Multi-view dense 3D modelling of untextured objects from a moving projector-cameras system

Jacques Harvent, Benjamin Coudrin, Ludovic Brèthes, Jean-José Orteu, Michel Devy

Machine Vision and Applications manuscript No.
(will be inserted by the editor)

Multi-View Dense 3D Modelling of Untextured Objects From a Moving Projector-Cameras System

Jacques Harvent · Benjamin Coudrin · Ludovic Brèthes · Jean-José Orteu · Michel Devy

Received: date / Accepted: date

Abstract Structured light methods achieve 3D modelling by observing with a camera system a known pattern projected on the scene. The main drawback of single projection structured light methods is that moving the projector changes significatively the appearance of the scene at every acquisition time. Classical multi-view stereovision approaches based on appearance matching are then not useable. The presented work is based on a two-cameras and one single slide projector system embedded in a hand-held device for industrial applications (reverse engineering, dimensional control, ...). We propose a method to achieve multi-view modelling for camera pose and surface reconstruction estimation in a joint process. The proposed method is based on the extension of a stereo-correlation criterion. Acquisitions are linked through a generalized expression of local homographies. The constraints brought by this formulation allow an accurate estimation of the modelling parameters for dense reconstruction of the scene and improve the result when dealing with detailed or sharp objects, compared to pairwise stereovision methods.

Keywords Multi-View Stereovision · 3D Modelling · Bundle Adjustment · Image Correlation · Pattern projection · Image-based modelling

J. Harvent · B. Coudrin · L. Brèthes
Noomeo, R&D Department, 425 rue Jean Rostand, 31670 Labège France
Tel.: +33 5 61 00 77 15
Fax: +33 9 74 76 02 52
E-mail: jacques.harvent@noomeo.eu

B. Coudrin
E-mail: benjamin.coudrin@noomeo.eu

L. Brèthes
E-mail: ludovic.brethes@noomeo.eu

J.J. Orteu
Université de Toulouse ; Mines Albi ; ICA (Institut Clément Ader) ; Campus Jarlard, F-81013 Albi, France
E-mail: jean-jose.orteu@mines-albi.fr

M. Devy
Université de Toulouse ; LAAS-CNRS ; 7 avenue du Colonel Roche, F-31077 Toulouse, France
E-mail: michel@laas.fr

1 Introduction

Modelling a scene in 3D from images consists in mapping image pixels with real world 3D points. Since the imaging process is a projection from a 3D space to a 2D space, some information is lost and one would need more than a single image to fully recover a real world point. The process is then the problem of retrieving a scene structure from a set of images. This is done by jointly estimating camera poses at acquisition times and inferring the structure from different point of views of the surface.

With the development of stereo vision methods and 3D modelling from vision, scanning devices and methods have greatly increased. Industrial applications actors are increasingly using 3D modelling for a wide range of problems. Among the main applications, one can note reverse engineering, dimensional control, simulation, design, industrial maintenance, etc.

In these applications, some objects can be challenging for vision technologies : untextured, with details or sharp edges, specular surfaces, and so on. Basic stereo vision setups – that is to say with simply a pair of cameras for instance – are often unable to accurately and densely model such objects. Untextured objects penalize matching process in stereo images, edges cause occlusions and ambiguities. To overcome these limitations, active vision systems propose to provide a measurable energy to the scene, often in visible or near

infrared regions of the light spectrum or using laser. Among active vision systems, structured light methods project a specific pattern on the scene. This pattern can be known and its projector calibrated to act like an inverted camera in a stereo system, or can be unknown and used to add texture in a peak detection [4] or in an appearance correlation approach [40]. The projection can be a set of patterns with temporal coding. This imposes that the system is not moving during the projection. Structured light methods based on a single projection are, on the contrary, useable in moving devices, such as hand-held 3D modelling systems.

Recently, important research efforts have been directed towards the design of small and portable 3D modelling systems [39,20]. Hand-held systems allow an eased scanning process due to a more dexter handling compared to fixed systems. Most hand-held 3D modelling devices are based upon laser triangulation techniques or time-of-flight measurements. Hand-held structured light devices remain rare due to the projector limitations [12]. If the projector is placed in the room, independently from the scanning device, the operator has to take care of the occlusions of the lights causing shadows in the scene. Moreover, the setup of such a system can be complicated (more than one projector may be needed, the projectors have to be placed judiciously to cover the entire object, ...) which is contrary to the initial purpose of simplicity of hand-held devices. If the projector is embedded in the device, the scene changes every time the device moves. Indeed, since the operator has to move the scanner to cover the entire scene, every acquisition leads to a new device position, and then to a new projector position, changing the projected light on the scene, that is to say changing the scene texture (figure 1).

A classical approach for multi-view 3D modelling from cameras is to match appearance information in a set of images and optimize a system consisting of camera poses and scene surface parameters. Using a hand-held structured light device with an embedded projector does not allow such approach due to the appearance changing when the device is moving. Frequently the modelling process operates as a range scanning task: 3D information – 3D point cloud – is retrieved from images at every acquisition, and registration is achieved using geometrical methods [33,37,7].

However, interest of using a multi-view approach has been demonstrated by a number of works [14,36,6]. This kind of approach utilizes more information in the modelling process. It tends to give more robust and accurate results when compared to standard pairwise approaches. Multi-view stereo (MVS) provides an efficient way to retrieve a complex, dense and accurate shape

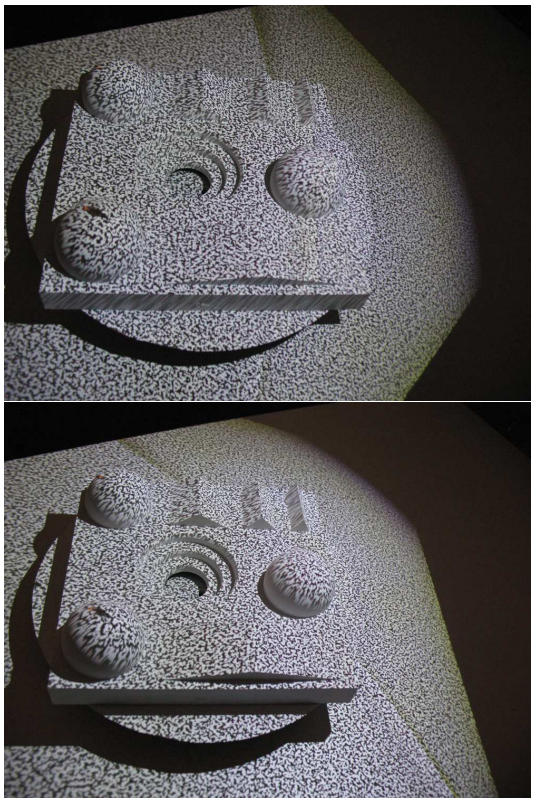


Fig. 1 The structured light device projects a specific pattern. The appearance of the scene changes when the device moves

from a set of images. A taxonomy and evaluation of MVS is presented by Scharstein and Szeliski [34]. Some of the best-performing methods are even available in a software package and allows to reconstruct very larges scenes [13,15]. Many MVS methods start from an initial surface, refined afterward during an optimization step. The initial shape can be acquired from *Structure-From-Motion* techniques or from traditional stereo methods [41]. In an industrial context, mechanical parts control applications for instance, a CAD model can be used as a reference in this process. Normals to the surface are often the optimized parameters as they encode the 3D position of points and the direction of the surface. MVS is known to perform particularly well for sufficiently textured scenes. Untextured areas are processed considering priors like surface smoothness. These priors may result in inaccurate results on the final 3D model. Most stereovision algorithms use local planar approximation models – correlation windows, for instance – that may oversmooth the surface. MVS approaches take advantage of the number of views to refine 3D surfaces using simple planar models. Other approaches use more complex surface models like surfels [5]. The complexity of these models make them hard to use in multi-view formulations.

Among important contributions, Pollefeys *et al.* [31] presented a complete *Structure-from-Motion* approach for off-line modelling of a scene acquired from a single uncalibrated mobile camera. This method tracks features in the image sequence to globally determine camera intrinsic and extrinsic parameters along with the geometry of tracked keypoints. In a second time, the 3D scene is densified by propagating pairwise-computed disparities to expected neighbours in other images. A recent evolution [23] of this kind of approach combined a probabilistic framework for real-time *Structure-from-Motion* and a densification using local *Bundle Adjustment* [42] to lower dynamic. The main drawback of these methods is the need of highly textured scenes.

Recently, Wu *et al.* [44] proposed to combine structure from shading and multi-view stereo. However, applying these methods to industrial applications is not straightforward as a lot of constraints must be considered. Objects in industrial applications are most of the time untextured as they are made of a single material. When silhouettes are available, they are sometimes embedded in the process [17] in order to add constraints on the reconstructed model.

Recent work on *Simultaneous Localization and Mapping* (SLAM) has been oriented towards the use of visual data. Localization using photometric techniques has been explored to introduce optical odometry [27, 28] but suffers from drifts and rapidely causing large uncertainty. Davison [10] proposed an approach to sparse localization and mapping using a single mobile camera based on Kalman filtering. It benefits from information federation and loop closure to refine the last camera pose and the sparse model. This work has been extended to propose real-time implementations using active search in images [11]. Montiel *et al.* [21] solved the features delayed-initialization problem of the method using inverse-depth parameterization of features. Paz *et al.* [30] proposed the use of stereo camera systems to fully observe features but limiting the observation field of the features. Sola *and al* [38] considers the stereo system as a multi-cameras system, adding in the map points fully observed in the stereo view field, and partially observed outside.

Again, these approaches require that the scene does not change – no mobile object or lighting condition changes are allowed – and are not adapted to a moving projector causing a change in the scene appearance.

Geometrical approaches have been proposed to solve the pose estimation of range laser sensors [26], in robotics applications for instance. Nüchter *et al.* [25] proposed a framework to solve the simultaneous localization and modelling problem for robot navigation tasks in difficult areas exploration missions. Poses uncertainties can be taken into account by representing spatial transformations as random variables. Lu and Milios [19] introduced a stochastic approach for robot navigation applications. Other authors improved it with stochastic filtering [24], using Extended Kalman Filter for instance. These methods have been designed for sensor pose refinement but do not refine the scene model. Moreover they impose a sequential acquisition at relatively high dynamics to be consistent. Other geometrical approaches propose initial mesh refinments to match an input data set. Curless and Levoy [9] use a ponderation on a signed distance function from input meshes to fill a global voxel grid approximating the actual 3D surface. Zharescu *et. al.* applied their topological approach [45] to 3D reconstruction. However, these methods can not jointly estimate camera poses and surface reconstructions.

Vision systems modelling highly specular surfaces can be considered as appearance-change problems since the specular reflections is a function of the camera pose and do not project the same way in all images. Some approaches [29] remove saturated image regions. The holes created are filled using a multiple view imaging system. Reflectance methods [35] propose to model the specular reflections when measuring the surface geometry. Zheng *et al.* [47] use a combination of several weighted functions to represent light reflections to retrieve reflectance properties of an object and retrieve its normal map. Pons *et al.* [32] proposed an original formulation of a scene flow problem to model non-Lambertian surfaces or illumination changes. However, the method requires a sequential acquisition and assumes that the scene is not completely modified by illumination perturbations. Vu *et. al.* [43] extended the method to the refinment of a coarse initial mesh using a similar formulation and considering the visibility of points. However, this method is appearance-based and considers the extrinsic camera parameters fixed.

Our approach is a *Bundle Adjustment* based global optimization with a novel observation scheme. It refines a coarse model previously reconstructed following a standard pairwise stereo approach. The initial approach adjusts the local homography around a correlated point minimizing the intensity differences in stereo images [3]. The optimization criterion is based on reducing the correlation error of points projected in pairs of stereo images. Correlation scores are observed through a luminance difference in images acquired at the same time and globally linked to every view observing the same points with a generalized formulation of homographies induced from tangent planes to the surface. This approach proposes a method for multi-view optimization of sensor poses and surface reconstruction

using a least-squares formulation. It is similar to the approach of Chang *et. al.* [6] but with a specific formulation for MVS problem. We show that our method brings a multi-view approach to the particular case of a texture changing over time. We also demonstrate that it is a consistent multiview correlation-based approach for generic purpose problems allowing an accurate 3D modelling of objects even with sharp edges or details. It is also robust to the correlation window size changes, even with very small sizes.

We first introduce our conventions and model in section 2. The experimental device used in our application and for the evaluations is also briefly introduced. Our method is presented in section 3. General considerations and our formulations are introduced with the simple stereo case (3.1). Then our method is generalized to the multiple stereo-images case (3.2) and some algorithm optimization is introduced (3.3). The method is evaluated in section 4 after presenting our experimental protocol. Section 5 gives some conclusions and perspectives for improvements.

2 System Model and Principle

The method we propose is designed to address the particular case of multi-view 3D modelling of a texture-changing scene. For the purpose of the method we used an experimental camera-projector setup. In the following chapter, we present this system and its operation along with some conventions.

2.1 System Model

The imaging process of a camera allows the mapping of a real world 3D point $\mathbf{m}_{\mathcal{W}} = (X, Y, Z, 1)^T$ to a 2D image pixel $\mathbf{m}_{\mathcal{I}} = (sx, sy, s)^T$. It is a relation of the form

$$\mathbf{m}_{\mathcal{I}} = \mathbf{K}\,[\mathbf{R}_{\mathcal{CW}}|\mathbf{t}_{\mathcal{CW}}]\,\mathbf{m}_{\mathcal{W}} \tag{1}$$

In equation (1), $\mathbf{K}$ is the intrinsic parameters matrix. It allows the mapping of a point expressed in the camera frame to an image pixel. The rigid transformation $[\mathbf{R}_{\mathcal{CW}}|\mathbf{t}_{\mathcal{CW}}] \in SE(3)$ allows the change between world and camera frames. In general, rigid transformations between points $\mathbf{p}_{\mathcal{A}}$ expressed in frame $\mathcal{A}$ and points $\mathbf{p}_{\mathcal{B}}$ expressed in frame $\mathcal{B}$ are noted $[\mathbf{R}_{\mathcal{AB}}|\mathbf{t}_{\mathcal{AB}}]$ such that

$$\mathbf{p}_{\mathcal{A}} = [\mathbf{R}_{\mathcal{AB}}|\mathbf{t}_{\mathcal{AB}}]\,\mathbf{p}_{\mathcal{B}} \tag{2}$$

In our system, we use two cameras rigidly attached. First camera frame is noted $\mathcal{C}_0$ and second camera frame is noted $\mathcal{C}_1$. Each camera has an independent intrinsic parameters matrix, respectively $\mathbf{K}_0$ and $\mathbf{K}_1$. The transformation between cameras is noted $[\mathbf{R}_{\mathcal{C}_1\mathcal{C}_0}|\mathbf{t}_{\mathcal{C}_1\mathcal{C}_0}]$.

The camera parameters are calibrated using a chessboard calibration target and following the method described by Zhang [46]. The chessboard has been modified to allow automatic initialization. Images are corrected to remove the effect of distorsions prior to any processing, allowing the use of the linear models described previously.

2.2 Coarse solution

The system is handheld. It is composed of two CCD 1024×768 cameras with $8mm$ lenses, a projector and an intertial sensor. The stereoscopic baseline is $140mm$ long and the cameras are oriented with a 15° angle. The whole setup has been described and evaluated by Coudrin *et al.* [8]. The cameras and the projector are synchronised. We use a pattern projection to infer the dense 3D information from the pair of cameras. When the system is triggered, the pattern is projected on the scene. Then the two cameras acquire images simultaneously. The pair of images is used in a reconstruction algorithm, providing a 3D point cloud by a stereo-vision surface growing method. Points are expressed with respect to the frame $\mathcal{C}_0$.

To model an object completely, one has to move the vision system around the object. An acquisition at time t provides a set of 3D points expressed in the current first camera frame, $\mathcal{C}_0^t$. Point sets have to be registered in a common frame to retrieve the actual geometry of the scene. Registration can be solved using inertial sensing or surface descriptors matching for initial alignment and *ICP* algorithm or variants for refinement [33, 20]. Combining these algorithms makes the system more robust to symmetric ambiguity.

2.3 Towards Multi-View operation

This approach can lead to good results but still suffers from some problems. Mostly the accuracy of the results can be too low for actual industrial applications. The 3D surface reconstruction is based on the observation of the image neighbourhood of every pixel and then produces an over-smoothing effect related to the neighbourhood size, or noise near model discontinuities. In the case of reverse engineering, for instance, the smoothing effect can be problematic since it alters the sharpness of edges or suppresses details on the surface.

Moreover, the registration process is independent. A multiview approach tries to estimate jointly the camera poses and the structure scene. In this way, both informations can benefit from a global optimization process. In the approach introduced previously, the registration, being independent and using the surface discretization from the reconstruction, directly suffers from reconstruction errors or inaccuracies. We propose to improve the previous reconstruction by refining 3D points through multi-view optimization.

3 Multi-View Optimization

Vision-based systems can capitalize a lot from using all images at a time to refine the 3D model. Since, in our problem, the pattern projector is moving along with the cameras, causing a change of the appearance of the scene at every acquisition, classical *Bundle Adjustment* methods cannot be used. Our method uses the estimation of the homography induced by the tangent plane to a surface point [16] in a pair of images taken at the same time. The expression of the homography is then generalized to every image pair observing the same point to produce a non-linear least-squares criterion to optimize.

3.1 Pairwise Correlation

Let us assume that the system is at the origin of the world frame $\mathcal{W}$. That is to say, in our notation, that the transformation between the world frame and the first camera is the identity transformation $[\mathbf{R}_{\mathcal{C}_0\mathcal{W}}|\mathbf{t}_{\mathcal{C}_0\mathcal{W}}] = [\mathbf{I}_{3\times3}|\mathbf{0}_{3\times1}]$. With this assumption, the relation between a 3D point expressed in world frame and an image pixel in image taken from camera $\mathcal{C}_0$ can be written from equation (1) as

$$\mathbf{m}_{\mathcal{W}} = \mathbf{K}_0^{-1}\mathbf{m}_{\mathcal{I}_0} Z \tag{3}$$

with Z being the third coordinate of the 3D point.

From equations (1) and (3), we can write a relation between the projections in stereo images from cameras $\mathcal{C}_0$ and $\mathcal{C}_1$

$$\mathbf{m}_{\mathcal{I}_1} = \mathbf{K}_1 \left[\mathbf{R}_{\mathcal{C}_1\mathcal{C}_0}|\mathbf{t}_{\mathcal{C}_1\mathcal{C}_0}\right] \begin{bmatrix} \mathbf{K}_0^{-1}\mathbf{m}_{\mathcal{I}_0} Z \\ 1 \end{bmatrix} \tag{4}$$

The structure of the observed scene is obtained by finding Z in equation (4). This is achieved by determining the point $\mathbf{m}_{\mathcal{I}_1}$ matching the point $\mathbf{m}_{\mathcal{I}_0}$. This is done by observing and matching luminance informations in both images. To avoid ambiguities, the matching point is found by comparison of local appearance around the pixel $\mathbf{m}_{\mathcal{I}_0}$. Surface can be approximated locally by tangent planes to observed points. Let W be a region around point $\mathbf{m}_{\mathcal{I}_0} = (x, y, 1)^T$. Matching point to every $\mathbf{p}^i_{\mathcal{I}_0} = \left(x + \delta^i_x, y + \delta^i_y, 1\right)^T$, $\forall i \in [0..card\,(W) - 1]$ in image $\mathcal{I}_1$ can be found using the homography $\mathbf{p}^i_{\mathcal{I}_1} = \mathbf{H}\left(\mathbf{p}^i_{\mathcal{I}_0}\right)$ induced by the tangent plane to the observed 3D point (figure 2).

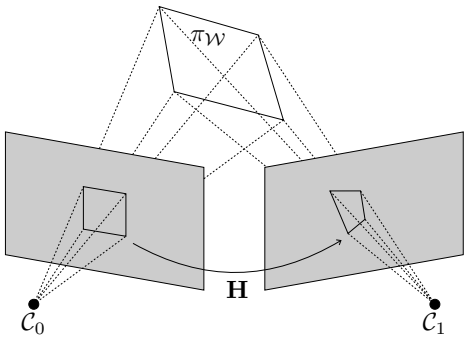


Fig. 2 An homography is a projective transformation relating the projections of a plane in two images

The tangent plane is defined using three points from region W. Let us define W as a square region in image $\mathcal{I}_0$ and choose, for convenience, its center $\mathbf{p}^0_{\mathcal{I}_0} = \mathbf{m}_{\mathcal{I}_0}$, and two corners $\mathbf{p}^1_{\mathcal{I}_0}$ and $\mathbf{p}^2_{\mathcal{I}_0}$ to define the plane. The 3D point associated to the pixel $\mathbf{p}^0_{\mathcal{I}_0}$ is defined by the parameter Z_0, as seen in equation (3). Let us define parameters Z_1 and Z_2 associated, respectively, to points $\mathbf{p}^1_{\mathcal{I}_0}$ and $\mathbf{p}^2_{\mathcal{I}_0}$ so that

$$\mathbf{p}^i_{\mathcal{W}} = \mathbf{K}_0^{-1}\mathbf{p}^i_{\mathcal{I}_0} Z_i, \quad \mathbf{p}^i_{\mathcal{W}} \in \mathbb{R}^3 \tag{5}$$

The normal vector $\mathbf{n}$ to the tangent plane is defined using our three points :

$$\mathbf{n} = \left(\mathbf{p}^1_{\mathcal{W}} - \mathbf{p}^0_{\mathcal{W}}\right) \times \left(\mathbf{p}^2_{\mathcal{W}} - \mathbf{p}^0_{\mathcal{W}}\right) \tag{6}$$

The tangent plane $\pi_{\mathcal{W}}$ is defined, using point $\mathbf{p}^0_{\mathcal{W}}$ and the normal vector $\mathbf{n}$, as all points $\mathbf{x} \in \mathbb{R}^3$ satisfying

$$\mathbf{n} \cdot \left(\mathbf{x} - \mathbf{p}^0_{\mathcal{W}}\right) = 0 \tag{7}$$

The coordinates of the tangent plane are then formalized as $\pi_{\mathcal{W}} = \left(\mathbf{n}^T, -\mathbf{n} \cdot \mathbf{p}^0_{\mathcal{W}}\right)^T$. The intersection between this plane and any ray coming from the camera center of camera $\mathcal{C}_0$ through a pixel of image $\mathcal{I}_0$ $\mathbf{p}^i_{\mathcal{I}_0}$, $\forall i \in [0..card\,(W) - 1]$ is written as

$$\begin{bmatrix} \mathbf{p}^i_{\mathcal{W}} \\ k \end{bmatrix} = \begin{bmatrix} -d & 0 & 0 \\ 0 & -d & 0 \\ 0 & 0 & -d \\ n_x & n_y & n_z \end{bmatrix} \mathbf{K}_0^{-1} \begin{bmatrix} \mathbf{p}^i_{\mathcal{C}_0} \\ 1 \end{bmatrix}, k \in \mathbb{R} \quad (8)$$

In equation (8), the normal vector is written as $\mathbf{n} = (n_x, n_y, n_z)^T$ and d is the fourth coordinate of the plane $\pi_{\mathcal{W}}$. Then, extending equation (4), matching point in image $\mathcal{I}_1$ to every pixel in the region W from image $\mathcal{I}_0$ can be expressed from

$$\begin{bmatrix} \mathbf{p}_{\mathcal{I}_1} \\ k \end{bmatrix} = \mathbf{H}\left(\mathbf{p}_{\mathcal{I}_0}\right) \quad (9)$$

$$= \mathbf{K}_1 \left[\mathbf{R}_{\mathcal{C}_1\mathcal{C}_0} | \mathbf{t}_{\mathcal{C}_1\mathcal{C}_0}\right] \begin{bmatrix} -d & 0 & 0 \\ 0 & -d & 0 \\ 0 & 0 & -d \\ n_x & n_y & n_z \end{bmatrix} \mathbf{K}_0^{-1} \begin{bmatrix} \mathbf{p}_{\mathcal{I}_0} \\ 1 \end{bmatrix} \quad (10)$$

This notation is equivalent to the one introduced by Hartley and Zisserman [16]. We introduce it because it allows a simpler adaptation of the homography expression to any configuration of cameras by simply composing rigid transformations, as it is shown in section 3.2.

In the following, coordinates of the plane are supposed to be normalized by the fourth coordinate $\pi_{\mathcal{W}} = \left(\mathbf{n}^T/d, 1\right)^T$ and we introduce the notation

$$\mathbf{S} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \\ \frac{n_x}{d} & \frac{n_y}{d} & \frac{n_z}{d} \end{bmatrix} \quad (11)$$

The generic expression of an homography in the two stereo cameras system is then expressed as

$$\mathbf{H} = \mathbf{K}_1 \left[\mathbf{R}_{\mathcal{C}_1\mathcal{C}_0} | \mathbf{t}_{\mathcal{C}_1\mathcal{C}_0}\right] \mathbf{S}\mathbf{K}_0^{-1} \quad (12)$$

The structure of the scene, from two images acquired at the same time, is obtained by minimizing dissimilarities between the luminance function $I_0\left(\mathbf{p}\right)$ in region W of image $\mathcal{I}_0$ and the luminance function $I_1\left(\mathbf{p}\right)$ in the transformed region $\mathbf{H}\left(W\right)$ of image $\mathcal{I}_1$. If we note $\mathbf{x} = (n_x/d, n_y/d, n_z/d)^T$, we try to find

$$\underset{\mathbf{x}}{argmin} \sum_{i \in W} \left(I_0\left[\mathbf{p}^i_{\mathcal{I}_0}\right] - I_1\left[\mathbf{H}_i\left(\mathbf{p}^i_{\mathcal{I}_0}, \mathbf{x}\right)\right]\right)^2 \quad (13)$$

Equation (13) introduces the minimization criterion used to find the parameters of a tangent plane from the correlation of two pixel regions in two stereo images. This problem can be solved using standard non-linear least-squares methods.

One should note that the assumption that the surface can be approximated by planes is valid only if the region W is small enough. If the region is large, the surface will be overly smoothed from this approximation, but if the region is too small ambiguities are introduced in appearance correlation, causing noise. The choice of the region size is then a trade off between details and noise levels. Extending this criterion to multiple image pairs, as introduced in section 3.2, not only allows multi-view stereo global optimization when appearance is changing between acquisitions but also allows to robustly estimate the scene geometry with a reduced region size.

3.2 Multi-Pairwise Correlation

We introduced in equation (12) the expression of the homography relating two projections of a 3D point in two images acquired at the same time from a calibrated stereo system. Let us suppose we take two acquisitions with our moving stereo system at time k_0 and k_1. The system is at the identity pose at time k_0, then moves to be at unknown pose $[\mathbf{R}_{k_1k_0}|\mathbf{t}_{k_1k_0}]$ at time k_1.

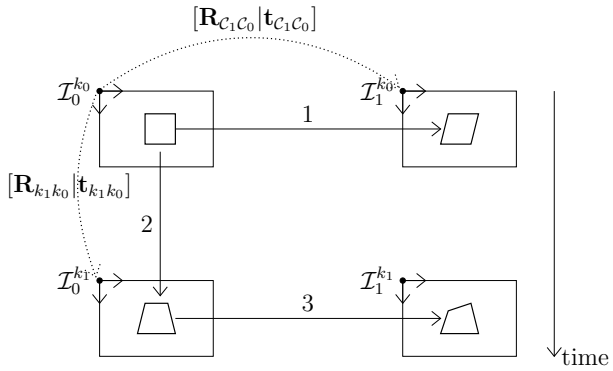


Fig. 3 Two acquisitions of a two cameras system

Figure 3 illustrates this setup and unfolds some relations that we need to introduce. Since frame $\mathcal{C}_0^{k_0}$ is the same as world frame, relation (1) has been introduced in (12). Relation (2) is similar to relation (1) but the rigid transformation between the two cameras is the unknown transformation $[\mathbf{R}_{k_1k_0}|\mathbf{t}_{k_1k_0}]$. Moreover the camera is the same, but at different acquisition times, so the intrinsic parameters matrix remains the same. The homography for a point i in this case is

$$\mathbf{H}^{i}_{\mathcal{I}_0^{k_1}\mathcal{I}_0^{k_0}} = \mathbf{K}_0 \left[\mathbf{R}_{k_1k_0}|\mathbf{t}_{k_1k_0}\right] \mathbf{S}_i \mathbf{K}_0^{-1} \tag{14}$$

Relation (3) is, also, similar to relation (1) but this time both cameras have moved. The relation between them is still the known calibrated rigid transformation $[\mathbf{R}_{\mathcal{C}_1\mathcal{C}_0}|\mathbf{t}_{\mathcal{C}_1\mathcal{C}_0}]$. In this case, the homography for a point i is written from equations (12) and (14) as

$$\mathbf{H}^{i}_{\mathcal{I}_1^{k_1}\mathcal{I}_0^{k_1}} = \mathbf{K}_1 \left[\mathbf{R}_{\mathcal{C}_1\mathcal{C}_0}|\mathbf{t}_{\mathcal{C}_1\mathcal{C}_0}\right] \left[\mathbf{R}_{k_1k_0}|\mathbf{t}_{k_1k_0}\right] \mathbf{S}_i \mathbf{K}_0^{-1} \tag{15}$$

Note that, using our formulation of the homography, relations are easily expressed by simply composing suitable rigid transformations. Therefore, from equation (15) we can introduce our general expression for the homography in a pair of stereo images. Let us suppose that a point i has been observed from a squared region in image $\mathcal{I}_0$, our two camera stereo system at this acquisition time is referenced by frame $\mathcal{L}$. After a motion, the system is referenced by frame $\mathcal{M}$. The homographies for the observation of the region in images $\mathcal{I}_0^{\mathcal{M}}$ and $\mathcal{I}_1^{\mathcal{M}}$ taken from the system in frame $\mathcal{M}$ are expressed as

$$\mathbf{H}^{i}_{\mathcal{I}_0^{\mathcal{M}}} = \mathbf{K}_0 \mathbf{P}_{\mathcal{M}\mathcal{L}} \mathbf{S}_i \mathbf{K}_0^{-1} \tag{16}$$

$$\mathbf{H}^{i}_{\mathcal{I}_1^{\mathcal{M}}} = \mathbf{K}_1 \left[\mathbf{R}_{\mathcal{C}_1\mathcal{C}_0}|\mathbf{t}_{\mathcal{C}_1\mathcal{C}_0}\right] \mathbf{P}_{\mathcal{M}\mathcal{L}} \mathbf{S}_i \mathbf{K}_0^{-1} \tag{17}$$

$$\mathbf{P}_{\mathcal{M}\mathcal{L}} = \left[\mathbf{R}_{\mathcal{M}\mathcal{W}}|\mathbf{t}_{\mathcal{M}\mathcal{W}}\right] \left[\mathbf{R}_{\mathcal{W}\mathcal{L}}|\mathbf{t}_{\mathcal{W}\mathcal{L}}\right] \tag{18}$$

with $\mathbf{P}_{\mathcal{M}\mathcal{L}}$ being the rigid transformation between frames $\mathcal{M}$ and $\mathcal{L}$, composed by the poses of the system at acquisition times relatives to the world frame.

From equations (16) and (17) we can formulate a global optimization criterion to refine points and camera poses from the model. Supposing the model has been initially reconstructed and registered (2.2), refining the model consists in building a relation with every point and every system pose. This involves to know how the pair of images taken from each system pose observes points. We have, then, to define a *visibility function* on a 2-tuple composed from a point and a system pose. Let us define the set of system poses V, $N_V = card\,(V)$, and the set of points P, $N_P = card\,(P)$. The visibility function $\phi\,(i,n)$ is defined for a point $i \in [1..N_P]$ and a pose $n \in [1..N_V]$ as

$$\phi\,(i,n) = \begin{cases} 0 \text{ if } i \text{ is not observable in } n \\ 1 \text{ if } i \text{ is observable in } n \end{cases} \tag{19}$$

A point is defined as observable from a pose (figure 4) if it is projected inside images $\mathcal{I}_0$ and $\mathcal{I}_1$, if the angles between the normal vector to the point and the cameras views directions is small enough, and if the point is not occluded. To determine the occlusion, the initial point cloud model is meshed and rays from camera centers to the considered point are tested against this mesh to check for intersections. If some intersections are found, the point is considered occluded. To avoid occlusions caused from a bad registration (multiple skins effect on coarsly registered objects) a tolerance threshold on intersection distances can be used.

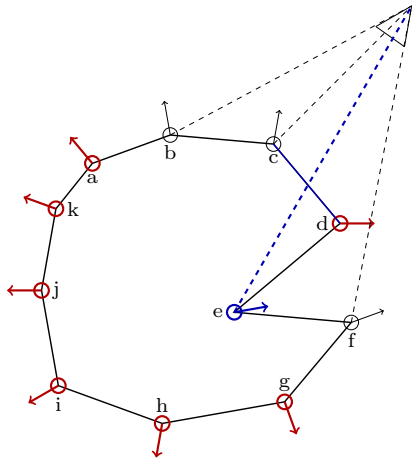


Fig. 4 For a given camera position, points to be used in global optimization are chosen regarding to the initial mesh and local normals of the object. In a two cameras setup, a point has to be observable by both cameras to be valid. *Red:* when the angle between normals and camera view direction is too large, point is not observable. *Blue:* the ray from the optical center to the point e intersects the model mesh, it is not observable

For the optimization process, we build a parameter vector $\mathbf{x}$. It is composed from system poses parameters $(\theta, \phi, \psi, x, y, z)^T \in \mathbb{R}^6$, respectively the three *Euler* angles and the three translation parameters, and from the reduced normal parameters of the points expressed by $(n_x/d, n_y/d, n_z/d)^T \in \mathbb{R}^3$. This parameter vector is estimated through a sparse Levenberg-Marquardt nonlinear least-squares optimization method [18], minimizing the criterion

$$\underset{\mathbf{x}}{argmin} \sum_{n=1}^{N_V} \sum_{i=1}^{N_P} \phi\,(i,n) \sum_{j \in W} (J - L)^2 \tag{20}$$

with

$$J = I_0^n \left[\mathbf{H}^{i}_{\mathcal{I}_0^n}\left(\mathbf{p}_{i,j}\right)\right] \quad j \in W \tag{21}$$

$$L = I_1^n \left[\mathbf{H}^{i}_{\mathcal{I}_1^n}\left(\mathbf{p}_{i,j}\right)\right] \quad j \in W \tag{22}$$

Since the scene changes between each acquisition, the dissimilarites measurement using luminance functions can only be done between images acquired at the same time. Equation (20) extends our two images correlation criterion. Points and poses are related between acquisitions by the inherent geometry of the observed scene. We optimize directly the normal parameters and observe their influence on the correlation results in regions transformed in a consistent way between every image. By doing this a robust estimation of our parameters can be achieved. This formulation still depends on the size of the region W but since the system becomes heavily constrained, it is more robust to small sizes. Indeed, ambiguities are reduced in small size regions by the geometrical constraints imposed by the formulation of our system. We demonstrate in section 4 that it allows a more accurate measurement of a scene by reducing the smoothing effect of correlation approaches and the ability to observe smaller details or sharper edges.

3.3 Densification

With a large number of points, the algorithm complexity leads to long computation times and memory size problems. Even with sparse implementations, refining every point from every acquisition is not easily manageable. However, not every points are needed to estimate finely the parameters of the problem.

We propose a sparse-to-dense approach to model an entire object with our method. Keypoints are selected in a sampling process that tends to pick a given number of points maximizing their visibility in the set of acquisitions. Moreover, the points having the worst correlation score after the initial step and the points located near the edges are discarded. Camera poses and keypoints parameters are estimated strongly using the method described in section 3.2. The refined model obtained is sparse and contains only refined 3D keypoints.

To densify the model, every remaining point from the initial model has to be refined. With a sufficient number of keypoints in the previous step, camera poses can be considered optimal[1].

Since points, in equation (20), have parameters independent from other points – points are only dependent on the camera poses – every point can be refined separatly. This reduces the complexity of the optimization process, and reduces memory issues and cache misses, improving execution time. Camera poses are fixed and each point is represented by a parameter vector $\mathbf{y} = (n_x/d, n_y/d, n_z/d)^T \in \mathbb{R}^3$, and are estimated by minimizing

$$\underset{\mathbf{y}}{argmin} \sum_{n=1}^{N_V} \phi(i,n) \sum_{j \in W} (J-L)^2 \quad \forall i \in [1..N_P] \qquad (23)$$

This approach is dependent on the initial process, since points have to be observed in the first place to be processed by our method. From the initial model, this approach allows to easily refine points in a time effective way, providing multi-view advantages to dense 3D modelling.

3.4 Initialization

Our multi-view optimization scheme uses the 3D model which is computed as described in Section 2.2. It requires that the different clouds be aligned so that the distance between the reprojection of the coarse 3D points and the reprojection of the final 3D points does not exceed the window size. If it does, the optimization does not converge. In order to initiate the multi-pairwise correlation step, it is possible to choose a larger window size when keypoints along the edges are discarded as described in Section 3.3. This does not violate the plane approximation. We observed that in this way the result of ICP is always far better than the maximum distance tolerated by our algorithm.

4 Experiments and Discussions

The presented method has been primarily designed for reverse engineering or control of mechanical parts applications. The evaluation is mainly focused on this type of objects. The experimental protocol is briefly introduced. A quantitative evaluation is proposed, observing the ability of our method and the improvement it brings compared to the standard stereovision approach. Some qualitative results are given at the end of the section.

4.1 Implementation Details

The first step consists in reconstructing a coarse model. This step is multithreaded and implemented in `C++`. The computation is fast enough for it to complete together with the scanning process. Models produced at this stage usually have around three to ten million points. Visibility map is computed using 3D fast intersection from CGAL [1]. Then, the model is sampled so that multi-pairwise correlation does not last more than a few minutes, which corresponds to a dozen thousand points.

[1] The poses are considered optimal from our method point of view, that is to say that they are fully observed and that no better convergence can be achieved from adding keypoints.

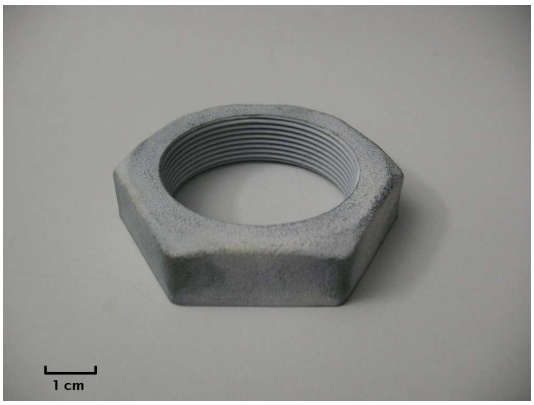


Fig. 5 The threaded nut used in our experiments. The thread is thin and can be diffult to observe

In this configuration, memory is not an issue. The minimization step uses the implementation of Lourakis [18]. The densification step takes a few more minutes. These two last steps run on a single core and could be improved.

4.2 Data and experimental protocol

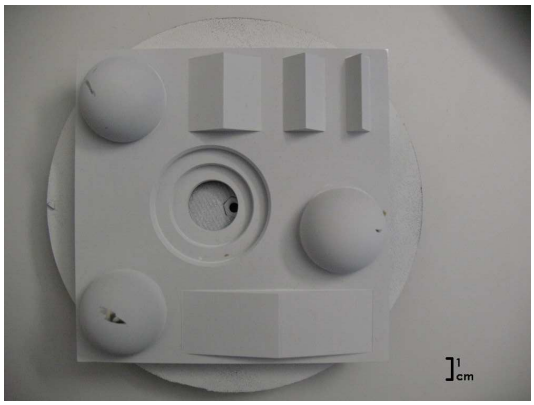
The device used for our experiments is made up of two cameras and a projector. The projector is built from a light projection and a glass tile printed with a speckle pattern. All the components are rigidly attached. The transformation between cameras is known and accurately calibrated. The intrinsic parameters of the cameras are also calibrated. The projector is not calibrated, it is simply set to provide an optimal focalization at the work distance of the stereo camera setup.

Each acquisition provides a pair of images illuminated with the speckle pattern from the projector. They are processed by a dense digital image correlation algorithm to match pixels between the two images, at subpixel resolution. Matched pixels are triangulated to provide a cloud of 3D points. Every reconstructed 3D point cloud is coarsely registered with previously acquired ones using a fast *ICP*-based approach [33].

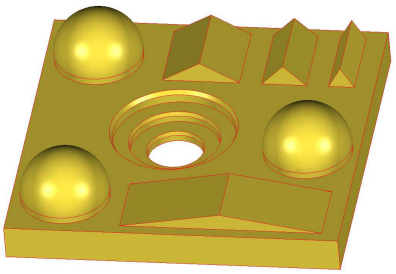
Threaded nut This is a mechanical part used for bolting in agricultural automotive applications. Even if the nut is large, the thread is fine and is not easy to observe using vision-based technologies (figure 5).

This object is used to compare the ability of our method to observe the thread depending on the number of views provided to the algorithm. It is also used to illustrate the contribution of the method compared to the standard pairwise stereovision method.

The part is made of steel. We covered it with a thin layer of matte white revealing powder to avoid specular reflections.



(a) Gauge block



(b) CAD model

Fig. 6 The gauge block composed from simple geometrical primitives. It contains several sharp edges. CAD model is used as ground truth.

Gauge block This part has been designed for measurement device calibration. It has been accurately manufactured and has an accurately known geometry. It is composed of several simple geometrical primitives (figure 6).

This object is used to observe the effect of our multi-view method on sharp edges. The CAD model of this object is used as ground truth in the evaluation of the results.

Turbine blade This part is a blade from a turbine (figure 7). It is made of metal and has also been covered with matte white revealing powder to avoid specular reflections.

Ball This ball has several seams (figure 8). It is used to evaluate the ability of our method to model fine details.

Maraca The maraca object is finely carved (figure 9). It is also used to evaluate the ability of our method to model fine details. The carving are shallow, approximately the size of the tip of a fingernail.

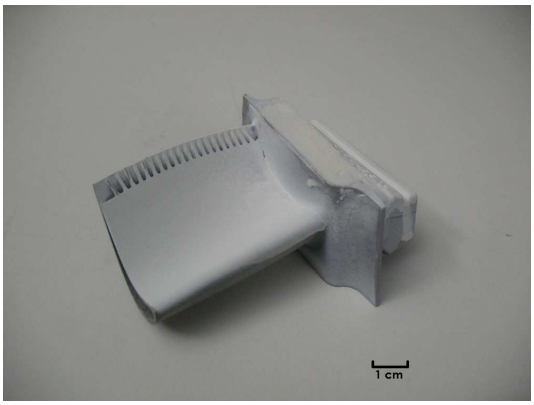


Fig. 7 Turbine blade

Fig. 8 Ball with details

Fig. 9 A carved maraca made of wood

4.3 Experiments

4.3.1 Influence of the number of image pairs

The first experiment uses the threaded nut object. Our method is evaluated in the reconstruction of the thread given various numbers of input pairs of images. The method has been tested providing three (3), five (5), seven (7) and nine (9) overlapping pairs of images. It as also been compared with the dense reconstruction provided by a single stereo pair. Figure 10 illustrates the results.

Following the columns shows the influence of a growing number of overlapping pairs, while following the rows shows the influence of increasing the size of the image correlation window W. The pairwise approach does not perform well with a small window size. Indeed, reducing the window size increases ambiguities in correlation and causes drifts in the stereo matching process, causing noise and holes. On the contrary, with an eleven (11) pixels window, the noise is reduced and no holes appear. On the other side, the window becomes too large compared to the size of the thread in the images causing a smoothing effect.

Adding two (2) overlapping pairs of images to the process contributes to greatly improve the result, even with very small size windows. The noise is largely reduced to a slightly granular effect on the surface while holes have almost completely disapeared. Increasing the window size also reduces the noise, but causing a slight smoothing effect. Comparing the result from one (1) image pair with an eleven (11) pixels window, and the result from three (3) image pairs with the same window size, one can note that the thread is more detailed with the larger number of views.

Figure 10 proves that increasing the number of image pairs tends to completely remove noise and holes while the details become fully observed. Indeed, the constraints added to the optimization system reduces drifts effect on the parameter vector. With five (5) pairs the result is already stable, and is not greatly improved with seven (7) pairs. Beyond five pairs the influence is not noticeable. Adding more pairs does not bring further improvements because the object is fully observed. This means that the system is able to reach its optimality with a given finite number of pairs, depending on the object and the quality of the points of view.

One can also note that, with an important number of pairs, the result is stable with any window size. As we said previously, an increasing number of pairs constrains the system and reduces ambiguities with small window sizes. On the other hand, the large number of views allows to observe completely the parameters of the system, and the least-squares solution can be efficiently found despite the approximation of the surface in tangent planes. It can be noted that using a smaller window size can be beneficial since it reduces the dimension of the problem and speeds-up the process.

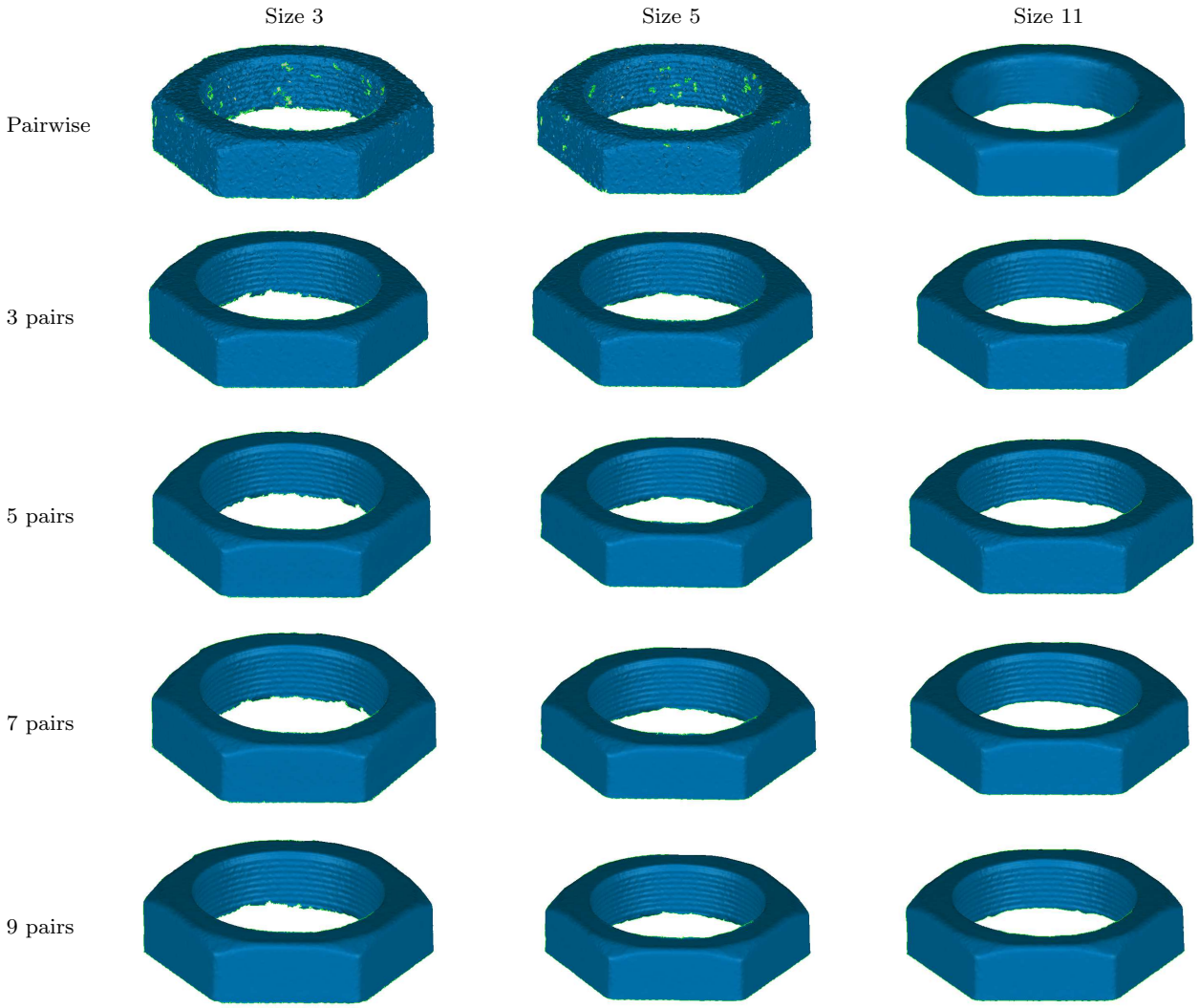


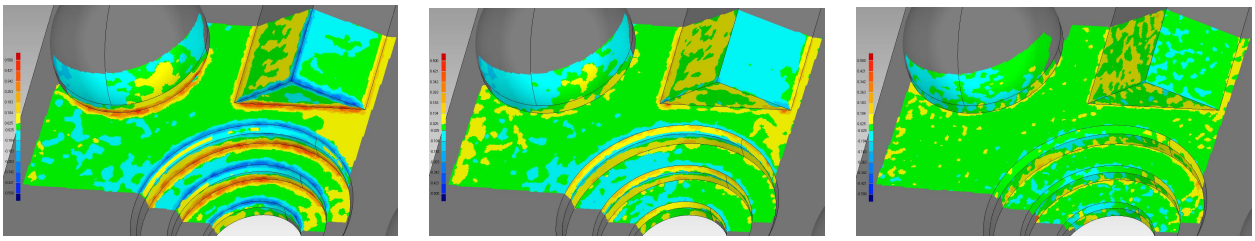
Fig. 10 Reconstruction of the threaded nut changing the number of input overlapping image pairs and correlation window size (in pixels). Increasing the number of pairs reduces noise and allows the algorithm to converge towards its optimal solution

4.3.2 Accuracy comparison

To give an absolute reference of the results provided by our method, we used a gauge block with a known geometry. The ground truth CAD model is available and used to evaluate reconstruction error. We compare our method to the standard pairwise stereo approach. As we have seen previously, this approach tends to smooth details and edges. The gauge block is made of a large plane and geometrical primitives intersecting the plane, creating sharp edges. Since the pairwise stereovision method performs poorly in this case, we improved the result using an edge sharpening method. We used a similar approach to that of Attene *et al.* [2] with a stereo image consistency check. This method allows to extend surface of geometrical primitives by surface subdivision. Subdivided polygons are checked in stereo images, to allow chamfers, rounds and fillets detections.

Models resulting from the three methods (pairwise stereovision, edge sharpening in pairwise stereovision and our multiview stereovision method) are registered on the CAD model of the gauge block. Points are projected orthogonally on the surfaces of the theoretical model and projection distance is measured for each point. This distance is considered as the reconstruction error of the point. Figure 11 represents these errors using a color map. Green areas are measured inside the tolerance range $[-25\mu m, +25\mu m]$. Points with a larger positive error are represented on a scale from yellow to red. Points with a larger negative error are represented on a scale from light to deep blue.

As expected, the error map of the pairwise stereovision method shows a large error near edges of the model.

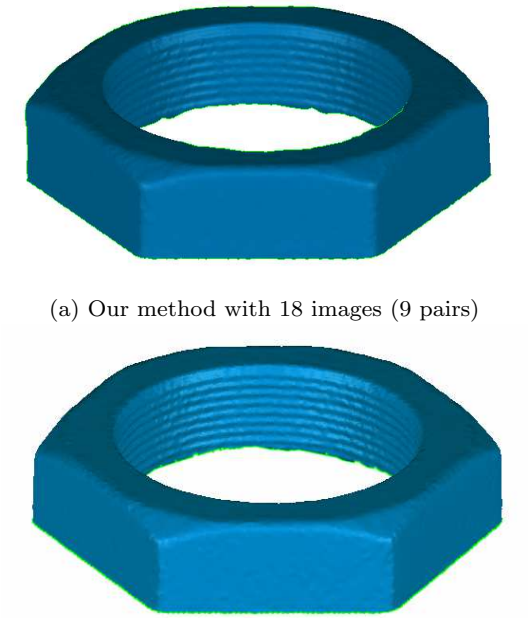


(a) Pairwise stereo (b) Edge optimization (c) Our MVS method

Fig. 11 Comparison of reconstruction error using pairwise stereovision, pairwise stereovision with edge detection and correction, and our multiview stereovision method using five pairs. Window size is set to 9 pixels.

	Pairwise stereo	**Edge optimization**	**Our method**
Max error (mm)	0.553	0.478	0.452
Min error (mm)	−0.606	−0.286	−0.446
Pos. mean (mm)	0.036	0.037	0.023
Neg. mean (mm)	−0.032	−0.031	−0.020
Std. dev. (mm)	0.066	0.049	0.033

Table 1 Reconstruction error to the ground truth of the gauge block. *Max/Min error* : maximal positive and negative errors. *Pos./Neg. mean* : means of, respectively, positive and negatives errors. *Std. dev* : standard deviation on the error set



(a) Our method with 18 images (9 pairs)

(b) PMVS with 18 images

Fig. 12 Comparison between the reconstruction of the threaded nut with our method and with PMVS2

Due to the smoothing effect of the surface approximation in the reconstruction process, edges are rounded. The error on the dominant plane is essentially contained in the tolerance range. The edge-optimized method performs a lot better along the edges. Maximal errors are now mainly located near complex edges – intersection of more than two surfaces – and along the chamfer of the central drilling. However a small error on the edges remains. With our method, error is still mainly located near edges, but the model is more homogeneously inside the tolerance. The chamfers in the central drilling are now more finely reconstructed. It is to be noted that the dominant plane is also more finely reconstructed. A significatively larger part of the object is considered inside the tolerance comparing to the other methods.

Table 1 summarizes the results for this experiment. The results are presented showing maximal and minimal errors. The mean error should be 0 *mm* since the models are registered minimizing the distance between point clouds and the surface of the CAD model. It is then more meaningful to present average errors in positive and negative parts. The standard deviation on the error set is also presented.

As it was pointed in figure 11, the edge-optimized algorithm and our MVS method present better results and a higher accuracy in terms of deviation to the theoretical surface. Our method offers the best results, with the lowest deviation.

4.3.3 Comparison with PMVS2

Our approach is not easily comparable with most state of the art methods since it is designed to address the problem of a moving projected texture. To provide a comparison we have used an external projector instead of the embedded one. The projection system is now fixed and projects a non moving pattern. We can then compare with appearance-based MVS methods. We decided to use PMVS2 software [15] because it is easily available as an open source software package and because it is one of the best performing algorithms in the evaluation of Seitz *et al.* [36]. PMVS2 is able to reconstruct a 3D structure of an object or a scene from a

set of images and camera parameters. We acquired nine (9) pairs of images of the threaded nut, used as eighteen (18) individual images by PMVS2 and as pairs by our method. We used the internal parameters of our system for both experiments. Figure 12 shows the results of the methods.

The result from PMVS2 is similar to the result of our method. This could be expected as both of these methods are patch based. The only difference is that we do not have a photometric relation between pairs whereas PMVS2 does have. This suggests that many multi-view stero algorithm could be tuned to work on a moving pattern.

Moreover, this experiment shows that our method can work perfectly on highly textured scenes or with external fixed projection and can compare to state-of-the-art methods.

Other examples of reconstructed objects with our MVS method are presented in figure 13.

5 Conclusion

In this paper we presented a novel method for correlation-based Multi-View Stereovision. This method is not dependant on the scene texture, that can change during the acquisition process. The method is based on the observation of correlation criterions linked through a generalized formulation of homographies. By writing a global criterion, this method can refine the poses of the visual sensors and the structure of the observed scene. After an initialization step, the model is refined. Due to an highly constrained formulation of the optimization problem with a non-linear least-squares scheme, we achieved an accurate modelling of objects, even with fine details or sharp edges.

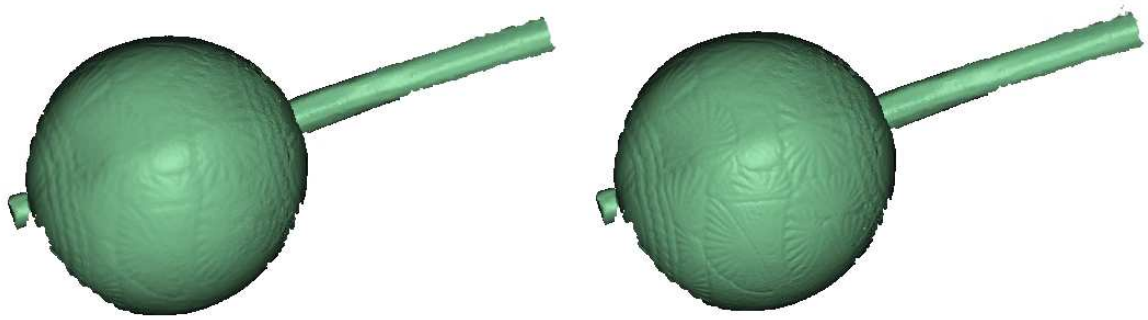
We evaluated our method on several sample objects and demonstrated a clear improvement compared to a pairwise stereovision approach. Comparing to a ground truth model, the effective reconstruction error has been evaluated and it has been shown that our method has the ability to provide results in a tolerance range consistent with most industrial applications.

This method is designed to work on the particular case of a projector moving along with the cameras, changing the scene at every acquisition time. The problem is not well addressed. While it is most of the time solved by pairwise or sequential approaches, our approach allows the use of multi-view stereovision.

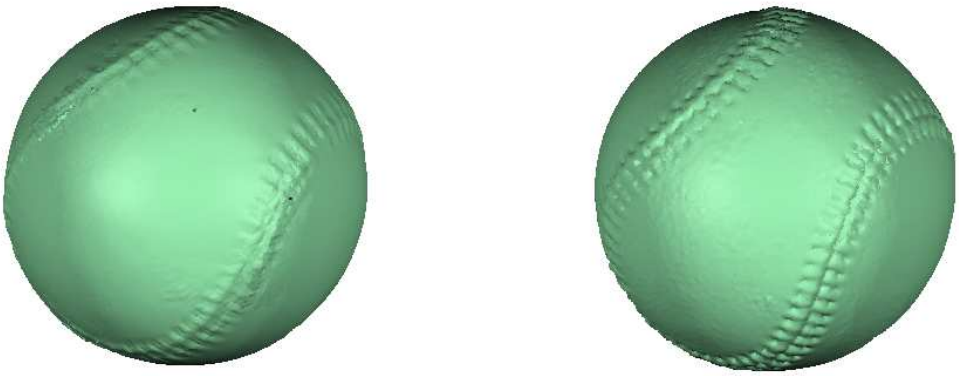
Our future works will focus on implementation improvements as general speed-up for solving large scale modelling problems.

References

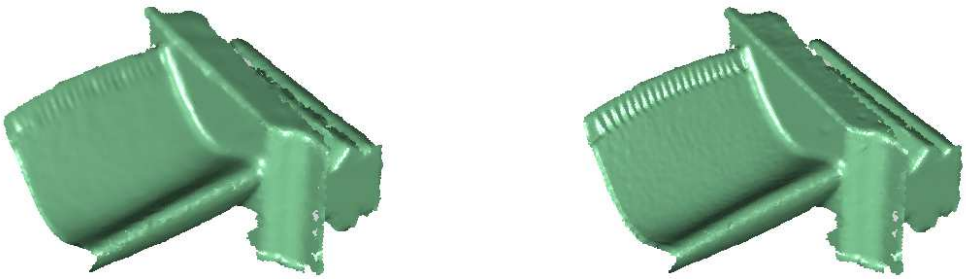
1. Alliez, P., Tayeb, S., and Wormser, C. 3d fast intersection and distance computation (aabb tree). http://www.cgal.org/Manual/latest/doc_html/cgal_manual/AABB_tree_ref/Chapter_introduction.html.
2. Attene, M., Falcidieno, B., Rossignac, J., and Spagnuolo, M. Edge-sharpener: Recovering sharp features in triangulations of non-adaptively re-meshed surfaces. In *ACM Symposium on Geometry Processing*, 2003.
3. Beder, C., Bartczak, B., and Koch, R. A comparison of pmd-cameras and stereo-vision for the task of surface reconstruction using patchlets. In *CVPR*, 2007.
4. Blais, F and Rioux, M. Real-time numerical peak detector. *Signal Processing*, 11(2):145–155, 1985.
5. Carceroni, R.L. and Kutulakos, K.N. Multi-view scene capture by surfel sampling: From video streams to non-rigid 3d motion, shape and reflectance. *International Journal of Computer Vision*, 49(2), 2002.
6. Chang, J.Y., Lee, K.M., and Lee, S.U. Multiview normal field integration using level set methods. In *CVPR*, 2007.
7. Coudrin, B., Devy, M., Orteu, J.J., and Brèthes, L. Registration strategies of 3d images acquired from a hand-held visual sensor. In *3D-IMS, Conference on 3D-Imaging of Materials and Systems*, September 2010.
8. Coudrin, B., Devy, M., Orteu, J.J., and Brèthes, L. An innovative hand-held vision-based digitizing system for 3D modelling. *Optics and Lasers in Engineering*, 49:1168–1176, 2011.
9. Curless, B. and Levoy, M. A volumetric method for building complex models from range images. In *SIGGRAPH*, 1996.
10. Davison, A.J. Real-time simultaneous localisation and mapping with a single camera. In *International Conference on Computer Vision*, pages 1403–1410, 2003.
11. Davison, A.J., Reid, I.D., Molton, N.D., and Stasse, O. Monoslam: Real-time single camera slam. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29:2007, 2007.
12. Ferreira, J.F., Lobo, J., and Dias, J. A 3D scanner – Three-dimensional reconstruction from multiple images. In *First International Symposium on 3D Data Processing Visualization and Transmission (3DPVT'02)*, 2002.
13. Furukawa, Y., Curless, B., Seitz, S.M., and Szeliski, R. Clustering views for multi-view stereo, 2011. http://grail.cs.washington.edu/software/cmvs.
14. Furukawa, Y. and Ponce, J. Accurate camera calibration from multi-view stereo and bundle adjustment. *Int. J. Comput. Vision*, 84:257–268, 2009.
15. Furukawa, Y. and Ponce, J. Patch-based multi-view stereo software, 2010. http://grail.cs.washington.edu/software/pmvs.
16. Hartley, R.I. and Zisserman, A. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.
17. Hernández, C. and Schmitt, F. Silhouette and stereo fusion for 3d object modeling. *Comput. Vis. Image Underst.*, 96:367–392, 2004.
18. Lourakis, M. Sparse non-linear least squares optimization for geometric vision. In *European Conference on Computer Vision*, volume 2, pages 43–56, 2010.
19. Lu, F. and Milios, E. Globally consistent range scan alignment for environment mapping. *Autonomous Robots*, 4:333–349, 1997.
20. Matabosch Geronès, C. *3D hand-held sensor for large surface registration*. PhD thesis, Universitat de Girona, Girona, Juin 2007.



(a) Engraved maraca. *Left* : Standard stereovision method. *Right* : Our multi-view stereo method



(b) Ball with seams. *Left* : Standard stereovision method. *Right* : Our multi-view stereo method



(c) Turbine blade. *Left* : Standard stereovision method. *Right* : Our multi-view stereo method

Fig. 13 Some reconstruction examples using pairwise stereovision method and our MVS method with five pairs and a window size of 7 pixels

21. Montiel, J.M.M., Civera, J., and Davison, A.J. Unified inverse depth parametrization for monocular slam. In *RSS*, 2006.
22. Murray, D. and Little, J.J. Patchlets: Representing stereo vision data with surface elements. In *WACV, IEEE Workshop on the Applications of Computer Vision*, 2005.
23. Newcombe, R.A. and Davison, A.J. Live dense reconstruction with a single moving camera. In *CVPR 2010*, 2010.
24. Nieto, J., Bailey, T., and Nebot, E. Scan-slam : Combining ekf-slam and scan correlation. In *FSR*, 2005.
25. Nüchter, A., Lingemann, K., Hertzberg, J., and Surmann, H. 6-d slam-3d mapping outdoor environments. *J.Field Robot.*, 24, 2006.
26. Nüchter, A., Surmann, H., Lingemann, K., Hertzberg, J., and Thrun, S. 6d slam with an application in autonomous mine mapping. In *IEEE International Conference on Robotics and Automation*, volume 2, pages 1998–2003, 2004.
27. Olson, C.F., Matthies, L.H., Schoppers, M., and Maimone, M.W. Rover navigation using stereo ego-motion. *Robotics and Autonomous Systems*, 43(4):215 – 229, 2003.
28. Pan, Q., Reitmayr, G., and Drummond, T.W. Interactive model reconstruction with user guidance. In *Mixed and Augmented Reality, IEEE / ACM International Symposium on*, pages 209–210, 2009.
29. Park, J. and Kak, A.C. Multi-peak range imaging for

accurate 3d reconstruction of specular objects. In *6th Asian Conference on Computer Vision*, 2004.
30. Paz, L.M., Pinies, P., Tardos, J.D., and Neira, J. Large-scale 6-dof slam with stereo-in-hand. *IEEE Trans. on Robotics*, 24, 2008.
31. Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J., and Koch, R. Visual modeling with a hand-held camera. *International Journal of Computer Vision*, 59:207–232, 2004.
32. Pons, J.P., Keriven, R., and Faugeras, O. Modelling dynamic scenes by registering multi-view image sequences. In *CVPR*, 2005.
33. Rusinkiewicz, S. and Levoy, M. Efficient variants of the ICP algorithm. In *Third International Conference on 3D Digital Imaging and Modeling (3DIM)*, Juin 2001.
34. Scharstein, D. and Szeliski, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47:7–42, 2002.
35. Wildes, R. Se, S., Jasiobedzki, P. Stereo-vision based 3d modeling of space structures. *Sensors and Systems for Space Applications*, 6555, 2007.
36. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 519–528, 2006.
37. Silva, L., Bellon, O.R.P., and Boyer, K.L. Multiview range image registration using the surface interpenetration measure. *Image and Vision Computing*, 25(1):114 – 125, 2007.
38. Sola, J., Monin, A., and Devy, M. Bicamslam: two times mono is more than stereo. In *2007 IEEE International Conference on Robotics and Automation (ICRA'07)*, pages 4795–4800, Rome (Italy), 2007.
39. Strobl, K.H., Mair, E., Bodenmüller, T., Kielhöfer, S., Sepp, W., Suppa, M., Burschka, D., and Hirzinger, G. The Self-Referenced DLR 3D-Modeler. In *IEEE International Conference on Intelligent Robots and Systems (IROS'09)*, October 2009.
40. Sutton, M.A., Orteu, J.J, and Schreier, H.W. *Image Correlation for Shape, Motion and Deformation Measurements – Basic Concepts, Theory and Applications*. Springer, 2009.
41. Szeliski, R. *Computer Vision: Algorithms and Applications*. Springer, 2011.
42. Triggs, B., Mclauchlan, P., Hartley, R., and Fitzgibbon, A. Bundle adjustment – a modern synthesis, 2000.
43. Vu, H.H., Labatut, P., Keriven, R., and Pons, J.P. High accuracy and visibility-consistent dense multi-view stereo. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2012.
44. Wu, C., Wilburn, B., Matsushita, Y., and Theobalt, C. High-quality shape from multi-view stereo and shading under general illumination. In *CVPR*, 2011.
45. Zaharescu, A., Boyer, E., and Horaud, R.P. Topology-adaptative mesh deformation for surface evolution, morphing, and multi-view reconstruction. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2011.
46. Zhang, Z. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1330–1334, 2000.
47. Chen, Z. Zheng, Z., Ma, L. An extended photometric stereo algorithm for recovering specular object shape and its reflectance properties. *ComSIS*, 7(1), 2010.