input attitude provided by an external system, we estimate only the translation part (3D-VME) that equals :

$$T = \sum_M (M_p - (R.M_c)) \;/\; |M| \quad (9)$$



(c) Acquisition model (rectilinear rig).



(d) Iterative matching. Only matches are represented. The algorithm inspects all combinations in practice.

This first estimate allows to update the overlapping areas and optionally run an ICP-like [11] motion estimation algorithm to refine the result. Once the elementary motion is computed, it can be integrated to the previous localization to get the current one. Here, a simple integration is done, but we could consider a Kalman filter.

4. ACQUISITION CAMPAIGN

4.1. Reference localization

Precise localization of the stereo bench at each acquisition is a key measurement to evaluate in details visual odometry implementations. We have designed a localized reference frame linked to the stereo bench. It is composed of four targets which positions, measured by a laser tracker, determine its localization 1(b). A preliminary alignment step computes the rigid transformation from this frame to the stereo bench frame. We expect a localization accuracy of the stereo bench of 0.05° in orientation and 0.5mm in position. This guarantees a good evaluation of the algorithm at each elementary step and as a whole. It also provides a reliable information that is used to tune algorithm parameters.

4.2. Acquisition environment

Acquisitions were done on the place of Serom at CNES (the french space agency). It is a testing ground that models the surface we expect to encounter on Mars. We planned trajectories on the area composed of small grain 1(b).

The stereoscopic bench is mounted on a tripod at this stage of the study. It was placed at 800mm height looking 70° down. The focal length is 8mm and the baseline length is 50mm. The length of elementary displacements is between 150mm and 250mm, i.e. 40% to 60% image overlapping. Sensors are composed of 1280 by 1024 cells on a surface of 7.68mm by 6.144mm. Images are subsampled by 4 to get working images of 320 by 256 pixels. We recorded three trajectories of respectively 6m, 10m and 3.5m. The last one is composed of two acquisitions by step. A strong shadow is artificially projected on the scene of second acquisitions. We can, then, build three scenarios from it : without shadows, with shadows and mix of both. The weather was cloudy. This is an advantage to tune exposure parameters. Yet, a sunny weather is expected to provide better landmarks given by high contrast.

5. EVALUATION

The evaluation aims at answering three questions. First, what is the accuracy we can expect from such a visual odometry system in an unknown rough environment ? Here, we are interested in short term (around 2m) and long term trajectories (around 100m). Next, what is the robustness of the method with respect to localization errors ? What are the performance differences between proposed options ? We now present evaluation criteria, the noise model, next experiments.

5.1. Criteria

The performance evaluation implies two criteria. First, overall performances in position are presented. Next, the usage of system resources is provided. Different system configurations are compared in the presence of various noise levels.

Overall performances are evaluated by providing the relative error in position. In this paper, we focus on the accuracy of short and long term localizations. We then evaluate system performances for trajectories of 2m length. It represents the distance between two path forward planning operations. In the long term scenario, we aim at localizing the rover during a one day trip, i.e. 100m. The localization is optionally updated in orientation by an external sensor. Long term trajectories are artificially constructed by integrating existing elementary motions. In order to avoid loops and forward/backward motions, we conduct path integration to get a rectilinear trajectory. This approach allows to have long trajectories, yet the reference localization looses accuracy due to dead reckoning.

The consumption of system resources is measured on an Intel Pentium 4, 3.20GHz, 1Go RAM, Linux 2.6.17-5mdv, KDE 3.5.4. We will extrapolate figures to give an approximate idea of the time consumption on Leon. This is the machine type that should be loaded on ExoMars.

5.2. Noise model

We introduce noise in input localizations to simulate the noisy outputs of the attitude sensor and wheel odometry. Indeed, the proposed implementation uses an initial guess of the elementary displacement, first to compute overlapping areas in previous and current images, next to match landmarks. Moreover, one option is to estimate with visual odometry, only the translation, and keeping the rotation provided by another system.

Noise in position, n_t, is a constant in percent that introduces a constant relative error in translation. Let $\vec{T}$ be the translation between two acquisitions, then $\vec{T'} = (1 + n_t)\vec{T}$. It models wheel sliding in the motion direction. Typical values are around 10% and extreme values reach 50%. The impact of this error is the imprecise computation of overlapping areas that may result in matching difficulties. One solution is to ignore the translation component of the input and work on whole images. It will work if overlapping areas are of 40% at least.

Noise in orientation from an attitude sensor is more complex to establish a model. We assume that three parameters are sufficient to characterize the error : an instantaneous measurement error θ_m, a bias error θ_b that evolves every 100s (˜ displacement of 2m at 0.02m/s) : $\mathcal{N}(0, \sigma_b)$, and a bias stability θ_s, similar to random walk, that evolves every 10s (˜ displacement of 200mm at 0.02m/s) : $\mathcal{N}(0, \sigma_s)$. At $t = I \times 100s + J_I \times 10s$ and assuming a rotation around a fix axis, the error is, then :

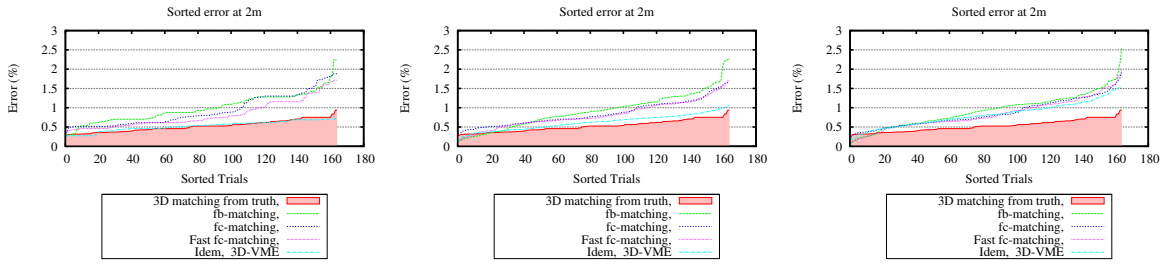
$$\theta = \theta_m + \sum_{i}^{I} \sum_{j}^{J_i} (\theta_{bi} + \sum_{k}^{j} \theta_{sk}) \tag{10}$$

This model assumes the attitude is obtained by integration of the angular velocity and the bias is regularly estimated. In experiments, this model is generalized to rotations around three axis attached to the sensor. We have retained $\sigma_m = 0.1°$, $\sigma_b = 0.01°$ and $\sigma_s = 0.01°$ as parameters of noise of small amplitude. These values are based on expected performances of the IMU : gyroscopic bias stability and bias accuracy of $0.002°/s$ with a 10sec averaging time. We also consider noise of moderate amplitude with proposed values multiplied by two. Thus, we take into account the uncertainty on noise model and variations of rover speed. The noise has a weak impact on short term localization, but it becomes more important on long trajectories without a regular global calibration as we will see in experiments. This model then completed by a short or long term calibration thanks to absolute references such as gravity and sun.

5.3. Experiments

Errors inherent to the perception chain and the detection of Harris' corners are present. They result in unavoidable 3D reconstruction uncertainties of matched landmarks. In this paper, we want to evaluate the system capacity to select best matches, yet we want to give an idea of ideal performances that could be targeted. For this purpose, we take as reference experiment the algorithm that matches landmarks by a 3D closest search of registered corners. The registration is achievable since we have a step by step accurate localization of the stereoscopic bench. This reference experiment is denoted 3D matching from truth. Its performances are limited by the actual perception chain.

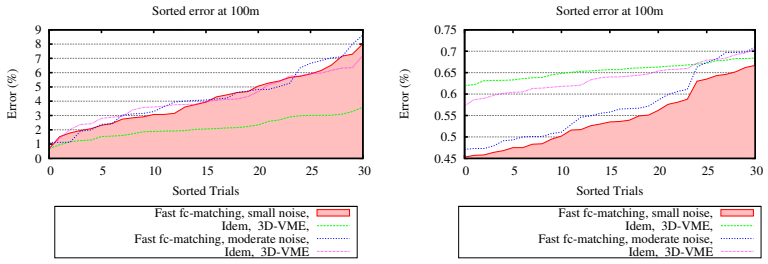
Fig. 1 shows VME performances on trajectories of 2m for each trial. Proposed matching options reveal efficient, allowing a gain in accuracy. 3D-VME seems the best choice at proposed noise levels. Translation noise was set to 10%. The current algorithm encounters difficulties over 20% since computed overlapping areas are not truly overlapping.



(e) No noise (f) Small noisy input (g) Moderate noisy input

Figure 1. Short term (2m) VME performances.

Fig. 2 shows VME performances on trajectories of 100m. Two scenarios are proposed. In the first one, we let the attitude sensor drift. In the second one, the attitude sensor is calibrated every 2m. As expected, localization accuracy is weak if we don't have a regular access to an absolute reference. In this latter case, its is preferable to estimate only the translation between acquisitions. This choice is confirmed by the step by step evaluation of the error in orientation of 6D-VME. The precision of VME in orientation is poor compared to what can be expected from an attitude sensor. We have a standard deviation of 0.2° in pitch and yaw, and 0.1° in roll (around optical axis). Mean values are small, close to 0.01°. The distribution of the translation error of 3D-VME can be modeled by a Gaussian centered around 2mm with a standard deviation of 1mm. Now, these values are specific to the presented experiments.



(a) No calibration (b) Orientation calibration every 2m

Figure 2. Long term (100m) VME performances.

5.4. Resources

Memory requirements fluctuate around 10MB for the perception chain (down sampling, rectification and stereoscopy) and around 7MB for the VME algorithm. The table 1 shows the execution time of tasks necessary to visual motion estimation. Times on Linux are obtained from experiment runs, while times on Leon are extrapolated. The CNES provided reference times on Linux and Leon of the navigation chain (perception, building of the digital elevation model). The critical subsystems are the stereo and the matching, all the more so FPGA implementation of the stereo and landmark detection is possible [12]. Fast matching leads to an important saving of computation time, however it may fluctuates with matching difficulties.

6. CONCLUSION

We presented a visual odometry system and its localization accuracy for a given configuration of the stereoscopic bench. The major weakness of VME is the orientation estimation. So, it is more appropriate to use the orientation provided by an external sensor at proposed noise levels. Localization accuracy is very good on long trajectories (one day trip) as far as a regular calibration thanks to absolute references, i.e. gravity or sun, is done. Otherwise, the localization accuracy is at the limit of the expected accuracy (5%) for the ExoMars mission. In addition, further experiments showed that a regular calibration allows to limit the effect of the instantaneous noise θ_m.

Table 1. Execution time of VME components. Gray cell points out the reference value used to extrapolate execution time on Leon. Gray values are extrapolated. They do not take into account memory access latencies on Leon.

System	Sub-system	Time on Linux (ms)	Time on Leon (ms)
Navigation		403	9340
Perception	Down sampling by 4	10	232
	Rectification	21	487
	Stereo	270	6258
VME	Landmarks	50	1159
	Geom. Match.	200	4635
	Fast fc-Matching	100	2318
Totals		554 or 454	12841 or 10524

The work plan is for now to better characterize the localization accuracy of VME with respect to configurations compatible with the ExoMars mission. The robustness of the proposed method needs also to be confirmed on more acquisitions and in motion.

7. ACKNOWLEDGMENT

This study was conducted in the robotics laboratory DCT/SB/VS of CNES at Toulouse. We would like to thanks all the staff that allowed the achievement of this work.

REFERENCES

[1] Il-Kyun JUNG. *Simultaneous Localization and Mapping in 3D environments with stereovision*. PhD thesis, LAAS (France), 2004. 1

[2] L. Matthies and Steven Shafer. Error modeling in stereo navigation. *IEEE Journal of Robotics and Automation*, 3(3):239 – 250, June 1987. 2

[3] S. Se, T. Barfoot, and P. Jasiobedzki. Visual Motion Estimation and Terrain Modeling for Planetary Rovers. In *Proceedings of the International Symposium on Artificial Intelligence, Robotics and Automation in Space*, volume 603 of *ESA Special Publication*, August 2005. 2

[4] Cheng YANG, Mark W. MAIMONE, and Larry MATTHIES. Visual odometry on the mars exploration rovers. *Proceedings of the IEEE International Conference on robotics and automation*, 13(2):54–62, 2006. 2

[5] A. Mallet, S. Lacroix, and L. Gallo. Position estimation in outdoor environments using pixel tracking and stereovision. *Proceedings of the IEEE International Conference on Robotics and Automation*, 4:3519–3524 vol.4, 2000. 2

[6] H. Hirschmuller, P.R. Innocent, and J.M. Garibaldi. Fast, unconstrained camera motion estimation from stereo without tracking and robust statistics. *Proceedings of the International Conference on Control, Automation, Robotics and Vision*, 2:1099–1104 vol.2, Dec. 2002. 2

[7] Thomas Lemaire, Cyrille Berger, Il-Kyun Jung, and Simon Lacroix. Vision-based slam: Stereo and monocular approaches. *Int. J. Comput. Vision*, 74(3):343–364, 2007. 2

[8] David G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision*, pages 1150–1157, 1999. 3

[9] C. Harris and M.J. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–152, 1988. 3

[10] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Trans. Pattern Anal. Mach. Intell.*, 9(5):698–700, 1987. 4

[11] D. Chetverikov, D. Svirko, D. Stepanov, and Pavel Krsek. The trimmed iterative closest point algorithm. In *Proceedings of the International Conference on Pattern Recognition*, page 30545, Washington, DC, USA, 2002. IEEE Computer Society. 5

[12] A. Benedetti and P. Perona. Real-time 2-d feature detection on a reconfigurable computer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–593, 1998. 7