In Proceedings of the 8th ESA Workshop on Advanced Space Technologies for Robotics and Automation
'ASTRA 2004' ESTEC, Noordwijk, The Netherlands, November 2 - 4, 2004

High resolution 3D terrain mapping with low altitude imagery

Simon Lacroix and Il-Kyun Jung
LAAS-CNRS
7, Av. du Colonel Roche
31077 Toulouse Cedex 4 France
Simon.Lacroix@laas.fr

Abstract

This paper presents an approach to build a very high resolution digitial terrain model with the imagery provided by a stereovision bench mounted on-board a balloon flying at low altitude. The approach is a "simultaneous localization and mapping" procedure, that enables to build a spatially consistent map, by allowing to recover the balloon position with a very high accuracy: only imagery is used for that purpose, and no position sensor is required. Results show the possibility to build a 5cm resolution digital elevation map of several thousands square meters areas from images acquired below 100m altitude.

1 Introduction

In the context of its field robotics activities, the Robotics and AI group of LAAS/CNRS is developping an autonomous blimp project. The long term objective is to tackle the various issues raised by the deployment of air/ground autonomous robots ensembles, in the context of exploration, surveillance and intervention missions.

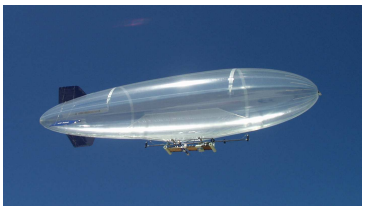

Figure 1: *The autonomous airship Karma. Note the two cameras mounted on the front and rear of the gondola.*

Aerial robots are essentially useful to gather precise information on the overflown environment: such information can help operators to analyse the terrain and detect potential areas of interest for instance, but are also very useful for ground rovers, by providing them a global model of the environment they are evolving in. For these purposes, the information gathered through the aerial imagery has to be structured into a spatially consistent model of the environment: a digital terrain map (DTM) is a well suited model, as it can faithfully represent the environment geometry, as well as color and texture information.

This paper summarizes our recent work on DTM building from low altitude stereovision imagery. It especially insists on the way the stereovision bench pose is recovered, using a "simultaneous localisation and mapping approach" that does not require any additional localization sensor. The next section presents the principle of the approach. A section then summarizes the basic vison algorithms on which we rely, and section 4 presents the implementation of a Kalman filter to solve the localization problem. Localization results and DTM built are presented in section 5.

2 Principle of the approach

Besides 3D data acquisition, the main difficulty to build a terrain model that gathers a set of data acquired during motion is to have a precise estimation of the sensor position: if this position is not precisely known, the built model is eventually distorted and contains discrepancies. A huge amount of work has of course been devoted to the localisation problem in robotics. In the absence of any external abosulte reference, the only way to guaranty a sound position esti-

mate during motions is to rely on environment features, that are detected and localized as they are perceived by the robot. This approach is known as "simultaneous localization and mapping" ("SLAM" - see *e.g.* [1, 9]): the robot position is concurently estimated with the position of *landmarks* that are detected by the robot exteroceptive sensor. The usual SLAM procedure is depicted in figure 2.

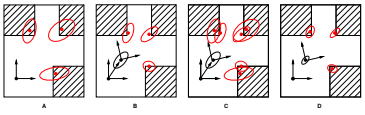

Figure 2: *Illustration of the SLAM problem with a robot evolving on a plane: (a), environment features (or landmarks) are perceived and located by an exteroceptive sensor with an uncertainty (observation). In (b), the robot moves, and estimates its position change via an other mean (e.g. odometry), with an associated uncertainty (prediction). The landmarks are re-observed. Thanks to matches of the landmarks perceived from the two positions (c: data association), the uncertainties on both the robot on the features can be reduced (d: fusion). The process goes on, and incrementally estimates the robot and landmarks position errors*

The approach presented here uses only stereovision to estimates both the robot motions (prediction) and the landmark positions (observation) [6]. Landmarks are *interest points*, *i.e.* visual features that can be matched when perceived from various positions, and whose 3D coordinates are provided by stereovision. We use an extended Kalman filter (EKF) as the recursive filter: the state vector of the EKF is the concatenation of the stereo bench position (6 parameters) and the landmark's positions (3 parameters for each landmark). The key algorithm that allows both motion estimation between consecutive stereovision frames (prediction) and the observation and matching of landmarks (data association) is a robust interest point matching algorithm [5].

The various algorithmic stages achieved every time a stereovision image pair is acquired are the following:

1. Stereovision: a dense 3D image is provided by stereovision (section 3.1).

2. Interest points detection and matching between consecutive frames (section 3.2).

3. Landmark selection: a set of selection criteria are applied to the matched interest points, in order to partition them in three sets: an a non-landmark set, a candidate-landmarks set and observed-landmark set (section 4.2).

4. Visual motion estimation (VME): the interest points retained as "non-landmarks" are used to estimate the 6 motion parameters between the previous and current steps (section 3.3).

5. Update of the Kalman filter state (section 4).

Finally, after every SLAM cycle defined by these steps, a digital elevation map is updated with the acquired images (section 5.2).

Step 3 is necessary for two reasons: first, only non-landmarks points should be used to to estimate the local motion, in order to de-correlate the prediction and update steps of the Kalman filter and second, new landmarks should be cautiously added to the filter state, in order to avoid a rapid growth of its dimension and to obtain a regular landmark coverage of the perceived scenes.

3 Basic algorithms

3.1 Stereovision

We use a classical pixel-based stereovision algorithm, that relies on an off-line calibrated binocular stereovision bench (figure 3). A dense disparity image is produced thanks to a correlation-based pixel matching algorithm, false matches being filtered out thanks to a reverse correlation. The 3D coordinates of the matched pixels are determined, with an associated uncertainty whose computation is depicted in section 4.1.

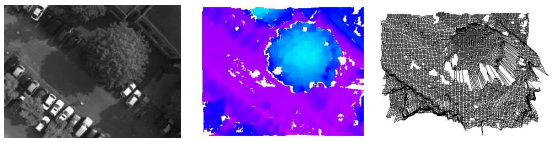

Figure 3: *A result of the stereovision algorithm, with an image pair taken at about* 30 *m altitude. From left to right: one of the original image, disparity map (shown here in a blue/close red/far color scale), and 3D image, rendered as a mesh for readability purposes. Pixels are properly matched in all the perceived areas, even the low textured ones.*

3.2 Interest points detection and matching

Visual landmarks must be invariant to image translation, rotation, scaling, partial illumination changes and viewpoint changes. *Interest points*, such as detected by the popular Harris detector, has proven to have good stability properties [8]. When a there is prior knowledge on the scale change, even approximate, a scale adaptive version of Harris detector yields a repeatability high enough to allow robust matches [2].

To match interest points, we use a matching algorithm that relies on local interest point groups matching, imposing a combination of geometric and signal similarity constraints, thus being more robust than approaches solely based on local point signal characteristics (details can be found in [5]). Figure 4 shows that this interest point matching algorithm generates a lot of good matches, even when the view point change between the considered images is quite high.

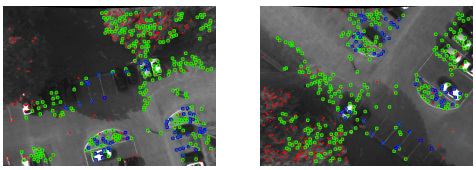

Figure 4: *A result of our interest point matching between two non registered aerial images*

3.3 Visual motion estimation

The interest points matched between consecutive images and the corresponding 3D coordinates provided by stereovision are used to estimate the 6 displacement parameters between the images, using the least square minimization technique presented in [3]. The important point here is to get rid of the outliers (wrong matches). The outliers could be rejected using a robustly estimated fundamental matrix computed on the basis of the matches, but this would not cope for stereovision errors, such as the ones that occur along depth discontinuities: inlier matches in the image plane might become outliers when considering the corresponding 3D coordinates.

Therefore, we use an outlier rejection method that consider both matching and stereovision errors. First, matches that imply a 3D point whose coordinates uncertainties are over a threshold are discarded (the threshold is empirically determined by statistical analysis of stereovision errors). Then, the remaining matches are analyzed according to the following procedure: A 3D transformation is determined by least-square minimization. The mean and standard deviation of the residual errors are computed, and aA threshold is defined as k times the residual error standard deviation: k should be at least greater than 3. The 3D matches whose error is over the threshold are then eliminated, k is set to $k-1$ and the procedure is re-iterated until $k = 3$.

This outlier rejection algorithm yields a precise 3D motion estimation between consecutive stereovision frames (see results in sections 4.1 and 5.1), which is used for the prediction stage of the Kalman filter.

4 Kalman filter setup

The EKF is an extension of the standard linear Kalman filter, that linearizes the nonlinear prediction and observation models around the predicted state. The discrete nonlinear system and the observations are modeled as:

$$x(k+1) = \mathbf{f}(x(k), u(k+1)) + v(k+1)$$

$$z(k) = \mathbf{h}(x(k)) + w(k)$$

where $u(k)$ is a control input and v, w are vectors of temporally uncorrelated observation errors with zero mean and covariance $\mathbf{P}_v(k)$, $\mathbf{P}_w(k)$.

In our approach, the state of the filter is composed of the 6 position parameters $\mathbf{x}_p = [\phi, \theta, \psi, t_x, t_y, t_z]$ of the stereovision bench and of a set of N landmarks 3D coordinates $\mathbf{m}_i = [x_i, y_i, z_i]$, $0 < i \leq N$:

$$\mathbf{x}(k) = [\mathbf{x}_p, \mathbf{m}_1 \cdots \mathbf{m}_N]$$

The associated state covariance has the following form:

$$\mathbf{P}(k) = \begin{bmatrix} \mathbf{P}_{pp}(k) & \mathbf{P}_{pm}(k) \\ \mathbf{P}_{pm}{}^T(k) & \mathbf{P}_{mm}(k) \end{bmatrix}$$

where $\mathbf{P}_{pp}$ represents the stereo bench pose covariance, $\mathbf{P}_{mm}$ the landmark covariance and $\mathbf{P}_{pm}$ the cross-covariance between the bench pose and landmark estimates. In the Kalman filter framework, the state estimation encompasses three stages: prediction, observation and update of the state and covariance estimates.

Prediction. Under the assumption that landmarks are stationary, the state prediction is:

$$\hat{\mathbf{x}}(k+1 \mid k) = \mathbf{f}(k)(\hat{\mathbf{x}}(k), \mathbf{u}(k+1))$$

where $\mathbf{u}(k+1) = (\Delta\phi, \Delta\theta, \Delta\psi, \Delta t_x, \Delta t_y, \Delta t_z)$ is the visual motion estimation result between k and $k+1$ positions. The associated state covariance prediction is written as:

$$\begin{aligned}\mathbf{P}_{pp}(k+1 \mid k) = \nabla_p \mathbf{f}(k) \mathbf{P}_{pp}(k) \nabla_p \mathbf{f}^T(k) + \\ \nabla_u \mathbf{f}(k) \mathbf{R}_u(k) \nabla_u \mathbf{f}^T(k) + \mathbf{P}_v(k+1)\end{aligned} \quad (1)$$

$$\mathbf{P}_{pm}(k+1 \mid k) = \nabla_p \mathbf{f}(k) \mathbf{P}_{pm}(k)$$

$$\mathbf{P}_{mm}(k+1 \mid k) = \mathbf{P}_{mm}(k)$$

where $\mathbf{R}_u$ represents *the error covariance of the visual motion estimation result*. Note that the covariance of landmarks is not changed in the prediction stage.

Observation. When observing the i^{th} landmark, the observation model and the Jacobian of the observation function are written as:

$$\hat{\mathbf{z}}_i(k+1 \mid k) = \mathbf{h}_i(k)(\hat{\mathbf{x}}(k+1 \mid k))$$

$$\nabla \mathbf{h}_i(k) = [\nabla_p \mathbf{h}_i(k), 0 \cdots 0, \nabla_{m_i} \mathbf{h}_i(k), 0 \cdots 0]$$

where $\mathbf{h}_i(k)(\hat{\mathbf{x}}(k+1 \mid k))$ is a function of the predicted robot state and the i^{th} landmark in the state vector of the filter, which maps the state space into the observation state. The innovation and the associated covariance is written as:

$$\boldsymbol{\nu}_i(k+1) = \mathbf{z}_i(k+1) - \hat{\mathbf{z}}_i(k+1 \mid k) \quad (2)$$

$$\mathbf{S}_i(k+1) = \nabla \mathbf{h}_i(k) \mathbf{P}(k+1 \mid k) \nabla \mathbf{h}_i{}^T(k) + \mathbf{R}_i(k+1) \quad (3)$$

where $\mathbf{R}_i$ represents *the error covariance of ith landmark observation*.

Update. The update stage fuses the prediction and the observation to produce and estimate of the state and its associated covariance, according to the following formulas:

$$\hat{\mathbf{x}}(k+1 \mid k+1) = \hat{\mathbf{x}}(k+1 \mid k) + \mathbf{K}_i(k+1) \boldsymbol{\nu}_i(k+1) \quad (4)$$

$$\mathbf{P}(k+1 \mid k+1) = \mathbf{P}(k+1 \mid k) - \mathbf{K}_i(k+1) \mathbf{S}_i(k+1) \mathbf{K}_i^T(k+1) \quad (5)$$

in which $\mathbf{K}_i(k+1) = \mathbf{P}(k+1 \mid k) \nabla \mathbf{h}_i^T(k) \mathbf{S}_i^{-1}(k+1)$ is the Kalman filter gain matrix. If no observation are made (*i.e.* if no already mapped landmarks are re-perceived), the observation and update stages are not activated: the state and its covariance are just updated by the prediction stage.

When detecting a new landmark, it is added to the state vector of the filter, that becomes $\hat{\mathbf{x}}(k) = [\hat{\mathbf{x}}_p(k), \hat{\mathbf{m}}_1(k) \cdots \hat{\mathbf{m}}_N(k), \hat{\mathbf{m}}_{N+1}(k)]$. The landmark initialization model is:

$$\hat{\mathbf{m}}_{N+1}(k) = \mathbf{g}(k)(\hat{\mathbf{x}}_p(k), \mathbf{z_{N+1}}(k)) \quad (6)$$

$$\mathbf{P}(k) = \begin{bmatrix} \mathbf{P}_{pp}(k) & \mathbf{P}_{pm}(k) & \mathbf{P}_{pz}(k)^T \\ \mathbf{P}_{pm}{}^T(k) & \mathbf{P}_{mm}(k) & \mathbf{P}_{mz}(k)^T \\ \mathbf{P}_{pz}(k) & \mathbf{P}_{mz}(k) & \mathbf{P}_{zz}(k) \end{bmatrix} \quad (7)$$

$$\mathbf{P}_{pz}(k) = \nabla_p \mathbf{g}(k) \mathbf{P}_{pp}(k),\ \mathbf{P}_{mz}(k) = \nabla_p \mathbf{g}(k) \mathbf{P}_{pm}(k)$$

$$\mathbf{P}_{zz}(k) = \nabla_p \mathbf{g}(k) \mathbf{P}_{pp}(k) \nabla_p \mathbf{g}^T(k) + \nabla_z \mathbf{g}(k) \mathbf{R}_m(k) \nabla_z \mathbf{g}^T(k)$$

where $\mathbf{z_{N+1}}(k)$ denotes the new landmark, $\mathbf{g}(k)$ represents the initialization function using the current robot pose estimate and $\mathbf{R}_m$ is *the error covariance of the new landmark*.

4.1 Errors identification

Error identification is crucial to set up a Kalman filter, as a precise determination of these errors will avoid the empirical "filter tuning" step. In our context, the following errors must be estimated:

- the landmark initialization error ($\mathbf{R}_m$),
- the landmark observation error ($\mathbf{R}_i$ for the observed landmark i),
- and the error of the input control u, which is the visual motion estimation result ($\mathbf{R}_u$).

Note that in our approach, the lumped process noise v is set to 0, landmarks being stationary and the robot pose prediction being directly computed with the current pose and the result of the visual motion estimation.

Landmark initialization errors. Landmarks are detected and matched on the video images, their 3D coordinates being computed by stereovision: the covariance matrix $\mathbf{R}_m$ on landmarks is totally defined by the stereovision error.

Statistics on image pairs acquired from the same position show that the distribution of the disparity computed on any given pixel can be well approximated by a Gaussian [7], and that there is a *strong correlation* between the shape of the similarity score curve around its peak and the standard deviation on the disparity: the sharper the peak, the more precise the disparity (figure 5). This relation is the basis of our error model: on line, during the stereo matching phase, a standard deviation σ_d is associated to each computed disparity d, using the curvature of the similarity score curve at its peak (in the matching phase, this curve is approximated at its peak by fitting a parabola to determine a sub-pixellic estimate of the disparity).

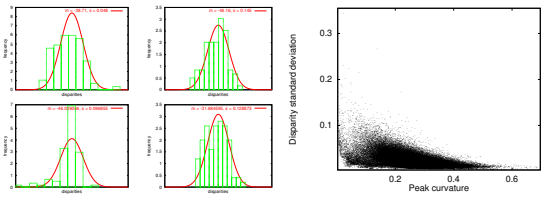

Figure 5: *Left: examples of some probability density functions of disparities computed on a set of 100 image pairs, with the corresponding Gaussian fit. Right: Standard deviation of the disparities as a function of the curvature of the similarity score curve at its peak.*

Once matches are established, the coordinates of the 3D points are computed with the usual triangulation formula: $z = \frac{b\alpha}{d}$, $x = \beta_u z$ and $y = \gamma_v z$, where z is the depth, b is the stereo baseline, and α, β_u and γ_v are calibration parameters (the two latter depending on (u, v), the considered pixel image coordinates). Using a first order approximation, it comes:

$$\sigma_z = \frac{\sigma_d}{b\alpha} z^2$$

The covariance matrix of the point coordinates is then derived from the triangulation equations. When a new landmark is observed, its coordinates are added to the filter state, and the state covariance is updated according to equations (6) and (7).

Observation error. In our case, landmark observation is based on interest point matching. Outliers being rejected (section 3.3), only interest point location errors are considered to determine the matching error.

With the precise Harris detector, the sub-pixellic coordinates of an interest point $\mathbf{p}$ belongs to one pixel. When perceived again from a very close point of view (*e.g.* in two consecutive images), most of the area corresponding to this pixel is mapped into an other pixel, and the matched interest point $\mathbf{p}'$ coordinates lie within this pixel. The expected matching error is therefore of the order of 0.5 pixel. But when the viewpoint is very different (*e.g.* when re-perceiving a landmark after a long loop), the projective deformation of the 3D scene and possible occlusions effects are much more important. In such cases, the expected matching error value is then set to 1 pixel, which is consistent with the maximum error of 1.5 pixel generally tolerated to assess good matches [8].

The interest point matching error is combined with the errors on the corresponding 3D estimates to define the observation error. The principle of this combination is illustrated in figure 6: the observation matching error is defined by the reprojection of the matching error in the 3D scene. When the 2D matching error is set to 1 pixel, the expected value of a matching point $\mathbf{p}_0$ is defined by its 8 closest neighbors $\mathbf{p}_k$, $k = 1, 2, , , 8$. The stereo error distribution being a zero mean normal one, the expected 3D coordinate and associated variance of the matching point is computed as follows:

$$\bar{\mathbf{X}} = \frac{1}{9}\sum_{i=0}^{8} \mathbf{X}_i, \qquad \boldsymbol{\sigma}^2_{\bar{\mathbf{X}}} = \frac{1}{9}\sum_{i=0}^{8} (\bar{\mathbf{X}} - \mathbf{X}_i)^2 + \boldsymbol{\sigma}_i^2$$

where $\mathbf{X}_0$ and $\mathbf{X}_k$ are the 3D point coordinates of $\mathbf{p}_0$ and its neighbors, and $\boldsymbol{\sigma}_0$ and $\boldsymbol{\sigma}_k$ are the corresponding variances.

When the expected matching error is set as 0.5 pixel, the 3D coordinates being only computed on integer pixels by stereovision, we assume the 3D surface variation is locally linear, and the expected 3D coordinate and corresponding variances of the observed point $\mathbf{p}_0$ are then:

$$\bar{\mathbf{X}} = \frac{1}{9}(\mathbf{X}_0 + \sum_{i=1}^{8} (\frac{\mathbf{X}_0 + \mathbf{X}_i}{2}))$$

$$\boldsymbol{\sigma}^2_{\bar{\mathbf{X}}} = \frac{1}{9}\sum_{i=0}^{8} (\frac{\bar{\mathbf{X}} - \mathbf{X}_i}{2})^2 + (\frac{\boldsymbol{\sigma}_0 + \boldsymbol{\sigma}_i}{2})^2$$

These coordinates and the associated variances are used in the equations (2) and (3).

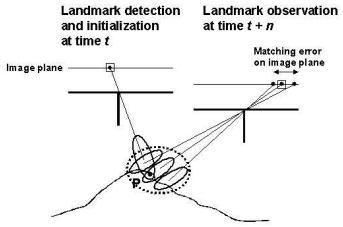

Figure 6: *Principle of the combination of the matching and stereovision errors. The points located in the square box are the projection of P on the image plane. Small ellipses indicate stereovision errors, the large dotted ellipsoid is the resulting observation error.*

Motion estimation errors. Given a set of 3D matched points $\hat{\boldsymbol{\mathcal{Q}}} = [\mathbf{X}_1, , , \mathbf{X}_N, \mathbf{X}_1, , , \mathbf{X'}_N]$, the function which is minimized to determine the corresponding motion is the following [3]:

$$J(\hat{\mathbf{u}}, \hat{\boldsymbol{\mathcal{Q}}}) = \sum_{n=1}^{N} (\mathbf{X'}_n - R(\hat{\phi}, \hat{\theta}, \hat{\psi})\mathbf{X}_n - [\hat{t}_x, \hat{t}_y, \hat{t}_z]^T)^2$$

where $\hat{\mathbf{u}} = (\hat{\phi}, \hat{\theta}, \hat{\psi}, \hat{t}_x, \hat{t}_y, \hat{t}_z)$. $\hat{\mathbf{u}}$ and $\hat{\boldsymbol{\mathcal{Q}}}$ can be written with random perturbations:

$$\hat{\mathbf{u}} = \mathbf{u} + \Delta\mathbf{u}, \qquad \hat{\boldsymbol{\mathcal{Q}}} = \boldsymbol{\mathcal{Q}} + \Delta\boldsymbol{\mathcal{Q}}$$

where the true $\mathbf{u}$ and $\boldsymbol{\mathcal{Q}}$ are not observed. In order to measure the uncertainty of local motion estimation, the uncertainties of 3D matching points set are propagated to the optimal motion estimate $\hat{\mathbf{u}}$. Assuming the optimal motion estimate extremizes the cost function and its Jacobian, the uncertainties of landmarks $\mathbf{X}_n$ and their observation $\mathbf{X}'_n$can be propagated by taking Taylor series expansion of the Jacobian around $\mathbf{u}$ and $\boldsymbol{\mathcal{Q}}$, as shown in [4]: considered that $\mathbf{X}_n$ and $\mathbf{X'}_n$ are not correlated, the covariance estimate $\mathbf{P}_{\hat{\mathbf{u}}}$ can be also written as:

$$\mathbf{P}_{\hat{\mathbf{u}}} = (\frac{\partial g}{\partial \mathbf{u}}(\hat{\mathbf{u}}, \hat{\boldsymbol{\mathcal{Q}}}))^{-1}(\Lambda_{\mathbf{X}} + \Lambda_{\mathbf{X}'})(\frac{\partial g}{\partial \mathbf{u}}(\hat{\mathbf{u}}, \hat{\boldsymbol{\mathcal{Q}}}))^{-1}$$

where $g = \frac{\partial J}{\partial \hat{\mathbf{u}}}$ is the Jacobian of the cost function and

$$\Lambda_{\mathbf{X}} = \sum_{n=1}^{N} \frac{\partial g}{\partial \mathbf{X}_n}(\hat{\mathbf{u}}, \mathbf{X}_n)\mathbf{P}_{\mathbf{X}_n}(\frac{\partial g}{\partial \mathbf{X}_n}(\hat{\mathbf{u}}, \mathbf{X}_n))^T$$

$$\Lambda_{\mathbf{X}'} = \sum_{n=1}^{N} \frac{\partial g}{\partial \mathbf{X'}_n}(\hat{\mathbf{u}}, \mathbf{X'}_n)\mathbf{P}_{\mathbf{X'}_n}(\frac{\partial g}{\partial \mathbf{X'}_n}(\hat{\mathbf{u}}, \mathbf{X'}_n))^T$$

$\mathbf{P}_{\hat{\mathbf{u}}} = \mathbf{R}_u$ is the input covariance matrix which is used in equation (1) to estimate the state variances during the filter prediction stage.

The mean of computed variances on the 6 motion parameters, and their dispersion are summarized in table 1. The visual motion estimation measures a few meters translations with a few centimeters accuracy, and measures rotations with a precision of the order of 0.1°, with a quite good regularity.

	Φ	Θ	Ψ	t_x	t_y	t_z
mean errors	0.098	0.089	0.037	3.6	3.8	1.1
σ of errors	0.024	0.030	0.008	1.3	1.1	0.3

Table 1: *Statistics on the estimated errors of the VME computed on about 50 local motion estimations (in degrees and centimeters).*

4.2 Landmark selection

As explained in the section 2, the 3D matches established after the interest point matching step are split into three sets. The observed-landmarks set is simply the points that corresponds to landmarks already in the state vector of the EKF. The rest of the matches are then studied, to select the set of candidate-landmarks according to the following three criteria:

• **Observability.** Good landmarks should be observable in several consecutive frames.

• **Stability.** The 3D coordinates of good landmarks must be precisely estimated by stereovision.

• **Representability.** Good landmarks must efficiently represent a 3D scene. The stereovision bench state estimation will be more stable if landmarks are regularly dispatched in the perceived scene, and this regularity will avoid a rapid growth of the EKF state vector size. The number of candidate landmarks that are checked is determined on the basis of the number of new interest point matches (*i.e.* the ones that do not match with an already mapped landmark). We use 10 % of the new interest points, as the visual motion estimation technique requires a lot of matches to yield a precise result. The landmark selection is made according a heuristic procedure so that they satisfy the above three criteria.

5 Results

Our developments have been tested with hundreds of images taken on-board a blimp, at al-

titudes ranging from 20 to 35 m. The cameras of the 2.2 m wide stereo bench are $1/2''$ 1024×768 pixels CCD sensors, with a a 4.8 mm focal length lens.

5.1 Positioning errors

We do not have any localization mean that could be used as reference on-board the blimp (such as a centimeter accuracy GPS). However, when the blimp flies over an already perceived area, the VME can provide an estimate of the relative positions between the first and last image of the sequence that overlaps. This reference is precise enough, as compared to the cumulation of errors induced with the VME applied on consecutive frames.

Figure 1 shows the evolution of the standard deviation of the 6 position parameters of the stereo bench when applying the EKF. Two phases can be seen on this latter figure: until image 200, the standard deviation grows, however much more slowly that when propagating only the errors of the VME. A few landmarks detected in the beginning of the sequence are re-perceived in the following images: the standard deviations decreases, and stabilizes for the subsequent images where some "old" landmarks are still observed. A drastic improvment on the position uncertainties can be seen after image 350, where landmarks mapped in the beginning of the trajectory are re-perceived.

The quantitative figures summarized in table 5.1 compare the results of the final position estimate with respect to the reference defined by the VME applied between images 1 and 40: the precision enhancement brought by the EKF is noticeable, and the absolute estimated errors are all bounded by twice the estimated standard deviations. The translation errors are below 0.1 m in the three axes after an about 60 m long trajectory, and angular errors are all below half a degree.

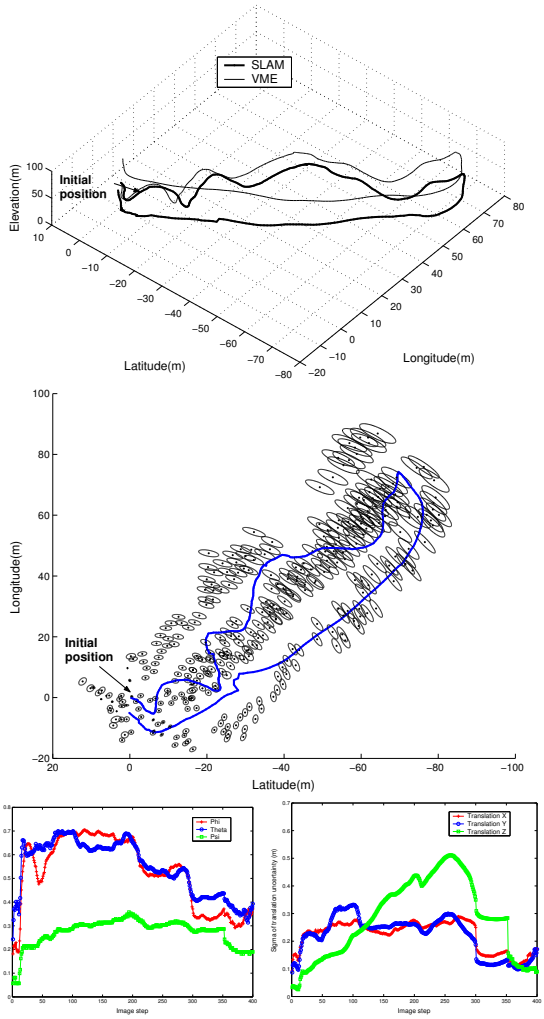

Figure 7: *A result of our SLAM implementation with a sequence of 400 stereovision pair. Top image show the reconstructed trajectory in orthogonal in 3D, the middle image show the 355 landmarks mapped, with 1σ uncertainty ellipses magnified by a factor of 40, and the two bottom figures show the evolution of the standard deviations of the blimp position parameters*

5.2 Digital elevation maps

Thanks to the precise positioning estimation, the processed stereovision images can be fused after every update of the EKF into a *digital elevation map* (DEM), that describes the environment as a function $z = f(x, y)$, determined on every cell (x_i, y_i) of a regular Cartesian grid.

Our algorithm to build a DEM simply computes the elevation of each cell by averaging the elevations of the 3D points that are vertically projected on the cell surface. Since a luminance value is associated to each 3D point produced by stereovision, it is also possible to compute a mean luminance value for each map cell. Figure 8 shows a digital elevation built from 100 images pairs, and figure 9 shows the DTM built using the 400 images used to recover the trajectory of figure 7: the resolution of the grid is here 0.1 m, and no map discrepancies can be detected in the corresponding orthoimage.

	Reference std. dev.	VME abs. error	SLAM std. dev.	SLAM abs. error
Φ	0.10°	3.30°	0.21°	0.20°
Θ	0.09°	2.40°	0.28°	0.61°
Ψ	0.04°	0.56°	0.09°	0.11°
t_x	0.04 m	0.91 m	0.11 m	0.24 m
t_y	0.04 m	1.47 m	0.08 m	0.07 m
t_z	0.01 m	0.91 m	0.04 m	0.10 m

Table 2: *Comparison of the errors made by the propagation of the visual motion estimation alone and with the SLAM EKF approach, using as a reference the VME applied between images 1 and 40*

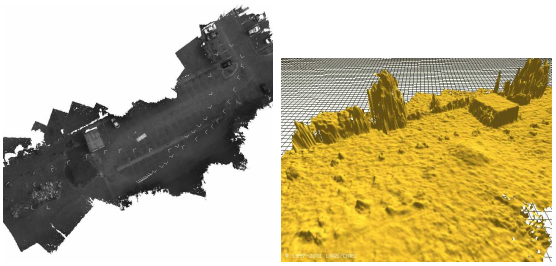

Figure 8: *The DEM computed with 100 images: orthoimage and 3D view of the bottom-left area. The map covers an area of about* 3500 m^2.

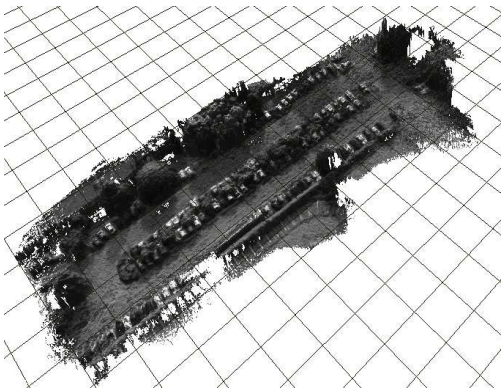

Figure 9: *A DTM computed with 400 images stereoscopic image pairs. The map covers an area of about* 6000 m^2.

6 Summary

We presented a vision-based SLAM approach that allows the building of large high resolution terrain maps. The use of interest point as landmarks allows an active selection of the landmarks to properly map the environment without any prior knowledge. Our interest point matching algorithm provides robust data associations which makes possible the matching of already mapped landmark and the precise visual motion estimation between consecutive frames. A rigorous study and identification of the various errors estimates involved in the filter allows to set it up properly, without any empirical tuning stage.

Current work aims at transferring this technique to a sequence of monocular images, with the help of a positionning sensor that provides the prediction estimate to the Kalman filter.

References

[1] G. Dissanayake, P. M. Newman, H-F. Durrant-Whyte, S. Clark, and M. Csorba. A solution to the simultaneous localization and map building (slam) problem. *IEEE Transaction on Robotic and Automation*, 17(3):229–241, May 2001.

[2] Y. Dufournaud, C. Schmid, and R. Horaud. Matching images with different resolutions. In *International Conference on Computer Vision and Pattern Recognition, Hilton Head Island, SC (USA)*, pages 612–618, Juin 2000.

[3] R. Haralick, H. Joo, C.-N. Lee, X. Zhuang, V.G. Vaidya, and M.B. Kim. Pose estimation from corresponding point data. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(6):1426–1446, Nov/Dec 1989.

[4] R.M. Haralick. Propagating covariances in computer vision. In *International Conference on Pattern Recognition*, pages 493–498, 1994.

[5] I-K. Jung and S. Lacroix. A robust interest point matching algorithm. In *8th International Conference on Computer Vision, Vancouver (Canada)*, July 2001.

[6] I-K. Jung and S. Lacroix. High resolution terrain mapping using low altitude aerial stereo imagery. In *International Conference on Computer Vision, Nice (France))*, Oct 2003.

[7] L. Matthies. Toward stochastic modeling of obstacle detectability in passive stereo range imagery. In *IEEE International Conference on Computer Vision and Pattern Recognition, Champaign, Illinois (USA)*, pages 765–768, 1992.

[8] C. Schmid, R. Mohr, and C. Bauckhage. Comparing and evaluating interest points. In *International Conference on Computer Vision*, Jan 1998.

[9] S. Thrun, D. Fox, and W. Burgard. A probabilistic approach to concurrent mapping and localization for mobile robots. *Autonomous Robots*, 5:253–271, 1998.