A Dense Structure Model for Image Based Stereo SLAM

Tommi Tykkälä and Andrew I. Comport

Abstract— **In this paper a dense structure model is developed for stereo image based Simultaneous Localization And Mapping (SLAM). It is proposed to model dense environment structure incrementally by robustly integrating disparity maps from current and previous time instants. In this way disparities can be refined over time to favor consistent 3D structure over noise. The analytical search bounds for disparities are transferred into the current map to allow efficient re-localization. The cost function is image-based and it is minimized by combining Iteratively Reweighted Least-Squares (IRLS) with exhaustive search for finding motion and disparity parameters respectively.**

I. INTRODUCTION

In the general SLAM problem, camera pose and environment structure are estimated simultaneously and incrementally in real-time using a combination of sensors. A SLAM approach is interesting in a wide range of robotics applications where a precice map of the environment does not exist or it is inconvenient to store. This work has been funded by a joint Thales-Alenia Space / CNRS project for an autonomous space exploration.

Recently camera based SLAM approaches have been an important and active area of study [9], [12]. Typically they rely on feature-based approaches, where a sparse set of points is extracted per image frame and the points are matched temporally based on their feature descriptors. The SLAM problem is then often solved by filtering motion and 3D point parameters from measured 2D motion using an Extended Kalman Filter or by performing local bundle adjustment.

Although feature descriptors have improved [11], robust SLAM systems relying purely on feature point tracking are rare. A major problem is to consistently extract and match projections of 3D points with current image features. Due to sparse environment modeling, these systems are also sensitive to occlusion. A recent improvement has been obtained by breaking the problem into one considering structure and motion separately [10].

Image based *direct* SLAM methods avoid the feature extraction process completely and base estimation on raw image data (Fig. 1). This is achieved by formulating SLAM problem as partial or full image registration task between subsequent frames. Direct methods for both planar and non-planar environment geometries have been studied [13], [14]. The main advantage of direct methods is that they minimize true error based on the actual measurements. They can also be made statistically robust due to the redundancy in information and they are scalable for real-time performance [15].

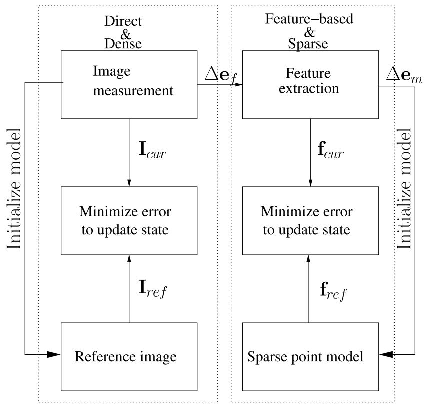

Fig. 1. The difference between direct dense method and feature-based sparse method. Feature-based method has intrinsic bias with feature extraction errors $\Delta\mathbf{e}_f$ as well as correspondence model errors $\Delta\mathbf{e}_m$.

Dense SLAM methods aim to use the entire image as measurements in order to increase robustness and accuracy in estimation [14],[16]. In addition to solving the SLAM problem, 3D reconstructions generated by a dense method are open to a wider range of applications than sparse ones, such as photometric environment mapping and Human-Computer Interaction (HCI). The main challenge of a dense method is to cope with a large number of mismatched image points efficiently in real-time.

The problem posed here is to estimate simultaneously the camera trajectory and disparities based on stereo video input in a direct and dense way. The optimal offline bundle adjustment minimization is approximated by decomposing the problem into a sequence of image-based cost optimization problems which maintains the local non-linearities of the problem. For each frame two optimization problems are solved: 1) structure optimization using spatial image registration and 2) motion parameter optimization using temporal image registration. In practise the first phase is solved by bounded exhaustive and search of disparities [4], and the second one by IRLS optimization [1].

One novelty of this method lies in the estimation of disparity maps. All past disparities of high quality are transformed into the current disparity map so that the local bundle adjustment problem is reduced to a incremental one involving only two stereo views. Due to uncertainties in motion and disparity parameters, transformed disparities are rematched in the current image. It will be shown that re-localization of the structure can be done efficiently by using search bounds defined by the analytical covariance of the cost function. This way the amplitude of the bounds varies

T. Tykkälä and Dr. A.I. Comport are with I3S UNSA CNRS, 2000 Route des Lucioles, 06903 Sophia-Antipolis Cedex, France

according to the uncertainty of the point associated.

The paper is organized in the following way. In Section II the problem is modeled mathematically, Section III presents the exhaustive disparity search technique and Section IV presents IRLS optimization of motion parameters. Section V explains how disparity maps are integrated using disparity uncertainties. Results are shown in Section VI followed by conclusions in Section VII.

II. Dense Stereo SLAM

A. Motion parameterization

The camera trajectory is defined by $\mathbf{X}_m = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$, where $\mathbf{x}_i = (\boldsymbol{\omega}_i, \boldsymbol{\upsilon}_i) \in \mathbb{R}^6$ encodes relative 3D motion between two camera poses. $\boldsymbol{\omega}$ is the rotation axis and $\|\boldsymbol{\omega}\|$ is the rotation angle in radians, and $\boldsymbol{\upsilon}$ is the velocity twist. The structure of the increment is the following:

$$\mathbf{A}(\mathbf{x}) = \begin{bmatrix} [\boldsymbol{\omega}]_\times & \boldsymbol{\upsilon} \\ \mathbf{0} & 0 \end{bmatrix} \quad e^{\mathbf{A}(\mathbf{x})} = \begin{bmatrix} \Delta\mathbf{R} & \Delta\mathbf{t} \\ 0 & 1 \end{bmatrix}, \tag{1}$$

where $\Delta\mathbf{R} \in \mathbb{SO}(3)$ and $\Delta\mathbf{t} \in \mathbb{R}^3$. The matrix exponential $e^{\mathbf{A}(\mathbf{x})}$ is used because it produces smooth and invertible transformations [8]. The right view pose matrix is $\mathbf{T}^R = \prod_{i=1}^{m} e^{\mathbf{A}(\mathbf{x}_i)}$ where multipliers come to the right. The left stereo view is obtained by $\mathbf{T}^L = \mathbf{T}^R\mathbf{T}^{RL}$, where $\mathbf{T}^{RL}$ represents the constant baseline between the stereo views. $\mathbf{T}^R$ and $\mathbf{T}^L$ are defined to map points from the local camera frame into the world reference frame.

B. Structure parameterization

General Lambertian 3D structure is represented as $3 \times n$ matrix of 3D points $\mathbf{G} = \begin{bmatrix}\mathbf{P}_1 \dots \mathbf{P}_n\end{bmatrix}$ associated with a vector of intensities $\mathbf{I}_g = \begin{bmatrix} I_1 \dots I_n \end{bmatrix}$, where $\mathbf{P}_k \in \mathbb{R}^3$, $I_k \in \mathbb{R}$ and n is the number of points. The projection of $\mathbf{G}$ into stereo view j is denoted as $4 \times n$ matrix $\mathbf{\Pi}_j = \begin{bmatrix} \mathbf{\Pi}_j^R & \mathbf{\Pi}_j^L \end{bmatrix}^T$, whose columns $\begin{bmatrix} \mathbf{p}_k^R \\ \mathbf{p}_k^L \end{bmatrix} \in \mathbb{R}^4$. In a rectified stereo view, the corresponding projection points $\mathbf{p}_k^L$ and $\mathbf{p}_k^R$ differ only in horizontal displacement $d_k \in \mathbb{R}^+$ called *disparity*. Thus $\mathbf{\Pi}_j$ is compactly represented as $3 \times n$ matrix $\mathbf{\Pi}_j^d = \begin{bmatrix} \mathbf{\Pi}_j^R & \mathbf{d}_j \end{bmatrix}^T$, where $\mathbf{d}_j$ is a column vector of disparity values. Visible $\mathbf{G}$ can be reconstructed from $\mathbf{\Pi}_j^d$ by *triangulation* [5].

A stereo image expressed as a one dimensional vector $\mathbf{I} = \begin{bmatrix} \mathbf{I}^R & \mathbf{I}^L \end{bmatrix} = \left(I_1^R \dots I_{w\times h}^R \; I_1^L \dots I_{w \times h}^L \right)$, spans the visible subset of discretized 3D geometry $\mathbf{G}$ if d_k can be estimated for each grid point $\mathbf{p}_k^R$. Occlusion, lack of texture, view-dependent reflection properties and complicated geometries can however prevent capturing $\mathbf{d}$ accurately. This is why $\widehat{\mathbf{\Pi}}^d$ may unfortunately contain an amount of non-physical geometry. The data of $\mathbf{I}^R$, $\mathbf{I}^L$, and $\mathbf{d}$ is bilinearly interpolated by associated functions $\mathcal{I}(\mathbf{p}) : \mathbb{R}^2 \Rightarrow \mathbb{R}$ and $\mathcal{D}(\mathbf{p}) : \mathbb{R}^2 \Rightarrow \mathbb{R}$.

C. Image-based motion model

When $\widehat{\mathbf{\Pi}}_j^d$ has been extracted from a stereo image, it is possible to simulate camera motion by reconstructing visible points $\mathbf{P}_k$ and projecting them into a new view (Fig. 2). This process is conveniently formalised by the *trifocal tensor* without explicitly reconstructing $\mathbf{P}_k$ [5]. If $\{\mathbf{T}^L, \mathbf{T}^R, \mathbf{T}^R e^{\mathbf{A}(\mathbf{x})}\}$ are the view matrices for the reference stereo view and a new point of view, a trifocal tensor $\begin{bmatrix} s_1(\mathbf{x}) & s_2(\mathbf{x}) & s_3(\mathbf{x}) \end{bmatrix}$, a cube of $3\times3\times3$ scalar values, can be generated, where each matrix slice is $s_j(\mathbf{x}) = \mathbf{a}_j \mathbf{b}_4^T(\mathbf{x}) - \mathbf{a}_4 \mathbf{b}_j^T(\mathbf{x})$, where $\mathbf{a}_j$ are the columns of $[\mathbf{T}^{RL}]_{3\times4}$ and $\mathbf{b}_j(\mathbf{x})$ are the columns of $[(e^{\mathbf{A}(\mathbf{x})})^{-1}\mathbf{T}^{RL}]_{3\times4}$.

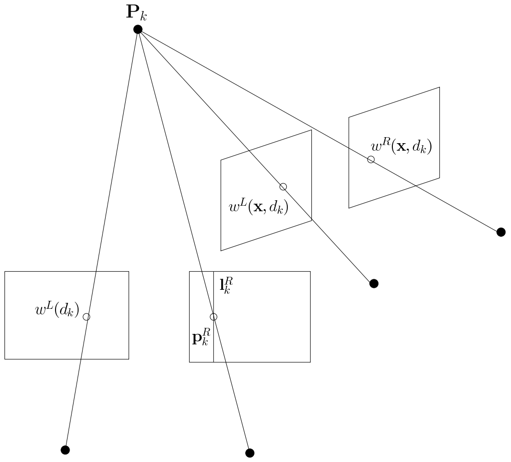

Fig. 2. The transfer of a matching point pair into the next stereo view. Intersection point $\mathbf{P}_k$ of a ray through $w^L(d_k)$ and a plane through $\mathbf{l}_k^R$ is projected into next view using trifocal tensor.

When the trifocal tensor is contracted with a point it produces a homography mapping $h_k(\mathbf{x})$, whose columns j are defined as

$$h_k^j(\mathbf{x}) = s_j^T(\mathbf{x})(\mathbf{K}^R)^T \mathbf{l}_k^R, \tag{2}$$

where $\mathbf{K}^R$ is the intrinsic matrix of the right camera and $\mathbf{l}_k^R = [1, 0, -x_k^R]^T$ defines a vertical line through point $\mathbf{p}_k^R = [x_k^R, y_k^R]^T$ [14]. For mapping all image points from a previous reference stereo view to the current view, the following two warping functions are required:

$$\begin{aligned} w^L(d_k) &= \mathbf{p}_k^R + (d_k, 0)^T & (3) \\ w^*(\mathbf{x}, d_k) &= N\left(\mathbf{K}^* h_k(\mathbf{x}) (\mathbf{K}^L)^{-1} \begin{bmatrix} w^L(d_k) \\ 1 \end{bmatrix} \right), & (4) \end{aligned}$$

where $\mathbf{K}^L$ is the intrinsic matrix of left camera, $\mathbf{K}^*$ the intrinsic matrix of target view and $N(\mathbf{p}) = (u/w, v/w)^T$ for $\mathbf{p} = (u, v, w)^T$. $w^L(d_k)$ warps points to the left stereo view and $w^*(\mathbf{x}, d_k)$ warps points under motion $\mathbf{x}$ into target stereo view $*$.

D. SLAM cost function

The dense image based objective to be minimized is

$$\mathbf{r}_d(k) = \mathcal{I}_m^L(w^L(d_k)) - \mathcal{I}_m^R(\mathbf{p}_k^R) \tag{5}$$

$$\mathbf{r}_x(k) = \mathcal{I}_{m+1}^R(w^R(\mathbf{x}_m, d_k)) - \mathcal{I}_m^R(\mathbf{p}_k^R) \tag{6}$$

$$\mathcal{O}(\mathbf{d}_m, \mathbf{x}_m) = \mathbf{r}_d^T \mathbf{W}_d \mathbf{r}_d + \mathbf{r}_x^T \mathbf{W}_x \mathbf{r}_x, \tag{7}$$

where $\mathbf{d}_m$ and $\mathbf{x}_m$ are the parameters to be estimated for frame m, and $\mathbf{r}_d$ and $\mathbf{r}_x$ are the corresponding image residuals. Both stereo views could be used to minimize motion error but in this paper only right image is used for simplicity. A few notes about how this approach relates to bundle adjustment can be found in Appendix B.

$\mathbf{W}_d$ and $\mathbf{W}_x$ are diagonal matrices given by the Tukey weighting function based on a robust statistical distribution of the associated residual. Tukey weighting is given by

$$\begin{aligned} u_k &= \frac{|r_k - median(\mathbf{r})|}{s}, & (8)\\ s &= c * median(|r_k - median(\mathbf{r})|), & (9)\\ \mathrm{w_k} &= \begin{cases} (1-(\frac{u_k}{b})^2)^2 & \text{if } |u_k| <= b \\ 0 & \text{if } |u_k| > b \end{cases}, & (10) \end{aligned}$$

where $\mathrm{w_k}$ are diagonal elements of $\mathbf{W}$, $c = 1.4826$ for robust standard deviation and $b = 4.6851$ for Tukey window [7].

Optimization is decomposed via marginalization into two separate non-linear minimization phases which are subsequentially executed for each stereo frame. This is the optimal formulation for the joint problem assuming that the initialization phase succeeds in finding close to optimal disparity parameters.

1) $\widehat{\mathbf{d}}_m = \underset{\mathbf{d}}{argmin}\ \mathbf{r}_d^T \mathbf{W}_d \mathbf{r}_d$.
2) $\widehat{\mathbf{x}}_m = \underset{\mathbf{x}}{argmin}\ \mathbf{r}_x^T \mathbf{W}_x \mathbf{r}_x$.
3) m = m + 1,

III. Disparity map initialization

The initial disparity $\widehat{d}_k$ is estimated for each $\mathbf{p}_k^R$ by performing an exhaustive search over an initially large bounded interval along the epipolar line,

$$\widehat{d}_k = \underset{d}{argmin}\ C(w^L(d_k), \mathbf{p}_k^R), d_k \in \mathbf{b}_k \quad (11)$$

$$C(\mathbf{p}^L, \mathbf{p}^R) = \sum_{\mathbf{u}_i \in \mathcal{S}} (\mathcal{I}^L(\mathbf{p}^L + \mathbf{u}_i) - \mathcal{I}^R(\mathbf{p}^R + \mathbf{u}_i))^2, \quad (12)$$

where $\mathbf{b}_k = [x_k^R, width] \in \mathbb{R}^2$ defines the 1D search region, and cost function C minimizes direct image error over the template region $\mathcal{S}$ by simple SSD. The initial guess is refined to sub-pixel accuracy by:

$$\begin{aligned} r_k &= \mathcal{I}^L(w^L(d_k)) - \mathcal{I}^R(\mathbf{p}_k^R),\ g_k = \frac{\partial r_k}{\partial d_k}, & (13)\\ \widehat{d}_k &= \widehat{d}_k + min(max(-r_k/g_k, -0.5), 0.5). & (14) \end{aligned}$$

Sub-pixel corrections are bounded to $[-0.5, 0.5]$ due to per-pixel evaluation of C. This results in $\widehat{\mathbf{d}}_0$ and the same procedure is used for initializing new points.

IV. Non-linear motion parameter estimation

$\widehat{\mathbf{x}}_m$ is obtained using IRLS optimization. Assuming sufficient frame rate, two subsequent images are similar and $\Delta\mathbf{x}_m = \mathbf{0}$ is a good initial guess. The local linearization of the cost function is

$$\begin{bmatrix} \mathbf{J}_x^T \mathbf{W}_x \mathbf{J}_x & \mathbf{J}_x^T \mathbf{W}_x \mathbf{J}_{xd} \\ \mathbf{J}_{xd}^T \mathbf{W}_x \mathbf{J}_x & \mathbf{J}_{xd}^T \mathbf{W}_x \mathbf{J}_{xd} + \mathbf{J}_d^T \mathbf{W}_d \mathbf{J}_d \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x}_m \\ \Delta\mathbf{d}_m \end{bmatrix} = - \begin{bmatrix} \mathbf{J}_x^T \mathbf{W}_x \mathbf{r}_x \\ \mathbf{J}_{xd}^T \mathbf{W}_x \mathbf{r}_x + \mathbf{J}_d^T \mathbf{W}_d \mathbf{r}_d \end{bmatrix}, \quad (15)$$

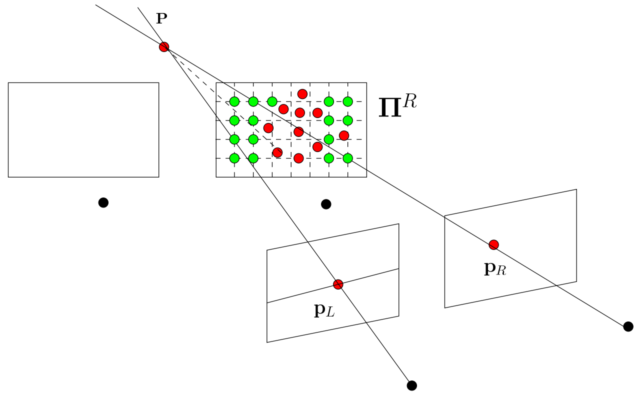

Fig. 3. $\mathbf{\Pi}^R$ is a combination of old and new points. New points are initialized by technique described in Section III whereas old points are efficiently re-located within tight search bounds.

where $\mathbf{J}_x = \frac{\partial \mathbf{r}_x}{\partial \mathbf{x}}$, $\mathbf{J}_{xd} = \frac{\partial \mathbf{r}_x}{\partial \mathbf{d}}$ and $\mathbf{J}_d = \frac{\partial \mathbf{r}_d}{\partial \mathbf{d}}$. Few notes about the derivation of $\mathbf{J}_x$ can be found in Appendix A.

Assuming $\widehat{\mathbf{d}}_m$ is already fixed by the exhaustive search, $\Delta\mathbf{d}_m$ can be marginalized out at each iteration using Gaussian elimination, and the equation for $\Delta\mathbf{x}_m$ becomes:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x}_m \\ \Delta\mathbf{d}_m \end{bmatrix} = \begin{bmatrix} \mathbf{E} \\ \mathbf{F} \end{bmatrix} \quad (16)$$

$$(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})\Delta\mathbf{x}_m = \mathbf{E} - \mathbf{B}\mathbf{D}^{-1}\mathbf{F}. \quad (17)$$

Motion increments are applied by $\mathbf{T}^R = \mathbf{T}^R e^{\mathbf{A}(\Delta\mathbf{x}_m)}$ until $|\Delta\mathbf{x}_m| < \epsilon$. The disparity covariance $\mathbf{C}_d$ is estimated by marginalizing out local uncertainty for the last $\Delta\mathbf{x}_m$ by

$$\mathbf{C}_d = (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}. \quad (18)$$

V. Disparity map integration

$\mathbf{\Pi}_{m+1}^R$ is always a combination of old points $\mathbf{p}_k^o$ and new points $\mathbf{p}_k^n$. Those $\mathbf{p}_k^o$, whose estimation weights are positive ($\mathrm{w}_k^d \mathrm{w}_k^x > 0$) with low variance ($\widehat{\sigma}_k < \frac{1}{16} width$), are mapped forward because they represent spatio-temporally reliable geometry. $\mathbf{p}_k^n$ are generated using the initial matching technique described in Section III to cover the regions where $\mathbf{p}_k^o$ do not warp. The combination is illustrated in Fig. 3.

$\widehat{\mathbf{\Pi}}_{m+1}^d$ is defined by an integrated disparity map, which is computed in the following way.

```
1) for each $\widehat{d}_k$ if $(\mathrm{w}_k^d \mathrm{w}_k^x > 0$ and $3\widehat{\sigma}_k < \frac{1}{16} width)$
     $b_k^s = \begin{bmatrix} 1 & 0 \end{bmatrix} \left( w^L(\widehat{\mathbf{x}}, \widehat{d}_k - 3\widehat{\sigma}_k) - w^R(\widehat{\mathbf{x}}, \widehat{d}_k) \right)$
     $b_k^e = \begin{bmatrix} 1 & 0 \end{bmatrix} \left( w^L(\widehat{\mathbf{x}}, \widehat{d}_k + 3\widehat{\sigma}_k) - w^R(\widehat{\mathbf{x}}, \widehat{d}_k) \right)$
     $d_k^+ = \underset{d^+}{argmin}\ C\left( w^L(d_k^+), w^R(\widehat{\mathbf{x}}, \widehat{d}_k) \right), d_k^+ \in [b_k^s, b_k^e]$
     $\mathcal{D}_{m+1}(\mathbf{p}) = d_k^+$, where $\mathbf{p} \in \mathbb{R}_{3\times3}(w^R(\widehat{\mathbf{x}}, \widehat{d}_k))$
2) Initialize un-assigned parameters in $\mathbf{d}_{m+1}$
3) Do subpixel refinement for $\mathbf{d}_{m+1}$
```

Bound transfer to a next frame is illustrated in Fig. 4. Deviations $\widehat{\sigma}_k$ are obtained as square-rooted diagonal values of covariance $\mathbf{C}_d$ and $\mathbb{R}_{3\times3}(\mathbf{p})$ is a set of nearest grid points in 3×3 region around point $\mathbf{p}$.

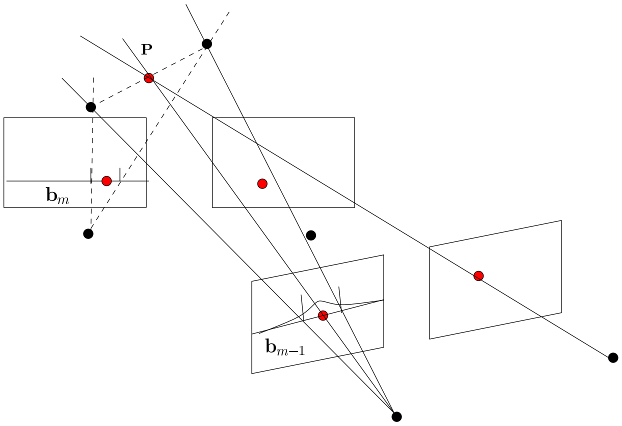

Fig. 4. The transfer of a bound into next stereo view. A bound is defined as $[-3\sigma_k, 3\sigma_k]$ interval of the 1D Gaussian uncertainty of associated d_k.

VI. Results

A. Dense SLAM Algorithm

1) Exhaustive search to initialize $\mathbf{d}_0$
2) Non-linear motion parameter estimation for finding $\widehat{\mathbf{x}}_m$
3) Determine $\mathbf{C}_d$ from cost function by marginalization
4) Find integrated $\widehat{\mathbf{d}}_{m+1}$ using bounded search
5) $m = m + 1$, jump to phase 2)

B. Extending convergence domain by multiple resolutions

As motion optimization starts at $\mathbf{x} = \mathbf{0}$, it is assumed that iteration starts near the global minimum. However, with faster motion or lower frame rate, the initial guess might be outside the convergence domain. It can however be extended spatially with a multi-resolution representation of the input images. At first $\mathbf{x}$ is estimated using low resolution images. Then the solution is refined by continuing the optimization at a higher resolution.

C. Hole filling

After the old points are re-localized within tight bounds, they do not naturally align with the existing grid points. This is why d_k^+ are extrapolated into discrete 3×3 neighborhood of the closest grid points. This effectively grows spatio-temporally reliable structure to cover also new points. After disparity map integration, all disparities are refined into sub-pixel accuracy and extrapolated geometry is locally optimized to match the image data.

D. Additional constraints

The image gradients are not valid for estimating the motion at discontinuous regions of the disparity map, such as at the boundaries of objects. Discontinuities can also originate from false stereo matches. It is considerable to neglect discontinuous surface regions that stretch dramatically or change winding order when mapping rectangle centered to each point $\mathbf{p}_k^R$ to left view. For these points weight w_k^d is set to zero.

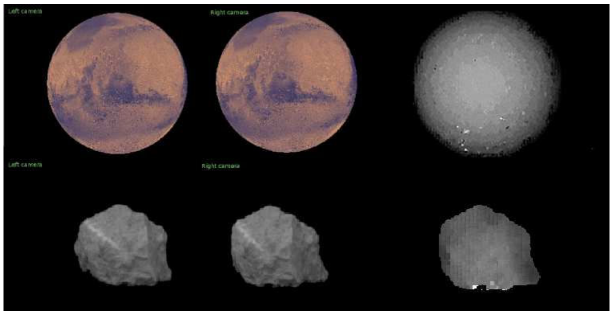

Fig. 5. Mars and Rock with the corresponding disparity images. Images are not in original scale.

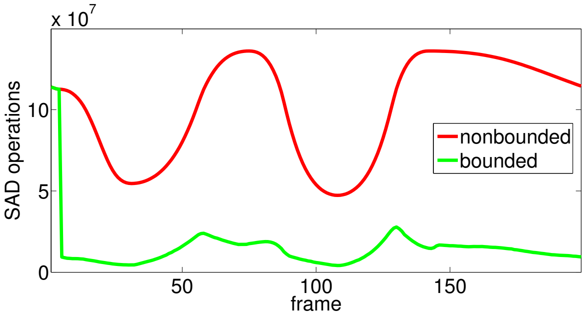

Fig. 6. Mars, The difference in amount of SAD operations with (green) and without disparity (red) bounding. Silhouette area varies during the sequence causing fluctuations to the required computational requirement.

E. Experiments

The proposed method is tested using a simulation and a real rock sequence (Fig. 5). An implementation of the proposed method was built with C++. Initially simple template matching was used for generating disparity maps, but as good template size was not found for real sequences, the exhaustive search had to be done using multi-resolution SAD [6], which also minimizes direct image error. The main validation is to compare with the classical visual odometry technique [14] that does not use bounding.

1) Simulation: The Mars sequence using a real texture [17] was rendered in 640×480 resolution with Blender along with ground truth trajectory. No lighting conditions were simulated and thus the model emitted texture colors directly. Three multi-resolution layers were used for motion estimation. A significant performance improvement is obtained by bounded search directly after the first frame (Fig. 6). Estimated trajectories for bounded and non-bounded case illustrated in Fig. 7. As can be seen, bounding does not degrade pose estimation quality. Disparity map accuracy is improved by bounding in the problematic cases where the camera is closer to homogeneously textured Mars (Fig. 8). Structural accuracy is measured by $\frac{1}{n}\Sigma(d_k - d_k^o)^2$, where d_k^o are the ground truth disparities. In practice they are obtained by projecting the intersection point between a ray and the object into the second view.

2) Real sequence: The rock sequence was recorded using a calibrated stereo camera and a turn-table (Fig. 9). The

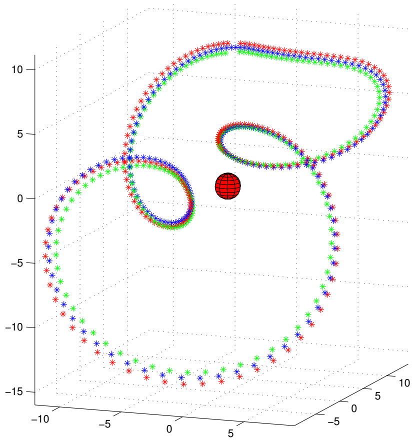

Fig. 7. Ground truth trajectory (blue) around Mars with estimated camera trajectories. Non-bounded in red ($\times 1.03$) and bounded in green ($\times 0.97$). As can be seen, the method can robustly track complicated 3D motion with low drift.

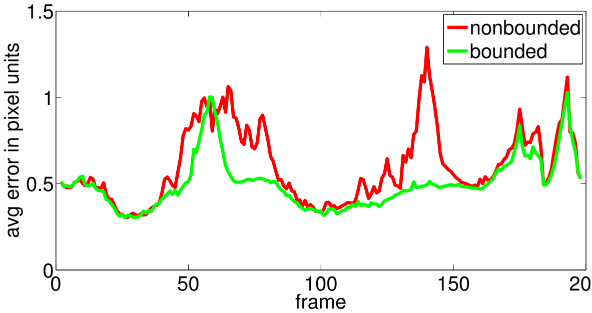

Fig. 8. Mars, The difference in disparity map accuracy with (green) and without (red) disparity bounding. In the problematic frames 60 and 140 camera is closer to Mars which, in general, has homogeneous texturing.

original resolution of 1280×960 was downsampled into 320×240 and two multi-resolution layers were used for motion estimation. Background subtraction and rectification was applied to the rock sequence as a post-process. Once again performance improvement is obtained by bounded search directly after the first frame (Fig. 10). The camera pose trajectory is marginally better for the bounded version as can be seen in Fig. 11.

VII. CONCLUSIONS

A dense structure model has been developed for image based stereo SLAM. The structure is parameterized as a set of disparities whose uncertainty defines bounds which are used for efficient and precise temporal re-localization of the structure. The simulations and real experiments show significant performance improvement in dense correspondency computation due to bounded search of disparities. Despite tighter bounding the accuracy of camera trajectory and disparities is still maintained and even improved. Fur-

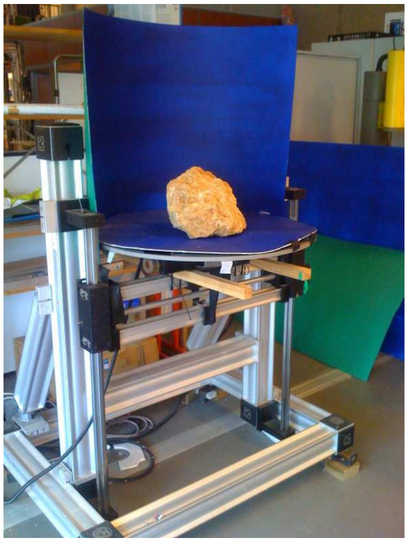

Fig. 9. The turn-table in INRIA was used for recording the rock sequence.

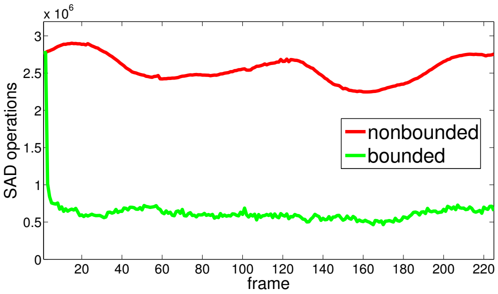

Fig. 10. A rock, The difference in amount of SAD operations with (green) and without disparity (red) bounding.

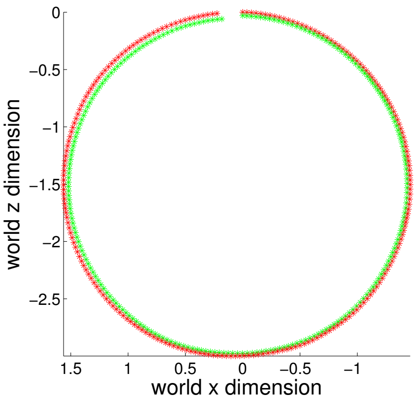

Fig. 11. A rock, The difference in bounded trajectory (green) and non-bounded trajectory (red). Bounded trajectory intentionally scaled by 0.98 for separation.

ther work will be done to handle unknown baseline case, varying lighting conditions and utilizing surface continuity constraints.

VIII. ACKNOWLEDGMENTS

The authors gratefully acknowledge CNRS and Thales-Alenia Space for funding the project and INRIA for providing a turn-table.

IX. APPENDIX

A. Motion Jacobian $\mathbf{J}_x$

$\mathbf{J}_x = \mathbf{J}_p * \mathbf{J}_g$ is a column-wise product of a photometric and a geometric component. $\mathbf{J}_g$ represents the optical flow of 3D structure and it can be precomputed once for each reference disparity map if the coordinate system for the motion increment is chosen to be the reference. Thus it is natural to define $\mathbf{J}_x$ for the warped image.

$$\mathbf{J}_g = \begin{bmatrix} \frac{\partial w^R(\mathbf{x},d_1)}{\partial x_1} & \cdots & \frac{\partial w^R(\mathbf{x},d_1)}{\partial x_6} \\ \vdots & \vdots & \vdots \\ \frac{\partial w^R(\mathbf{x},d_N)}{\partial x_1} & \cdots & \frac{\partial w^R(\mathbf{x},d_N)}{\partial x_6} \end{bmatrix} \tag{19}$$

$\mathbf{J}_p$ is the intensity gradient at the points of the warped image.

$$\mathbf{J}_p = \begin{bmatrix} \frac{\partial \mathcal{I}(\mathbf{p}_0)}{\partial \mathbf{p}} & \cdots & \frac{\partial \mathcal{I}(\mathbf{p}_N)}{\partial \mathbf{p}} \end{bmatrix}^T \tag{20}$$

In each iteration of the optimization, $\mathbf{J}_x$ must be evaluated only at $\mathbf{x} = \mathbf{0}$, because the previous motion increments can be stacked. This means $\widehat{\mathbf{x}}_m$ is not explicitly required, but the relative transformation is generated as a sequence of small transformation steps $\Delta\widehat{\mathbf{x}}_i$. In this form it is also possible to use the special rule $\frac{\partial e^{\mathbf{A}(\mathbf{x})}}{\partial \mathbf{x}}(\mathbf{0}) = \frac{\partial \mathbf{A}(\mathbf{x})}{\partial \mathbf{x}}$.

B. Relation to full bundle adjustment

In an intensity based bundle adjustment problem, the parameter vector $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{d}_1 \ldots \mathbf{d}_m]^T$ is optimized, where $[\mathbf{x}_1, \ldots, \mathbf{x}_m]^T$ are the motion parameters of full trajectory organized into a column vector of $6m \times 1$ elements and $[\mathbf{d}_1 \ldots \mathbf{d}_m]^T$ are the disparities for all views organized into a column vector of $mn \times 1$ elements. The bundle adjustment problem is by nature local for SLAM due to visibility and occlusion. Disparities as a local structure descriptor are suitable.

The objective function is defined as

$$\begin{aligned} \mathbf{e}_{ij}(k) &= \mathcal{I}_j\left(w(\mathbf{x}_{ij}, \mathbf{d}_i(k))\right) - \mathcal{I}_i\left(\mathbf{p}_i(k)\right), i \neq j, && (21) \\ \mathbf{r}_g &= \left[\mathbf{e}_{12}\ \mathbf{e}_{13} \cdots \mathbf{e}_{(m-1)m}\right]^T && (22) \\ \mathcal{O}(\mathbf{x}) &= \mathbf{r}_g^T \mathbf{W}_g \mathbf{r}_g, && (23) \end{aligned}$$

where $\mathbf{x}_{ij}$ is the motion parameterization between views i and j, $\mathbf{r}_g$ is the full residual and $\mathbf{W}_g$ it's weight matrix.

Assuming a good initial guess, *Gauss-Newton* iteration rule for finding the optimum is

$$\mathbf{H}_g \Delta\mathbf{x} = \mathbf{J}^T \mathbf{W}_g \mathbf{r}_g, \tag{24}$$

where $\mathbf{J} = \frac{\partial \mathbf{r}_g}{\partial \mathbf{x}}$ and $\mathbf{H}_g = \mathbf{J}^T \mathbf{W}_g \mathbf{J}$.

For incremental SLAM, the state is decomposed into old and new parts

$$\mathbf{x} = \left[\mathbf{x}_o \,\middle|\, \mathbf{x}_n\right]^T = \left[\mathbf{x}_1 \cdots \mathbf{x}_{m-1}\ \mathbf{d}_1 \cdots \mathbf{d}_{m-1} \,\middle|\, \mathbf{x}_m\ \mathbf{d}_m\right]^T. \tag{25}$$

$\mathbf{H}_g$ is decomposed into old and new blocks

$$\mathbf{H}_g = \begin{bmatrix} \mathbf{H}_{oo} & \mathbf{H}_{on} \\ \mathbf{H}_{no} & \mathbf{H}_{nn} \end{bmatrix}. \tag{26}$$

And the iteration rule of Equation 24 can be re-written as

$$\begin{aligned} \mathbf{H}_n \Delta\mathbf{x}_n &= \mathbf{g} && (27) \\ \mathbf{H}_n &= \mathbf{H}_{nn} - \mathbf{H}_{on}^T \mathbf{H}_{oo}^{-1} \mathbf{H}_{on} && (28) \\ \mathbf{g} &= \mathbf{J}_n^T \mathbf{W}_n \mathbf{r}_n - \mathbf{H}_{on}^T \mathbf{H}_{oo}^{-1} \mathbf{J}_o^T \mathbf{W}_o \mathbf{r}_o. && (29) \end{aligned}$$

In this paper the incremental estimation related to the global one is obtained by assuming elements of $\mathbf{H}_{oo}^{-1} \approx \mathbf{0}$ and previous state residual $\mathbf{r}_o \approx \overline{\mathbf{0}}$. This means past trajectory is assumed to be correct without variance and the optimization problem boils down to a one concerning $\begin{bmatrix}\mathbf{x}_m & \mathbf{d}_m\end{bmatrix}^T$.

REFERENCES

[1] P.W. Holland and R.E. Welsch, Robust Regression Using Iteratively Reweighted Least-Squares, *Comm. Statistics Theory and Methods*, vol. A6, pp. 813–827, 1977.

[2] R.M. Haralick, Propagating Covariance in Computer Vision, *In Proc. Workshop on Performance Characteristics of Vision Algorithms*,1996,pp 493-498.

[3] B. Triggs, P. Mclauchlan, R. Hartley and A. Fitzgibbon, Bundle Adjustment – A Modern Synthesis, 2000

[4] D. Scharstein and R. Szeliski, A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms, Vol 47, 2002, pp. 7–42.

[5] R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, ISBN 0-521-54051-8, 2003

[6] R. Yang and M. Pollefeys, Multi-Resolution Real-Time Stereo on Commodity Graphics Hardware, *IEEE Computer Societ Conference on Computer Vision and Pattern Recognition*, vol. 1, 2003, pp. 211–217

[7] A. I. Comport, Statistically robust 2D visual servoing, In *IEEE Transactions on Robotics*, Vol 22(2), 2006, pp. 415-421

[8] Y. Ma, S. Soatto, J. Kosecka and S. Sastry, An Invitation to 3-D Vision, Springer, ISBN 978-0-387-00893-6, 2006

[9] A.J. Davison, I.D. Reid, N.D. Molton and O. Stasse, MonoSLAM: Real-Time Single Camera SLAM, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, 2007, pp 1052–1067.

[10] G. Klein and D. Murray, Parallel Tracking and Mapping for Small AR Workspaces, In *Proceedings of the International Symposium on In Mixed and Augmented Reality* (ISMAR), 2007, pp. 225–234.

[11] H. Bay, A. Ess, T. Tuytelaars and L. V. Gool, SURF: Speeded Up Robust Features, In *Computer Vision and Image Understanding* (CVIU), Vol. 110, No. 3, pp. 346–359, 2008.

[12] K. Konolige and M. Agrawal, FrameSLAM: from Bundle Adjustment to Realtime Visual Mappping, In *IEEE Transactions on Robotics*, Vol 24, 2008, pp. 1066–1077.

[13] G. Silveira, E. Malis and P. Rives, An Efficient Direct Approach to Visual SLAM, In *IEEE transactions on robotics*, Vol 24, 2008, pp. 969–979.

[14] A.I. Comport, E. Malis and P. Rives, Real-time Quadrifocal Visual Odometry, *The International Journal of Robotics Research*, vol. 29, 2010, pp 245-266.

[15] M. Meilland, A.I. Comport and P. Rives, A Spherical Robot-Centered Representation for Urban Navigation, *IROS*, 2010.

[16] R.A. Newcombe and A.J. Davison, Live Dense Reconstruction with a Single Moving Camera, *CVPR*, 2010.

[17] High resolution color texture map of Mars produced by NASA/USGS, `http://www.solarviews.com/cap/mars/marscyl1.htm`