Volumetric multi-camera scene acquisition with partially metric calibration for wide-area tele-immersion

X. Zabulis, J. P. Barreto, N. Kelshikar, R. Molana, K. Daniilidis

GRASP Laboratory,University of Pennsylvania,
Levine Hall L402,3330 Walnut Street, Philadelphia, PA 19104-6228, USA
{zabulis,jpbar,nikhil,molana,kostas}@grasp.cis.upenn.edu

Abstract. Tele-immersion is a new medium that tries to create the illusion of virtual collocation among physically distant places. To create this sense of co-presence, tele-immersion has to be visually compelling and run in real-time, putting, thus, high performance constraints in all three areas of computer vision, graphics, and networking. We describe our newest results in the scene acquisition component of tele-immersion: algorithms for volumetric scene reconstruction from multiple cameras and for calibration of camera clusters. The reconstruction algorithm is volumetric, makes no assumption on camera loci and is based on the detection of local correlation maxima in 3D. It outputs an occupancy voxel grid, with occupied voxels being accompanied by a surface normal; a fact that improves reconstruction quality. The calibration of cameras distributed in a wide area is a challenging task because it is impossible to use reference objects visible to all cameras and because wide field-of-view cameras suffer under radial distortion. Our calibration method uses a reference object to calibrate a minimum of two cameras in order to provide a euclidean world coordinate system. Then, we deliberately move an LED in front of all cameras to obtain correspondances between views. The projection matrices and radial distortion parameters of all cameras are computed using two-step factorization. The combination of the two algorithms alleviates "misregistration" [8] artifacts, encountered when concatenating reconstructions/calibrations, obtainbed from independent stereos.

1 Introduction

Reconstruction with multiple cameras over a wide area is a challenge for stereo algorithms. Multiple viewpoints contribute to scene completeness by capturing otherwise occluded areas and enabling reconstructions of occluding contours. There is no notion of a single "depth" dimension or a 2+1/2D map since for an arbitrary placement of cameras there is no semantic difference in x, y, and z directions of the world. We need a volumetric representation and algorithms that can correlate images from multiple cameras and handle visibility.

The state of the art is characterized by three main streams: (1) silhouette-based approaches (e.g. [7]) with their weaknesses in concave scene components, (2) photo-consistency and space carving [**?**] with their constraining assumptions on color constancy among cameras and (3) combination of correlation stereos, with possible optimization over depths [6] or volume occupancies [4], but which impose restrictions on camera loci or are not guaranteed to converge respectively.

A volumetric algorithm can not perform without accurate calibration including intrinsic parameters, radial distortion, and spatial registration (aka extrinsic parameters).

Wide-area calibration is challenging due to the non-existent overlapping field of view which does not allow the use of a metric reference object like a dotted floor or a checkerboard. Camera calibration errors cause the same world point to be reconstructed in different loci, in independent reconstructions. The error is systematic [8] and, when concatenating reconstructions, its effect is a "ghosting" of surfaces making them appear multiple times one in front of each other. To fix the world coordinate frame we use a euclidean calibration of a minimum of two cameras. Then we deliberately move an LED in thousands of positions in front of the sixty cameras. Projection matrices are recovered without any assumption on the 3D position of the points and a novel scheme recovers the radial distortion. The new calibration outperforms metric calibration due to the large effective field of view.

2 A new approach to surface reconstruction

A operator is introduced that when applied at a world point $\boldsymbol{p} \in R^3$ outputs confidence score $s(\boldsymbol{p})$ (strength) and a 3D unit normal $\boldsymbol{\kappa}(\boldsymbol{p})$ (orientation), given a strongly calibrated image pair (I_1, I_2). Let a planar surface patch $\mathcal{S}_{\boldsymbol{p}}^{\boldsymbol{n}}$, which size is $\alpha \times \alpha$ units of length (mm), centered at $\boldsymbol{p}$, with unit normal $\boldsymbol{n}$. Backprojecting I_1 and I_2 onto $\mathcal{S}_{\boldsymbol{p}}^{\boldsymbol{n}}$ yields two images $w_1(\boldsymbol{p},\boldsymbol{n})$ and $w_2(\boldsymbol{p},\boldsymbol{n})$. Their formation rule is: $w_i(\boldsymbol{p},\boldsymbol{n}) = I_i\left(P_i \cdot \left(\boldsymbol{p} + R(\boldsymbol{n}) \cdot [\Delta x\, \Delta y\, 0]^T\right)\right),\ i \in \{1,2\}$, where P_i is the projection matrix of camera i, $R(\boldsymbol{n})$ is a 3×3 rotation matrix that maps $[0\,0\,1]^T$ to $\boldsymbol{n}$, and $\Delta x,\ \Delta y \in [0, \alpha]$ are local horizontal and vertical coordinates of a point on the patch. Fig. 1(a) illustrates the above process, which in effect is a collineation.

To obtain $s(\boldsymbol{p})$ and $\boldsymbol{\kappa}(\boldsymbol{p})$, the correlation of $w_1(\boldsymbol{p},\boldsymbol{n})$ and $w_2(\boldsymbol{p},\boldsymbol{n})$, $Corr(w_1(\boldsymbol{p},\boldsymbol{n}), w_2(\boldsymbol{p},\boldsymbol{n}))$, is optimized with $\boldsymbol{n}$ as the free variable: $s(\boldsymbol{p}) = \max_{\boldsymbol{n}}\left(Corr(w_1(\boldsymbol{p},\boldsymbol{n}), w_2(\boldsymbol{p},\boldsymbol{n}))\right)$, $\boldsymbol{\kappa}(\boldsymbol{p}) = argmax_{\boldsymbol{n}}\left(Corr(w_1(\boldsymbol{p},\boldsymbol{n}), w_2(\boldsymbol{p},\boldsymbol{n}))\right)$. Scalar $s(\boldsymbol{p})$ and vector $\boldsymbol{\kappa}(\boldsymbol{p})$ are combined into vector $\boldsymbol{v}(\boldsymbol{p})$, so that $|\boldsymbol{v}(\boldsymbol{p})| = s(\boldsymbol{p})$ and $\boldsymbol{v}(\boldsymbol{p})/|\boldsymbol{v}(\boldsymbol{p})| = \boldsymbol{\kappa}(\boldsymbol{p})$ hold. At $\boldsymbol{p}$'s where $s_i(\boldsymbol{p}) < 0$ we set $\boldsymbol{v} = \boldsymbol{0}$. We note $w_i'(\boldsymbol{p}) = w_i\left(\boldsymbol{p}, \kappa(\boldsymbol{p})\right)$ and assume that I_1 and I_2 image Lambertian and locally planar textured surfaces.

When I_1 and I_2 are continuous and $\mathcal{S}_{\boldsymbol{p}}^{\boldsymbol{n}}$ is tangent at a world surface, $w_1'(\boldsymbol{p})$ and $w_2'(\boldsymbol{p})$ are identities of the world texture on this surface, because they collineate the same world points in a world-coordinate regular parameterization. Thus $I_1(P_1\boldsymbol{x}) = I_2(P_2\boldsymbol{x}),\ \forall \boldsymbol{x} \in \mathcal{S}$ and therefore correlation is optimal. In contrast, correlation is suboptimal when $\mathcal{S}_{\boldsymbol{p}}^{\boldsymbol{n}}$ is not tangent to the surface, because $w_1(\boldsymbol{p},\boldsymbol{n})$ and $w_2(\boldsymbol{p},\boldsymbol{n})$ image different world surface patches. In practice, the proposed operator is defined in the discrete domain. Let a lattice $\mathcal{L}$ of $r \times r$ points on $\mathcal{S}_{\boldsymbol{p}}^{\boldsymbol{n}}$. We project the points of $\mathcal{L}$ on I_1 and I_2 and sample at these projections, to form images $w_1(\boldsymbol{p},\boldsymbol{n})$ and $w_2(\boldsymbol{p},\boldsymbol{n})$. Then $\boldsymbol{s}(\boldsymbol{p})$ and $\boldsymbol{s}(\boldsymbol{p})$ are computed as denoted above.

The number of pixels from which the collineations are sampled is a monotonically decreasing function of $\mathcal{S}$'s obliqueness, and thus, so is their sample variance. The effect is a spurious increment of the correlation value. We restrict $\mathcal{S}$ from obtaining very oblique postures by introducing threshold τ_O, measured in units of angle.

Using multiple cameras Let the M binocular pairs defined within a *tuple* of cameras. The output of the operator is function $\boldsymbol{t}$. For each $\boldsymbol{p}$, vectors $\boldsymbol{v}_i(\boldsymbol{p}), i \in \{1,2,3,\ldots,M\}$, are computed, where $\boldsymbol{v}_i$ is the result of the operator for pair i. At $\boldsymbol{p}$'s where $s_i(\boldsymbol{p}) < 0$ we set $\boldsymbol{v}_i = \boldsymbol{0}$. The magnitude of the result vector $\boldsymbol{t}(\boldsymbol{p}) = \prod_{i=1}^{M} |\boldsymbol{v}_i(\boldsymbol{p})|$ and its unit direction $\boldsymbol{\kappa}(\boldsymbol{p}) = \sum_{i=1}^{M} \boldsymbol{v}_i(\boldsymbol{p}) / \sum_{i=1}^{M} |\boldsymbol{v}_i(\boldsymbol{p})|$. Multiplying the magnitudes of vectors

$\boldsymbol{v}_i$ is a conservative approach towards reconstructing surfaces: only surface segments that are viewed by *all* cameras are likely to exceed a high correlation threshold.
Best tuple (view) selection Given T tuples, we compute $\beta(\boldsymbol{p}) = argmax_j(|\boldsymbol{v}_j(\boldsymbol{p})|)$, $\beta(\boldsymbol{p})$ and $j \in \{1, 2, \ldots, T\}$ and $\boldsymbol{V}(\boldsymbol{p}) = \max_j(\boldsymbol{v}_j(\boldsymbol{p}))$. Vector $\boldsymbol{\kappa}_\beta(\boldsymbol{p})$ $(= \boldsymbol{V}(\boldsymbol{p})/|\boldsymbol{V}(\boldsymbol{p})|)$ corresponds to the highest correlation value ($s_\beta(\boldsymbol{p})$) and, thus, is the best estimation of all tuples for a voxel. That is because tuple j exhibits the most matching backprojections which, in turn, best match to the underlying surface pattern. Different estimations of the normal can be obtained due the value of τ_O. For the tuples that the surface normal is out of their search range, backprojections are unequal and we obtain a small $t(\boldsymbol{p})$.
Occupancy detection Spatially local maxima of function $\boldsymbol{v}(\boldsymbol{p})$ that exhibit a high correlation value are regarded as corresponding to occupied voxels. Their detection is performed by an extension of Canny's edge detection algorithm [3], in 3D. Vectors $\boldsymbol{V}(\boldsymbol{p})$ are considered as gradient vectors. Two correlation thresholds (τ_L, τ_H) are introduced for the, last, hysteresis thresholding step.

Occupied voxels are reconstructed by centering a surface patch at their centers ($\boldsymbol{p}$) and orienting it as $\boldsymbol{\kappa}(\boldsymbol{p})$. Texture is backprojected on it from a camera of tuple $\beta(\boldsymbol{p})$. Thus reconstruction is more detailed than fulfilling an occupied voxel with a particular "color". Surface normal information accelerates rendering and considerably simplifies recovery of surface connectedness.

Experiments We compared our operator to the traditional approach of correlating images patches (as implemented in [8]), as to their efficacy in correspondence establishment, using identical experimental conditions and input, for a binocular pair. We observed that more surface was reconstructed using the proposed operator, as a result of the 2-degree rotational freedom of $\mathcal{S}$. The additional surface tends to be located at depth discontinuities and slanted surfaces In this paper, we demonstrate our algorithm in room-size scenes, acquired from 8 cameras tuples ($M = 3$). Fig. 1(b,c) shows reconstructions of a $2 \times 2.22 \times 2.7\, m^3$ volume in a tessellation of $d = 10\, mm$. No misregistration occured in the results, and wall regions that were not reconstructed were not visible by any camera. The reconstruction required 16 hours on a $2.4\, GHz$, Pentium 4 DELL machine. We then tested our reconstruction algorithm as to its behavior in the misregistration of tuples. We first computed maps $\boldsymbol{v}$ for the same volume, for 5 camera tuples ($M = 3$). We then obtained 2 reconstructions: one by concatenating the reconstructions of each $\boldsymbol{v}_j$, and one for the combined $\boldsymbol{v}$. Tuples were horizontally arranged against the curtain at the back end of the room; their eccentricities within $[0°, 90°]$. Fig. 1(d-k) shows the detected local maxima for the independent and combined strength maps, across a horizontal slice. We explain the alleviation of the ghosting effect as the result of the non-maxima suppression step of Canny's algorithm. An occluding object created "holes" in the individual strength maps, but the occlusion has been compensated in the combined map.

3 Wide-area Multiple Camera Calibration Including Lens Distortion

In order to perform volumetric reconstruction the cameras must be accurately calibrated. Wide-area calibration is challenging due to the non-existent overlapping field of view which does not allow the use of a metric reference object like a dotted floor or a

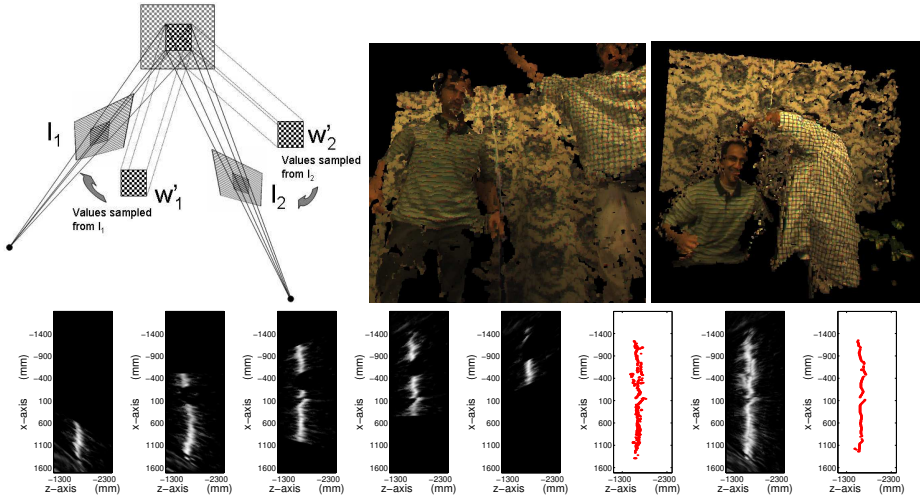

Fig. 1. Top to botton and left to right: (a) A surface is projectively distorted in images ($I_{1,2}$), but the collineations ($w_{1,2}$) from a planar patch tangent to this surface are not. Values of this collineation are regularly sampled from the images at the points where this planar patch projects. (b,c) Reconstructions of a room with two persons ($d = 10\,mm$, $a = 10\,mm$, $r = 11$ pixels, $\tau_H = 0.5$, $\tau_L = 0.3$, $\tau_O = 60^\circ$, $\sigma = 1$, $\tau_V = 0$, image size 480×640 pixels). (d-h) Individual strength maps for tuples 1 to 5. (i) Concatenated local maxima, detected from individual strength maps. (j) Combined correlation map. (k) Detected local maxima from combined strength map. ($d = 10\,mm$, $a = 10\,mm$, $r = 11$ pixels, $\tau_H = 0.5$, $\tau_L = 0.3$, $\tau_O = 60^\circ$, $\sigma = 1$, and $\tau_V = 0$).

checkerboard. Moreover, since the cameras usually have a large field of view (FOV), the lens distortion can not be neglected. Previous works, like [9], try to solve the problem by: *a)* Calibrate, up to a colineation $\mathbf{H}$, using multi-view factorization [11, 10] or pair wise fundamental matrices; *b)* Apply euclidean stratification to estimate $\mathbf{H}$; *c)* Compute the radial distortion using non-linear optimization.

Our method, descibed in detail in [1], presents the following novelties

- The option of pre-calibrating a minimum of two cameras in order to avoid euclidean stratification. The pre-calibration is performed using standard techniques [2].
- Two step multi-view factorization to simultaneously recover the projection matrices and radial distortion. The method is computationally efficient since the solution is computed by solving an eigen-problem and no nonlinear minimization is required.

The method requires synchronous image acquisition and can be used to calibrate any network where every camera has overlapping FOV with at least another camera. The proposed approach is completely new and, as far as we know, it has never appeared in the literature.

3.1 Image formation model

The scheme of Fig. 2 shows the assumed model for image formation. Point $\mathbf{u} = (u_x, u_y, u_z)^t$ is the image of a generic 3D point $\mathbf{X}$. The projection matrix $\mathbf{P}$ has dimension 3×4. It depends on the intrinsic parameters and on the camera pose with

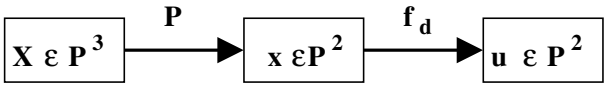

Fig. 2. Mapping model including radial distortion

respect to the world system of coordinates. A point $\mathbf{X}$ in the scene is transformed into a 2D point $\mathbf{x} = (x, y, z)^t$ such that $\lambda\mathbf{x} = \mathbf{PX}$ (λ denotes the affine depth). If the camera lens is distortion free then $\mathbf{u} = \mathbf{x}$. However, in the presence of distortion, $\mathbf{x}$ and $\mathbf{u}$ are related by a non-linear mapping function f_d which must be taken into account. We model the radial distortion using the so called division model [5]. Consider points $\mathbf{x}$ and $\mathbf{u}$ expressed in a 2D coordinate system with origin at the distortion center. The relation between the two points is given by equation 1, where the radial distortion is parameterized by ξ. Notice that the model requires the distortion center to be known. In the absence of any other information, we can place it at the image center without significantly affect the correction

$$\mathbf{x} = \mathbf{f_d^{-1}}(\mathbf{u}, \xi) = \mathbf{u} + \xi \underbrace{\left(0, 0, \frac{u_x^2 + u_y^2}{u_z}\right)^t}_{\mathbf{v}} \tag{1}$$

3.2 Calibration Algorithm

Consider a set K of cameras $\mathbf{C_1}$, $\mathbf{C_2}$, $\mathbf{C_3}$, ..., $\mathbf{C_K}$ spatially distributed in a room as shown in Fig. 3(a). For each camera $\mathbf{C_i}$ we aim to estimate the corresponding projection matrix $\mathbf{P_i}$ and distortion parameter ξ_i. Assume that the first M cameras are calibrated in advance using standard techniques [2] ($M \geq 2$).

The input for the calibration procedure is the multiple view of a set of N_p points in the scene. However, finding points $\mathbf{u}$ and establishing correspondences between multiple images is a difficult task. If the cameras are synchronized then the problem can be solved using a a laser pointer or a LED in a similar way as proposed in [9]. The user is required to move the laser/LED throughout the working volume that should be dark. The illumination conditions provide enough contrast such that the point projection can be accurately measured by performing a simple image threshold.

The pre-calibration of a minimum of two cameras is required in order to estimate the affine depth λ with respect to the first camera. The depth of all the points simultaneously viewed by $\mathbf{C_1}$ any other calibrated camera, can be easily computed by solving an eigenproblem [1, 11]. The projection matrix $\mathbf{P_i}$ and the distortion parameter ξ_i are estimated from the point correspondances between views $\mathbf{C_i}$ and $\mathbf{C_1}$. In [1], we propose the first algorithm in the literature for radial distortion estimation from multiple views. The algorithm is based on the application of two subspace approximation steps. At these steps, the estimated approximate solution for a matrix can be projected to the manifold of the parameter space by adjusting the singular values.

3.3 Experimental Results

Fig. 3(a) is an image of our experimental setup used for tele-immersion. There are 52 cameras, grouped in 13 clusters of 4 cameras each, spread all over the room. Fig. 3(b), generated from the final calibration results, shows the location of the different

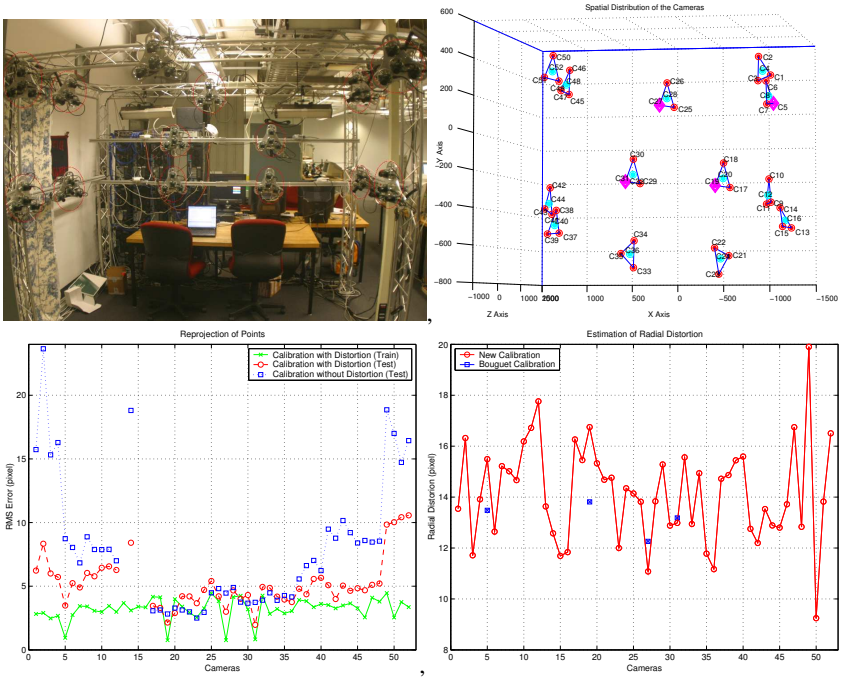

Fig. 3. Experimental results in calibrating a set of 52 cameras for tele-immersion application. a) Experimental setup. b) Camera location. c) Reprojection error. d) Radial distortion estimation

cameras. An LED is waved through the room, in order to generate 2500 virtual points. The corresponding image points are determined by simple thresholding. Cameras 5, 19, 27 and 31 are initially calibrated using the Bouguet Calibration Toolbox [2].

The 3D points visible in at least two calibrated views are reconstructed. From this set we select 100 test points, uniformly spread in the working volume. The remaining points (training set) are used in order to determine the projection matrices and the radial distortion parameters of the 52 cameras. Fig. 3(d) compares the radial distortion, estimated using our method, with the distortion measured with Bouguet for the 4 initial cameras. The results present the same order of magnitude which is reasonable taking into account that the cameras are all similar. After calibrating, the training set is reconstructed and reprojected in the image plane. Fig. 3(c) shows, for each view, the corresponding RMS error. The error in reprojecting the test points, reconstructed using the Bouguet calibration, is also exhibited. Despite the fact that these test points are not ground truth, it is reasonable to claim that the obtained calibration is Euclidean. The third curve refers to the reprojection of the test points when the camera set is calibrated without taking into account the radial distortion. As can be observed the RMS error is significantly higher for the cameras laterally positioned in the room. The test points lying in the FOV of these cameras are projected at the image sides in the views initially calibrated with Bouguet. Due to the effect of distortion these points are not correctly reconstructed. This explains the observed RMS error and shows the importance of taking into account the radial distortion.

References

[1] Joao P. Barreto and Kostas Daniilidis. Wide-area multiple camera calibration and estimation of lens radial distortion. In *Technical Report*, GRASP Lab, University of Pennsylvania, Philadelphia, September 2003.

[2] J Y Bouguet. Camera calibration toolbox for matlab. 2003.

[3] J. F. Canny. A computational approach to edge detection. *PAMI*, 8(6):679–698, 1986.

[4] O. Faugeras and R. Keriven. Complete dense stereovision using level set methods. In *Proc. ECCV 98*, volume 1, pages 379–393, 1998.

[5] A. Fitzgibbon. Simultaneous linear estimation of multipleview geometry and lens distortion. In *In Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Haway, USA, 2001.

[6] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *Proc. ECCV 02*, volume 1, pages 379–393, 1998.

[7] A. Laurentini. The visual hull concept for silhouette-based image understanding. *PAMI*, 16(2):150–162, 1994.

[8] J. Mulligan, X. Zabulis, N. Kelshikar, and K. Daniilidis. Stereo-based environment scanning for immersive telepresence. *IEEE Transactions on Circuits and Systems for Video Technology.* (to appear).

[9] T. Svoboda. Quick guide to multi-camera self calibration. In *Technical Report BiWi-TR-263*, Computer Vision Lab, Swiss Federal Institute of Technology, Zurich, August 2003.

[10] Bill Triggs. Factorization methods for projective structure and motion. In *In Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, 1996.

[11] J. Kosecka Y. Ma, S. Soatto and S. Sastry. *An Invitation to 3D Vision. From Images to Geometric Models*. Interdisciplinary Applied Mathematics. Springer, 2003.