Unified Direct Visual Tracking of Rigid and Deformable Surfaces Under Generic Illumination Changes in Grayscale and Color Images

Geraldo Silveira, Ezio Malis

International Journal of Computer Vision, **Vol. 89, No. 1, pp. 84–105, 2010**
This is a preprint of the accepted article

Unified direct visual tracking of rigid and deformable surfaces under generic illumination changes in grayscale and color images

Geraldo Silveira · Ezio Malis

Received: 1 September 2009 / Accepted: 5 February 2010 / Published online: 24 February 2010

Abstract The fundamental task of visual tracking is considered in this work as an incremental direct image registration problem. Direct methods refer to those that exploit the pixel intensities without resorting to image features. We propose new transformation models and optimization methods for directly and robustly registering images (including color ones) of rigid and deformable objects, all in a unified manner. We also show that widely adopted models are in fact particular cases of the proposed ones. Indeed, the proposed general models combine various classes of image warps and ensure robustness to generic lighting changes. Finally, the proposed optimization method together with the exploitation of all possible image information allow the algorithm to achieve high levels of accuracy. Extensive experiments are reported to demonstrate that visual tracking can indeed be highly accurate and robust despite deforming objects and severe illumination changes.

Keywords Visual tracking · Image registration · Nonlinear optimization · Direct methods · Nonrigid objects · Lighting changes · Robust techniques · Robotics

Electronic Supplementary Material The online version of this article contains supplementary material, which is available to authorized users.

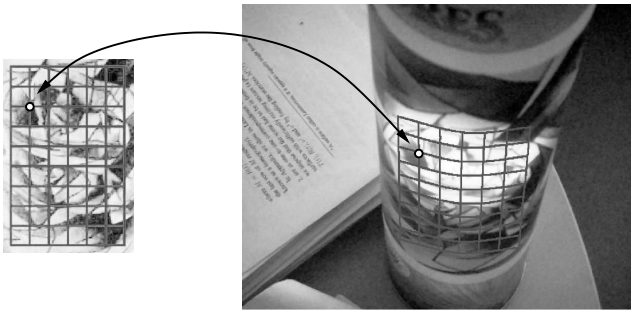
G. Silveira
CTI Renato Archer
Rod. Dom Pedro I, km 143,6, Amarais
CEP 13069-901, Campinas/SP, Brazil
E-mail: Geraldo.Silveira@cti.gov.br

E. Malis
INRIA Sophia-Antipolis
2004 Route des Lucioles, BP 93
06902 Sophia-Antipolis Cedex, France
E-mail: Ezio.Malis@sophia.inria.fr

1 Introduction

Visual tracking of an object of interest can be formulated as an incremental image registration task. In other terms, as the problem of estimating the incremental transformations which optimally align a reference image with successive frames of a video sequence (see Fig. 1). In this case, the reference image is also called the fixed image, and the current image can also be referred to as the moving one. Frequently, only a region of interest (also called template) within the entire reference image is to be aligned with successive frames. Image registration is a fundamental component in a variety of vision-based applications, e.g., in medical image analysis, augmented reality and vision-based robot control. Given its importance, a huge body of literature has been elaborated (Brown, 1992; Maintz and Viergever, 1998). An exhaustive description of this production is beyond the scope of this article. Therefore, let us start by making explicit the context on which this paper focuses. Then, we shall present the state-of-the-art methods and our contributions to the field.

First of all, the solutions to this problem can in general be classified into feature-based or direct methods (Irani and Anandan, 1999; Szeliski, 2005). Feature-based methods require first extracting and matching a set of geometric primitives (e.g., points, lines, contours, etc.) from the two images. The estimation problem is afterward solved. Direct methods exploit the pixel intensities without having to extract and match image features. They are also called, e.g., intensity-based, appearance-based, and texture-based. Direct methods simultaneously solve for the estimation problem and for the pixel correspondences. They can be highly accurate mainly owing to the exploitation of all image information, even from areas where no features exist. On the other hand, these methods assume that the two images have a sufficient overlapping (Lucas and Kanade, 1981). This paper considers robotic applications, such as visual servoing (Sil-



(a) (b)

Fig. 1 (a) Reference image (template) superimposed by a grid. (b) Current image superimposed by the aligned grid. Image registration consists in estimating the appropriate parameters to optimally align all pixels within a reference template to another image of the same object, taken at different imaging conditions.

veira and Malis, 2007a) and visual SLAM (Silveira et al., 2008). Hence, we can suppose that the frame rate is sufficiently high such that only relatively small interframe displacements of the object are observed. Moreover, high accuracy is often needed for these applications. Thus, we focus here on direct image registration methods.

Further, this article concentrates only on uncalibrated direct algorithms that are both robust to (at least a certain degree of) illumination changes and potentially real-time for a robotic system. Therefore, methods that rely on the Brightness Constancy Assumption (BCA) (Lucas and Kanade, 1981; Benhimane and Malis, 2007), or that perform a bundle adjustment are not considered here. Bundle adjustment techniques are not considered within those robotic applications because of its noncausal estimation.

Moreover, since in most cases only local nonlinear optimization techniques can be used in a real-time setting, we suppose that an initial estimate sufficiently close to the true solution is available. This is the case when either the images present a sufficient amount of overlapping, or a suitable prediction is available (this issue will be discussed later). However, methods based on optical flow computation (Negahdaripour, 1998; Black et al., 2000; Haussecker and Fleet, 2001) are also not considered here since they assume a too small interframe displacement of the objects.

In addition, we consider applications where off-line learning steps are not possible to be executed prior to the registration task. Hence, the techniques proposed, e.g., by Hager and Belhumeur (1998), La Cascia et al. (2000) and Nastar et al. (1996) cannot be applied. The image registration must start immediately after that the reference image is selected. This selection can be made either manually or automatically.

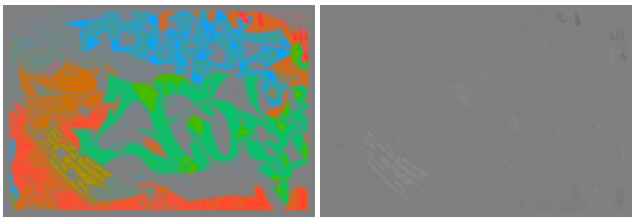
Very importantly, the solution to our problem must support all classes of image transformations, including perspective deformations. This is crucial to developing a general scheme. In particular, this enables the control of all six degrees-of-freedom of a robot. Thus, the visual tracking technique proposed, e.g., by Comaniciu et al. (2000), though effective, is not sufficient for our purposes since it provides up to a similarity transformation. Moreover, this technique only works for color images. We investigate techniques that can work with both grayscale and color images.

1.1 Related Work

Initial works within that context specially focused on registering images of planar surfaces (Shum and Szeliski, 2000), and on using the iterative Gauss-Newton minimization of their sum of squared differences. The same optimization approach can be used for the direct alignment of deformable surfaces (Bartoli and Zisserman, 2004). Contributions of the present article are both in the field of direct transformation models, and on the efficiency issue of the registration.

One important step toward real-time applications consists in improving the efficiency of the Gauss-Newton optimization method. Two approaches are possible for building efficient algorithms. The first one is to keep the same convergence rate (the number of iterations needed to obtain the minimum of the cost function), whilst reducing the computational cost per iteration. This can be achieved by precomputing partially (Hager and Belhumeur, 1998) or completely (Baker and Matthews, 2001) the Jacobian used in the minimization. The main limitation of these approaches is that they can only be applied to certain classes of warps. Another limitation concerns the visibility issue. The object of interest must be fully visible in the image, otherwise the Jacobians must be recalculated. Furthermore, in the case of surfaces in the 3D space, the convergence rate of the widely used algorithm (Baker and Matthews, 2004) is not equivalent to the convergence rate of (Lucas and Kanade, 1981). An alternative approach for building efficient algorithms is to keep the same computational cost per iteration, whilst increasing the convergence rate. This can be achieved, e.g., by using the efficient second order minimization method (Malis, 2004). This approach has been applied by Benhimane and Malis (2007) for visual tracking planar surfaces under the BCA. An approach derived from this latter is proposed by Mégret et al. (2008). Here, we propose a flexible and efficient algorithm that can be used for the alignment of rigid and deformable surfaces, whilst completely relaxing the BCA, all in a unified manner. Compared to existing techniques, a great efficiency is obtained by reducing the number of iterations needed to converge to the minimum of the cost function.

Indeed, we also tackle here an important issue to all vision-based algorithms: the robustness to generic lighting changes. We address the efficient tracking of Lambertian and non-Lambertian objects under unknown imaging conditions, also in a unified manner. To this end, a possible scheme to increase the robustness to variable illumination



(a) (b)

Fig. 2 (a) Original color image (please see it in the monitor, or print it in color, so as to verify how rich this image is) and (b) after its conversion to grayscale. Almost all information has been lost in this example, what illustrates the need to work with the color image directly.

is by performing a photometric normalization. For example, the images may be normalized using the mean and the standard deviation. However, this method provides inferior performance, especially when the interframe displacements are large (Baker and Matthews, 2004). Another widely used technique is to model the change in illumination as an affine transformation (Jin et al., 2001). Despite the fact that improved results are obtained, only global changes are modeled and thus specular reflections, for example, are not taken into consideration. A possible strategy to deal with local changes is to use a robust error function (Huber, 1981). Nevertheless, they are shown to be inefficient in the case of direct tracking (Baker and Matthews, 2004). The reasons are twofold. First, they may discard important, pertinent information that could be easily modeled and thus, exploited. Hence, the convergence rate of the algorithm tends to slow down or, even worse, the tracking may fail. Second, in this case there is an ambiguity in the interpretation of the intensity differences between those caused by motion and those caused by lighting changes (Jurie and Dhome, 2002). On the other hand, those robust functions may be applied to handle unknown occlusions since their realistic modeling is a rather difficult task.

Finally, we are interested in improving the robustness to generic illumination changes not only in grayscale images, but also in color images. Color images can be of particular importance in many scenarios. As a matter of fact, extreme cases exist where all visual information is lost when grayscale cameras are used (see Fig. 2). Even if this is an unlikely situation in practice, we can conjecture that in many cases color cameras provide much richer information than their grayscale counterparts. Hence, their application should be studied in more depth. Another motivation to work with color images is owing to the possibility of removing specularities in these images (Tan and Ikeuchi, 2005; Klinker et al., 1990). Color constancy (also referred to as chromatic adaptation) is indeed an active research topic, which seeks illuminant-invariant color descriptors. A closely related problem is to find illuminant-invariant relationships between color vectors. Given two images of a Mondrian world[1] under specific conditions,[2] Finlayson et al. (1994) claim that a multiplication of each tristimulus value (in an appropriate basis) by a scale factor is sufficient to support color constancy in practice. This framework has been exploited in color-based point tracking (Montesinos et al., 1999; Gouiffès et al., 2006) and in color image registration (Bartoli, 2008). Here, we show that such a framework corresponds to a particular case of the proposed general model.

1.2 Contributions

This article proposes a direct technique to visual tracking various classes of objects despite challenging lighting variations. To this end, we propose a new model of illumination changes and a new geometric model of image motion. The resulting photogeometric transformation model is general. On effect, the proposed model overcomes the limitations of both the Mondrian world[1] and those working conditions,[2] whilst naturally encompassing the graylevel case. Furthermore, it does not require prior knowledge of the imaging sensors (e.g., spectral response characteristics), of the light sources (e.g., number, power, pose), or of the object (e.g., albedos, shape). As for the proposed geometric model of image motion, we show here how to encompass both rigid and deformable objects whilst still preserving that robustness property. The related geometric variables are parametrized using the Lie algebra. The transformation model can be adapted such that the real-time constraint is satisfied, at an eventual expense of robustness/accuracy. Furthermore, we extend the efficient second order approximation method to simultaneously obtain the optimal global and local parameters related to all those models. Hence, large rates and domains of convergence are achieved.

This article is a revised and extended version of the approaches proposed in part in (Silveira and Malis, 2007b) and in (Malis, 2007). In particular, we show here that widely adopted models are in fact particular cases of the proposed general transformation models. Another contribution of this paper is the generalization to the case of color images. In other terms, we show that the photometric model proposed in (Silveira and Malis, 2007b) for grayscale images is also a particular case of a more general model of illumination changes. Given the parametric models, we demonstrate how a hierarchical scheme in terms of number of parameters can be devised. Another discussion only present here concerns the important aspect of surface modeling. Typically, it represents a compromise between computational complexity, robustness and accuracy. Finally, this article presents the main

[1] A Mondrian is a planar surface composed of Lambertian patches, and is after Piet Mondrian (1872-1944) whose paintings are similar.

[2] For example, the light that strikes the surface has to be of uniform intensity and spectrally unchanging, no interreflections, etc.

limitations of the proposed framework, as well as some possible solutions to overcome them.

Results are provided using various real-world sequences of images under large ambient, diffuse and specular reflections, which vary in power, type, number and space. Another complication that can arise concerns the occurrence of off-specular peaks (glints) and interreflections. Results demonstrate that the proposed approach also accommodates them without making any additional change. For the experiments, representative rigid and deformable surfaces were chosen which range from smooth to rough, including metal and dielectric objects. Existing efficient direct techniques are not able to cope with such a challenging scenario, especially when the object is not near-Lambertian and/or relatively large interframe displacements of the object are carried out. Supplemental multimedia material is provided so as to better support and demonstrate the generality, robustness and reliability of the proposed visual tracking technique.

1.3 Paper Organization

This article is organized as follows. Section 2 briefly recalls standard modeling aspects and techniques related to image registration. The proposed models are introduced in Section 3, whereas the proposed methods are presented in Section 4. Section 5 contains comparison results with the related state-of-the-art techniques. Many other experimental results are reported in Section 6, which also describes the supplemental multimedia material. Finally, Section 7 presents the main conclusions and some directions for future research.

2 Theoretical Background

2.1 Notations

Unless otherwise stated, scalars are denoted either in italics or in lowercase Greek letters, e.g., v, λ; vectors in lowercase bold fonts, e.g., $\mathbf{v}$; whereas matrices are represented in uppercase bold fonts, e.g., $\mathbf{V}$. Also, $\mathbf{0}$ (resp. $\mathbf{1}$) denotes a matrix of zeros (resp. ones) of appropriate dimensions, and $\{\mathbf{v}_i\}_{i=1}^n$ corresponds to the set $\{v_1, v_2, \ldots, v_n\}$. We follow the standard notations $\widehat{\mathbf{v}}$, $\overline{\mathbf{v}}$, $\widetilde{\mathbf{v}}$, and $\|\mathbf{v}\|$ to respectively represent an estimate, its true value, an increment, and the Euclidean norm of $\mathbf{v}$. Here, a superscripted asterisk, e.g., $\mathbf{v}^*$, is used to characterize that a variable is defined with respect to the reference frame; whereas a superscripted circle, e.g., $\mathbf{v}^\circ$, denotes its optimal value relative to a given cost function. Further, $\mathbf{v}'$ represents a transformed, modified or a normalized version of the original $\mathbf{v}$. Finally, the gradient operator applied to a vector-valued function $\mathbf{d}(\mathbf{v})$ with respect to $\mathbf{v}$ is denoted $\nabla_{\mathbf{v}}\mathbf{d}(\mathbf{v})$. This matrix of first order partial derivatives is also referred to as the Jacobian matrix $\mathbf{J}(\mathbf{v})$.

2.2 Image Formation

Consider throughout this article the pinhole camera model. According to major illumination models, both experimental (Blinn, 1977) and physically-based ones (Cook and Torrance, 1982), the intensity (i.e., irradiance) at a pixel $\mathbf{p} = [u, v, 1] \in \mathbb{P}^2$ is due to specular, diffuse and ambient reflections. These models can be concisely expressed as

$$\mathcal{I}(\mathbf{h}_m, \mathbf{p}) = \mathcal{I}_s(\mathbf{h}_s, \mathbf{p}) + \mathcal{I}_d(\mathbf{h}_d, \mathbf{p}) + \mathcal{I}_a(\mathbf{h}_a) \quad \geq 0, \quad (1)$$

where $\mathbf{h}_m = \{\mathbf{h}_s, \mathbf{h}_d, \mathbf{h}_a\}$ comprises the respective parameters, which depend on a given illumination model. For example, the Blinn-Phong model is a function of the object pose relatively to the viewing direction, the spatial distribution of the light sources and their radiance (per-wavelength), of the diffuse and specular albedos of each surface point (per-wavelength), the specular exponent and camera gain. In the case of the Cook-Torrance model, other parameters include the Fresnel reflectance and the surface roughness.

Case 2.2.1 (Lambertian surfaces) These particular surfaces do not change appearance depending on the viewing direction. The specular term in (1) is thus null: $\mathcal{I}_s(\mathbf{h}_s, \mathbf{p}) = 0,\ \forall \mathbf{p} \in \mathcal{I}$.

2.3 Two-view Epipolar Geometry

The epipolar geometry establishes the relations between corresponding image points in a pair of images. Let us consider in this section uncalibrated views of *rigid* objects, defined with respect to the current frame $\mathcal{F}$ and the reference one $\mathcal{F}^*$. In this case, the geometric relation between corresponding image points $\mathbf{p} \leftrightarrow \mathbf{p}^*$ is given by (Faugeras et al., 2001; Hartley and Zisserman, 2000)

$$\mathbf{p} \propto \mathbf{G}\,\mathbf{p}^* + \rho^*\,\mathbf{e} \quad \in \mathbb{P}^2, \quad (2)$$

where the symbol '$\propto$' indicates proportionality up to a nonzero scale factor, $\mathbf{G} \in \mathbb{R}^{3\times3}$ is a homography relative to an arbitrary plane $\mathit{\Pi}$ not going through the origin of $\mathcal{F}^*$, $\mathbf{e} \in \mathbb{R}^3$ denotes the epipole (strictly speaking, $\mathbf{e} \in \mathbb{P}^2$), and $\rho^* \in \mathbb{R}$ is the parallax (relative to $\mathit{\Pi}$) of the 3D point projected in the reference image $\mathcal{I}^*$ as $\mathbf{p}^*$. This projective parallax also encodes the inverse of the depth of this 3D point.

2.4 Purely Geometric Direct Image Registration

Let us consider in this section the particular case of a *planar* object for simplicity. In this case, a warping function can be defined from (2) by setting $\rho^* = 0$:

$$\mathbf{w}\colon \mathbb{R}^{3\times3} \times \mathbb{P}^2 \rightarrow \mathbb{P}^2 \quad (3)$$

$$(\mathbf{G}, \mathbf{p}^*) \mapsto \mathbf{p} = \mathbf{w}(\mathbf{G}, \mathbf{p}^*), \quad (4)$$

The problem of purely geometric direct image registration consists in searching for the geometric parameters that best warp the current image such that each pixel intensity $\mathcal{I}(\mathbf{p})$ is matched as closely as possible to the corresponding one in the reference image $\mathcal{I}^*(\mathbf{p}^*)$. More formally, given an estimate $\widehat{\mathbf{G}}$ (it can be the identity element) of $\mathbf{G}$, and a geometric transformation model (i.e., image warping function)

$$\mathcal{I}'_g(\widetilde{\mathbf{G}}\widehat{\mathbf{G}}, \mathbf{p}^*) = \mathcal{I}(\mathbf{p}) = \mathcal{I}\big(\mathbf{w}(\widetilde{\mathbf{G}}\widehat{\mathbf{G}}, \mathbf{p}^*)\big) \quad \geq 0, \tag{5}$$

a typical purely geometric direct image registration system seeks the incremental $\widetilde{\mathbf{G}}$ to solve nonlinear optimization problems of the type

$$\min_{\widetilde{\mathbf{G}}\in\mathbb{R}^{3\times 3}} \frac{1}{2}\sum_i \big[\, \mathcal{I}'_g(\widetilde{\mathbf{G}}\widehat{\mathbf{G}}, \mathbf{p}_i^*) - \mathcal{I}^*(\mathbf{p}_i^*) \,\big]^2, \tag{6}$$

for the case of planar objects. If (5) returns a pixel coordinate $\mathbf{p}_i$ out of the image boundaries, this pixel is discarded. Of course, the cost function can be different, but the sum of square differences in (6) is the most widely used for registering images of the same modality without aberrant measures. In the sequel, let us focus on monomodality registration. Moreover, if unknown instances of those aberrant measures (e.g., unknown occlusions) may be present in the data, a robust function (Huber, 1981) may be considered in (6).

Various solutions to the problem expressed in (6) are available in the literature (Baker and Matthews, 2004). However, the solution proposed by Benhimane and Malis (2007) has been compared favorably in the case of efficiently (in terms of both domain and rate of convergence) registering images of planar objects under the Brightness Constancy Assumption (BCA). The keys to its efficiency are owing both to the parametrization of $\mathbf{G}$ as an element of the Lie group $\mathbb{SL}(3)$ (the special linear group of (3×3) matrices having determinant one), and to the efficient second order approximation method proposed by Malis (2004).

In this article, we show first how to efficiently extend the registration to rigid and deformable surfaces. The extension naturally encompasses that planar case. Then, we propose a technique to relax the BCA in a way that the images can present arbitrary illumination variations, even in the case of color images. Finally, we show how to apply the efficient second order approximation method to recover all related parameters of the proposed models.

3 Proposed General Models

3.1 Geometric Transformation Model

Consider a 3D point $\mathbf{m}^* = [x^*, y^*, z^*]^\top \in \mathbb{R}^3$ defined relatively to the reference frame $\mathcal{F}^*$. Eventually, this 3D point is deformed to the coordinate $\boldsymbol{m}^*$, also defined with respect to $\mathcal{F}^*$. We propose to model this change of position as:

$$\boldsymbol{m}^* = \frac{1}{\kappa^*}\,\mathbf{m}^* + \boldsymbol{\eta}^* \quad \in \mathbb{R}^3, \tag{7}$$

(a) (b)

Fig. 3 (a) Reference image of a planar surface. (b) An invariant deformation of the surface. Its 3D structure has been changed (i.e., the surface is not planar anymore), but it is still possible to obtain (a) by moving the viewpoint only.

where $\kappa^* \in \mathbb{R}_+$ takes into account only invariant deformations, and $\boldsymbol{\eta}^* = [\eta_x^*, \eta_y^*, 0]^\top \in \mathbb{R}^3$ captures the remaining deformations. Invariant deformations refer to those that change the 3D structure of the object with respect to $\mathcal{F}^*$ but do not alter the reference image. See Fig. 3 for an example.

Thus, by applying the equations of motion (with respect to some projective coordinate system) on (7) and using the perspective projection, we can generalize the geometric model expressed in (2) as

$$\mathbf{p} \propto \mathbf{G}\,(\mathbf{p}^* + \boldsymbol{\delta}^*) + \rho^*\,\mathbf{e} \quad \in \mathbb{P}^2, \tag{8}$$

where $\boldsymbol{\delta}^* = [\delta_u^*, \delta_v^*, 0]^\top \in \mathbb{R}^3$ is an image coordinate deformation vector which encompasses $\boldsymbol{\eta}^*$ and κ^*. We note that the parallax $\rho^* \in \mathbb{R}$ in (8) also takes into consideration the deformation imposed by κ^*. Again, $\mathbf{e} \in \mathbb{R}^3$ denotes the epipole, and $\mathbf{G} \in \mathbb{SL}(3)$ is a homography relative to an arbitrary plane not going through the origin of $\mathcal{F}^*$.

The general relation (8) allows for defining a hierarchical unified geometric modeling. Indeed, easy transition between models is assured as follows.

Case 3.1.1 (Planar objects) The planar case represents the simplest class with respect to the number of parameters. Indeed, for this case we have

$$\boldsymbol{\delta}^* = \mathbf{0} \quad \text{and} \quad \rho^* = 0. \tag{9}$$

Case 3.1.2 (Rigid objects) The class of nonplanar rigid surfaces has a higher degree of complexity relatively to the planar case since more parameters are required to fully model them. However, once their structure parameters are correctly estimated they may be fixed for all times on:

$$\boldsymbol{\delta}^* = \mathbf{0} \quad \text{and} \quad \dot{\rho}^* = 0. \tag{10}$$

Case 3.1.3 (Objects under invariant deformation) Increasing the degree of complexity, the next class comprises the deformable surfaces such that

$$\boldsymbol{\delta}^* = \mathbf{0} \quad \text{and} \quad \dot{\rho}^* \neq 0. \tag{11}$$

In the most general case (i.e., the one with the highest degree of complexity), the class of general deformable objects has

$$\boldsymbol{\delta}^* \neq \mathbf{0} \quad \text{and} \quad \dot{\rho}^* \neq 0 \tag{12}$$

within the Desideratum (8).

Finally, we can also generalize the warping operator (3) using (8) as

$$\mathbf{w} \colon G \times \mathbb{P}^2 \rightarrow \mathbb{P}^2 \tag{13}$$

$$(\mathbf{g}, \mathbf{p}^*) \mapsto \mathbf{p} = \mathbf{w}(\mathbf{g}, \mathbf{p}^*). \tag{14}$$

where G is an appropriate group and

$$\mathbf{g} = \{\mathbf{G}, \mathbf{e}, \rho^*, \boldsymbol{\delta}^*\} \quad \in G \tag{15}$$

encodes the geometric description of the scene structure, of the camera itself, and of its motion. This allows for defining a general geometric transformation model as

$$\mathcal{I}'_g(\mathbf{g}, \mathbf{p}^*) = \mathcal{I}\big(\mathbf{w}(\mathbf{g}, \mathbf{p}^*)\big). \tag{16}$$

Another important aspect concerns the parametrization of (15). Our proposed one will be discussed in Section 4.1, along with the corresponding photometric quantities.

3.2 Photometric Transformation Model

3.2.1 The Case of Grayscale Images

For image registration purposes, the photometric modeling aims at explaining the lighting changes between views. In other terms, it concerns the recovery of which lighting variations have to be applied to the current image $\mathcal{I}$ (1) such that the photometrically transformed one $\mathcal{I}'_h$ reproduces as closely as possible the illumination conditions at the time of acquiring the reference image $\mathcal{I}^*$.

A possible photometric transformation model to act on $\mathcal{I}$ (1) can be defined as

$$\mathcal{I}'_h(\boldsymbol{\alpha}_s, \boldsymbol{\alpha}_d, \beta, \mathbf{p}) = \boldsymbol{\alpha}_s(\mathbf{p})\, \mathcal{I}_s(\mathbf{p}) + \boldsymbol{\alpha}_d(\mathbf{p})\, \mathcal{I}_d(\mathbf{p}) + \beta \ \geq 0, \tag{17}$$

where $\boldsymbol{\alpha}_s(\mathbf{p}), \boldsymbol{\alpha}_d(\mathbf{p}) \in \mathbb{R}$ and $\beta \in \mathbb{R}$ aim to counterbalance the variations caused by specular, diffuse and global lighting changes, respectively. The latter also includes the shift in the camera bias. Notice that the first two variables depend on the albedos of each point on the surface, as well as its shape, the camera parameters and other imaging conditions. These variables can be seen as a function of the changes in $\mathbf{h}_m = \{\mathbf{h}_s, \mathbf{h}_d, \mathbf{h}_a\}$ between views. In this way, it represents a difficult, computationally intensive problem where many images and priors are required to consistently recover those parameters. Indeed, two assumptions are widely adopted by photogeometric direct image registration algorithms (Jin et al., 2001; Baker and Matthews, 2004). The first assumption is to consider that the surface is Lambertian (see the particular Case 2.2.1) so that $\boldsymbol{\alpha}_s(\mathbf{p}) = 0, \forall \mathbf{p} \in \mathcal{I}$. Additionally, they assume that the entire surface holds the same reflectance properties so that, $\forall \mathbf{p} \in \mathcal{I},\ \boldsymbol{\alpha}_d(\mathbf{p}) = \alpha$ is a constant. Although suited to some applications, both assumptions are obviously violated in many cases.

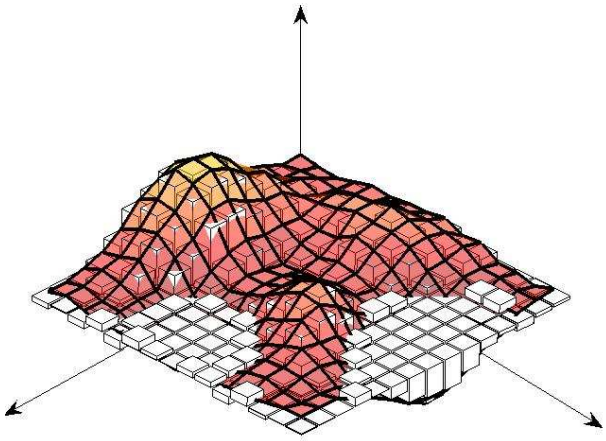


Fig. 4 (Color online) The illumination changes are viewed as an evolving three-dimensional surface $\mathcal{S}$. Thus, local lighting variations are also captured by this model.

In this paper, we develop a general model of illumination changes. Instead of using (17), we seek an elementwise multiplicative lighting variation $\mathcal{S}$ over the current $\mathcal{I}$, and a global $\beta \in \mathbb{R}$, such that the resulting $\mathcal{I}'_h$ matches as closely as possible to the reference $\mathcal{I}^*$. Indeed, we propose the following general (in the case of grayscale images) photometric transformation model:

$$\mathcal{I}'_h(\mathcal{S}, \beta, \mathcal{I}) = \mathcal{S} \cdot \mathcal{I} + \beta, \tag{18}$$

where the dot operator '$\cdot$' denotes the elementwise multiplication. Hence, the lighting variation $\mathcal{S}$ is viewed as a surface that evolves with time. Notice that, while the offset β captures global variations only, the surface $\mathcal{S}$ also models local illumination changes (e.g., produced by specular reflections). See Fig. 4 for an illustration. Very importantly, this model allows the registration to be performed without prior knowledge of the object (e.g., albedos, shape) or of the light sources (e.g., number, power, pose) or of the camera.

The proposed model (18) is also different from the one presented by Negahdaripour (1998), where the offset is also as a function of the pixels. This existing model is overparametrized, but is shown in that work to give satisfactory results in the case of computing optical flow. This computation is not our primary objective, though registration methods also recover that flow simultaneously. A strategy to reduce the problems related to that overparametrized model (e.g., convergence issues) is given by Lai and Fang (1999).

Case 3.2.1 (Affine model) It is easy to verify that the affine case corresponds to a particular model of the general one (18). In this case, the surface is described by:

$$\mathcal{S} = \gamma \mathbf{1}, \tag{19}$$

with $\gamma \in \mathbb{R}$. This model is appropriate if that previously mentioned prior knowledge of the imaging conditions and of the object is available.

In the general case, if the alignment involves only two images and robustness to generic illumination changes is sought, an underconstrained system is obtained (more unknowns than equations). Surface reconstruction algorithms classically solve underconstrained problems through a regularization of the surface. The basic idea is to prevent pixel intensities from changing independently of each other. Given that the model the illumination changes is viewed as an evolving surface, the same technique can be applied to the registration at hand. Indeed, the surface $\mathcal{S}$ is supposed to be described by a parametric function

$$\mathcal{S} \approx f_h(\boldsymbol{\gamma}, \mathbf{p}), \quad \forall \mathbf{p} \in \mathcal{I}, \tag{20}$$

where the real-valued vector $\boldsymbol{\gamma}$ contains less parameters than the available equations. Then, one has to choose an appropriate low-dimensional approximation $f_h(\boldsymbol{\gamma}, \mathbf{p})$ of the actual surface. This will be discussed in Section 3.3.

Highlights and shadows These particular effects can be interpreted as well-structured types of occluders. The characterization as an occluder is well-justified in the case where *all* information which are useful for registration purposes is hidden. In this case, they are also well-structured because a saturation pattern is exhibited either to zero or to the highest intensity level. Therefore, they can be filtered suitably: one only needs to check whether or not those homogeneous patterns appear in each warped image region.

3.2.2 Generalization to Color Images

It is shown here how to extend the photometric model presented in Section 3.2.1 to the case of color images. On effect, this new photometric transformation model overcomes the limitations of both the Mondrian world and various working conditions (see Section 1.1), whilst naturally encompassing the graylevel case. Furthermore, the extension will be made for any color image, i.e., other multispectral images such as those that include the infrared band.

Let $\boldsymbol{\mathcal{I}}$ represent a color image, which is obtained by stacking the channels $\mathcal{I}_k$, $k = 1, 2, \ldots, n$. The main idea consists in respecting all intrinsic couplings that may be present between channels so as to be as general as possible. Indeed, we propose to obtain a photometrically transformed n-channel color image $\boldsymbol{\mathcal{I}}'_h$ that best matches the reference one $\boldsymbol{\mathcal{I}}^*$ through the model

$$\begin{bmatrix} \mathcal{I}'_{h1}(\mathbf{h}, \boldsymbol{\mathcal{I}}) \\ \mathcal{I}'_{h2}(\mathbf{h}, \boldsymbol{\mathcal{I}}) \\ \vdots \\ \mathcal{I}'_{hn}(\mathbf{h}, \boldsymbol{\mathcal{I}}) \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} \mathcal{S}_{1j} \cdot \mathcal{I}_j + \beta_1 \\ \sum_{j=1}^{n} \mathcal{S}_{2j} \cdot \mathcal{I}_j + \beta_2 \\ \vdots \\ \sum_{j=1}^{n} \mathcal{S}_{nj} \cdot \mathcal{I}_j + \beta_n \end{bmatrix}, \tag{21}$$

where the full set of photometric variables

$$\mathbf{h} = \{\boldsymbol{\mathcal{S}}, \boldsymbol{\beta}\} \quad \in \mathbb{R}^{pn^2+n}, \tag{22}$$

where p is the number of image pixels, comprises the surfaces related to the illumination changes

$$\boldsymbol{\mathcal{S}} = \begin{bmatrix} \mathcal{S}_{11} & \mathcal{S}_{12} & \cdots & \mathcal{S}_{1n} \\ \mathcal{S}_{21} & \mathcal{S}_{22} & \cdots & \mathcal{S}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{S}_{n1} & \mathcal{S}_{n2} & \cdots & \mathcal{S}_{nn} \end{bmatrix}, \tag{23}$$

and the per-channel shift in the ambient lighting changes and camera bias, which is captured by the real-valued variable

$$\boldsymbol{\beta} = \left[\beta_1 \mathbf{1}, \beta_2 \mathbf{1}, \ldots, \beta_n \mathbf{1}\right]^\top. \tag{24}$$

In the sequel, let the Desideratum (21) be concisely written

$$\boldsymbol{\mathcal{I}}'_h(\mathbf{h}, \boldsymbol{\mathcal{I}}) \;=\; \boldsymbol{\mathcal{S}} \bullet \boldsymbol{\mathcal{I}} + \boldsymbol{\beta}, \tag{25}$$

where the operator '$\bullet$' represents the linear combination of the color channels, elementwise multiplied by the corresponding surface.

The proposed fully coupling photometric model (25) allows the registration to be performed without prior knowledge of the characteristics (including the spectral ones) of the light sources, of the object (which can be non-Lambertian), and of the camera sensors. Nonetheless, these priors can be easily applied to that general model if they are available. For example, prior knowledge of the spectral response of the camera sensors (e.g., from its datasheet) allows for suitably uncoupling the lighting variation $\boldsymbol{\mathcal{S}}$. This particular case is described below.

Case 3.2.2 (Known spectral characteristics) If the color camera's datasheet specifies that the n sensors are narrow-band, then a fully uncoupled model can be used by adopting

$$\boldsymbol{\mathcal{S}} = \operatorname{diag}(\mathcal{S}_{11}, \mathcal{S}_{22}, \ldots, \mathcal{S}_{nn}). \tag{26}$$

If only some of them are narrow-band, it is possible to devise other particular models from the general one (25) so as to suitably uncouple the corresponding channels. For example, given that at least the Red and the Blue channels are only weakly coupled in many RGB cameras, one may set $\mathcal{S}_{13} = \mathcal{S}_{31} = \mathbf{0}$ in (23). In addition, if a symmetry between a particular coupling is present, then a reduction on

the number of surfaces to be estimated can also be achieved by setting $\mathcal{S}_{12} = \mathcal{S}_{21} = \mathcal{S}_2$ and/or $\mathcal{S}_{23} = \mathcal{S}_{32} = \mathcal{S}_3$, i.e.,

$$\boldsymbol{\mathcal{S}} = \begin{bmatrix} \mathcal{S}_{11} & \mathcal{S}_2 & \mathbf{0} \\ \mathcal{S}_2 & \mathcal{S}_{22} & \mathcal{S}_3 \\ \mathbf{0} & \mathcal{S}_3 & \mathcal{S}_{33} \end{bmatrix}. \tag{27}$$

Case 3.2.3 (Affine model) Similarly to the grayscale case (see Case 3.2.1), it is easy to verify that affine models for color images also correspond to particular cases of the proposed general photometric transformation model (25). A first possibility (Finlayson et al., 1994) consists in changing the current and reference images to an appropriate basis $\mathbf{B} \in \mathbb{R}^{3\times 3}$ (i.e., to a suitable color space) and then to solve for a real diagonal matrix $\mathbf{D}$. This possibility corresponds to the affine model

$$\boldsymbol{\mathcal{I}}'_h = \left[\left(\mathbf{B}^{-1}\,\mathbf{D}\,\mathbf{B}\right) \otimes \mathbf{1} \right] \bullet \boldsymbol{\mathcal{I}} + \boldsymbol{\beta}, \tag{28}$$

where the symbol '$\otimes$' denotes the Kronecker product. If it is too difficult to estimate or choose the basis $\mathbf{B}$, another option is to directly estimate the matrix

$$\mathbf{B}^{-1}\,\mathbf{D}\,\mathbf{B} = \boldsymbol{\Lambda} \quad \in \mathbb{R}^{3\times 3}. \tag{29}$$

This corresponds to a particular case of (25) where

$$\boldsymbol{\mathcal{S}} = \boldsymbol{\Lambda} \otimes \mathbf{1}. \tag{30}$$

In the general case, if the alignment involves only two images and robustness to generic illumination changes is sought, an underconstrained system is still obtained even if n-channel images are considered. Thus, following the same technique for the graylevel case, we suppose that $\boldsymbol{\mathcal{S}}$ can be described by parametric functions

$$\boldsymbol{\mathcal{S}} \approx \boldsymbol{f}_h(\boldsymbol{\Gamma}, \mathbf{p}), \quad \forall \mathbf{p} \in \boldsymbol{\mathcal{I}}, \tag{31}$$

where $\boldsymbol{\Gamma} = \{\boldsymbol{\gamma}_{kj}\}$, $k, j = 1, 2, \ldots, n$. One then has to choose an appropriate low-dimensional approximation $\boldsymbol{f}_h(\boldsymbol{\Gamma}, \mathbf{p})$ of the actual $\boldsymbol{\mathcal{S}}$. This will be discussed in the next section. An efficient optimization procedure to estimate all those parameters is devised in Section 4.2.

Highlights and shadows Similarly to the grayscale case, saturations due to highlights and shadows are also interpreted as well-structured types of occluders. In the case of color images each channel is independently filtered.

3.3 Surfaces Modeling

The modeling of surfaces is an important design step within estimation methods from visual data. Besides the scene structure, illumination changes are also modeled here as a surface. Additionally, we showed that regularization techniques are needed in both cases so as to avoid constructing an underconstrained system. We remark that the total characterization of the surfaces to be estimated depends both on the complexity of the data and on the task-specific requirements. To this end, besides the number of surfaces, design parameters also include both the function itself and the number of samples to define each surface. They typically represent a compromise between computational complexity, robustness and accuracy.

Let us first discuss the number of surfaces. Consider an n-channel image, $n \geq 1$. Of course, the case where $n = 1$ corresponds to a graylevel image. In the simplest case of a planar object and fully decoupled surfaces for the illumination changes, we have a total of n surfaces to be estimated. On the other hand, in the most general case of a general deformable object along with a fully coupled model of lighting variations, a total of $n^2 + 3$ surfaces are required to accurately explain the image motion. They represent:

- the surface related to the projective parallax

$$\rho^* = f_\rho(\boldsymbol{\lambda}_\rho, \mathbf{p}); \tag{32}$$

- the surface related to the general deformation in the u-direction

$$\delta_u^* = f_\delta(\boldsymbol{\lambda}_u, \mathbf{p}); \tag{33}$$

- the surface related to the general deformation in the v-direction

$$\delta_v^* = f_\delta(\boldsymbol{\lambda}_v, \mathbf{p}); \tag{34}$$

- and finally the surface(s) related to the illumination changes

$$\mathcal{S}_{kj}(\mathbf{p}) = f_h(\boldsymbol{\gamma}_{kj}, \mathbf{p}), \quad k, j = 1, 2, \ldots, n. \tag{35}$$

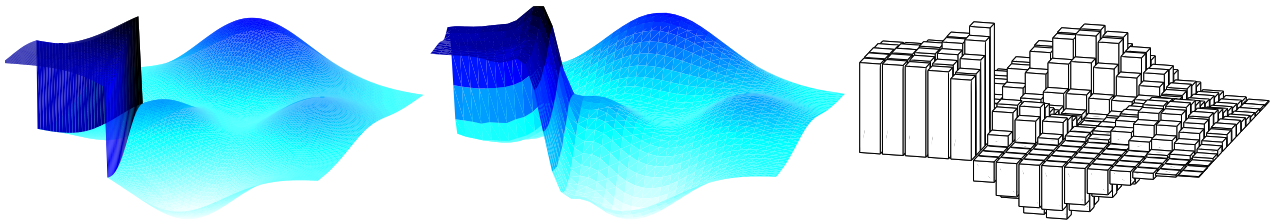
To approximate those surfaces, an appropriate choice of each function

$$f_{(\cdot)} \colon \mathbb{R}^{q_{(\cdot)}} \times \mathbb{P}^2 \to \mathbb{R}, \tag{36}$$

where $q_{(\cdot)}$ denotes its number of parameters, has to be made. This choice depends on several factors, as discussed next. Of course, different choices can be made for each one of the surfaces. We present below two possible functions:

1. a widely used technique to regularize a surface is via Radial Basis Functions (RBF) (Carr et al., 1997). In this case, the function in (36) may be defined, for example, as a thin plate spline $\varphi(x) = x^2 \log(x)$, $\forall x \in \mathbb{R}_+$, along with a first-degree polynomial, i.e.

$$f(\boldsymbol{\gamma}, \mathbf{p}) = [\gamma_{s+1},\, \gamma_{s+2},\, \gamma_{s+3}]^\top \mathbf{p} + \sum_{i=1}^{s} \gamma_i\, \varphi(\|\mathbf{p} - \mathbf{q}_i\|), \tag{37}$$



(a) Original surface (b) Approx. by a RBF (c) Approx. by discretizing

Fig. 5 Some possibilities to approximate a surface. (b) Radial Basis Functions (RBF) regularize it but do not capture discontinuities. (c) Discretization deals with discontinuities and yields a computationally efficient system, but ignores smoothness.

where $\{\mathbf{q}_i \in \mathbb{P}^2\}_{i=1}^{s}$ are image points (called centers) that can be selected, for example, on a regularly spaced grid or be interest points. The side conditions can be easily imposed by solving a linear system, whilst the interpolation conditions are indirectly imposed by minimizing a similarity measure (e.g., the sum of square differences). The use of RBFs allows for regularizing the surface, but they may fail to accurately capture discontinuities since the function (37) has a global support.

2. A possible strategy for dealing with discontinuous surfaces is to approximate it via a discretization into s sufficiently small $(\Delta u \times \Delta v)$ regions with

$$f(\boldsymbol{\gamma}, \mathbf{p}) \approx \begin{cases} \gamma_i, & \forall \mathbf{p} \in \Delta u_i \, \Delta v_i, \\ 0, & \text{otherwise}, \end{cases} \tag{38}$$

such that

$$\iint_{\mathcal{I}} \mathcal{S}(\mathbf{p}) \, du \, dv \approx \sum_{i=1}^{s} f(\boldsymbol{\gamma}_i, \mathbf{p}) \, \Delta u_i \, \Delta v_i. \tag{39}$$

This discretization leads to a computationally efficient solution since sparse Jacobians are obtained. On the other hand, this approximation ignores eventual surface smoothness.

Hence, the appropriateness of a particular approximation depends on various factors, such as the assumptions concerning the surface smoothness and on the required system's performance. See Fig. 5 for illustrative examples. Other possible approximations include the use of bivariate polynomials, and a combination of discretization followed by a suitable (e.g., cubic) interpolation.

Additionally, the number of parameters to define each surface, i.e., $q_\rho = \dim(\boldsymbol{\lambda}_\rho)$, $q_u = \dim(\boldsymbol{\lambda}_u)$, $q_v = \dim(\boldsymbol{\lambda}_v)$ and $q_\gamma = \dim(\boldsymbol{\gamma}_{kj})$, obviously has an impact on the system's performance as well. Nevertheless, a hierarchical approach can be applied to find a suitable number, starting from a planar surface to higher dimensional approximations.

4 Proposed Efficient Methods

4.1 The Full System

The full system is composed of the proposed transformation model, along with its parametrization, and of the nonlinear optimization method.

As for the modeling, a general photogeometric transformation model can be defined from the general model of illumination changes (25), along with the general warping model (14). More formally, the action of the proposed general transformation model on pixels is given by

$$\begin{aligned} \boldsymbol{\mathcal{I}}'_{gh}(\mathbf{x}, \mathbf{p}^*) &= \boldsymbol{\mathcal{I}}'_h\big(\mathbf{h}, \boldsymbol{\mathcal{I}}(\mathbf{p})\big) && (40) \\ &= \boldsymbol{\mathcal{I}}'_h\big(\mathbf{h}, \boldsymbol{\mathcal{I}}\big(\mathbf{w}(\mathbf{g}, \mathbf{p}^*)\big)\big) && (41) \\ &= \boldsymbol{\mathcal{S}}(\boldsymbol{\Gamma}, \mathbf{p}^*) \bullet \boldsymbol{\mathcal{I}}\big(\mathbf{w}(\mathbf{g}, \mathbf{p}^*)\big) + \boldsymbol{\beta} \quad \geq \mathbf{0}, && (42) \end{aligned}$$

where $\mathbf{x} = \{\mathbf{g}, \mathbf{h}\}$ comprises the geometric and photometric variables, respectively, $\mathbf{g} = \{\mathbf{G}, \mathbf{e}, \rho^*, \boldsymbol{\delta}^*\}$ (15) and $\mathbf{h} = \{\boldsymbol{\mathcal{S}}, \boldsymbol{\beta}\}$ (22), and the operator '$\bullet$' stands for a linear combination of the n channels of $\boldsymbol{\mathcal{I}}$, $n \geq 1$, elementwise multiplied by the corresponding surface.

Let us now discuss the important issue of parametrizing those geometric and photometric quantities. In other terms, we need to define the most appropriate set of parameters $\mathbf{z} = \{\mathbf{z}_g, \mathbf{z}_h\}$ to describe the variables $\mathbf{x} = \{\mathbf{g}, \mathbf{h}\}$:

$$\mathbf{x} = \mathbf{x}(\mathbf{z}) = \big\{\mathbf{g}(\mathbf{z}_g), \mathbf{h}(\mathbf{z}_h)\big\} \quad \in G \times \mathbb{R}^{pn^2+n}, \tag{43}$$

where p is the number of pixels considered for processing. Whereas the parametrization of the photometric quantities $\mathbf{h} = \mathbf{h}(\mathbf{z}_h)$ imposes no difficulties with

$$\mathbf{z}_h = \{\boldsymbol{\Gamma}, \boldsymbol{\beta}\} = \{\boldsymbol{\gamma}_{kj}, \boldsymbol{\beta}\} \quad \in \mathbb{R}^{q_\gamma+n}, \tag{44}$$

the adequate characterization of the geometric ones $\mathbf{g} = \mathbf{g}(\mathbf{z}_g)$ is a little more involved. Consider the (4×4) matrix

$$\mathbf{Q} = \begin{bmatrix} \mathbf{G} & \mathbf{e} \\ \mathbf{0} & 1 \end{bmatrix} \quad \in \mathbb{SA}(3). \tag{45}$$

The Lie group $\mathbb{SA}(3)$, i.e., the special affine group, is homeomorphic to $\mathbb{SL}(3) \times \mathbb{R}^3$. The Lie group $\mathbb{SE}(3) = \mathbb{SO}(3) \times$

$\mathbb{R}^3$, i.e., the special Euclidean group, is in fact a subspace of $\mathbb{SA}(3)$. The natural local parametrization of $\mathbf{Q} \in \mathbb{SA}(3)$ is through the related Lie algebra $\mathfrak{sa}(3)$, whose coordinates[3] are here denoted by $\boldsymbol{\upsilon} = [\upsilon_1, \upsilon_2, \ldots, \upsilon_{11}]^\top \in \mathbb{R}^{8+3}$, i.e., $\mathbf{Q} = \mathbf{Q}(\boldsymbol{\upsilon}) \in \mathbb{SA}(3)$. The mechanism for passing information from the Lie algebra to the related Lie group is the exponential mapping

$$\exp\colon \mathfrak{sa}(3) \to \mathbb{SA}(3) \tag{46}$$

$$\mathbf{A}(\boldsymbol{\upsilon}) \mapsto \exp\big(\mathbf{A}(\boldsymbol{\upsilon})\big) = \mathbf{Q}(\boldsymbol{\upsilon}), \tag{47}$$

where $\mathbf{A}(\boldsymbol{\upsilon})$ can be written as a linear combination of the canonical basis $\mathbf{A}_i, i = 1, 2, \ldots, 11$, of the Lie algebra $\mathfrak{sa}(3)$ (Warner, 1987; Varadarajan, 1974):

$$\mathbf{A}(\boldsymbol{\upsilon}) = \sum_{i=1}^{11} \upsilon_i \mathbf{A}_i \quad \in \mathfrak{sa}(3). \tag{48}$$

The exponential mapping (46) is smooth and one-to-one onto, with a smooth inverse, within a very large neighborhood around the origin of $\mathfrak{sa}(3)$ and the identity element of $\mathbb{SA}(3)$. This parametrization is then highly suitable to express incremental displacements. The set of geometric quantities $\mathbf{g} = \mathbf{g}(\mathbf{z}_g)$ can hence be fully parametrized by

$$\mathbf{z}_g = \{\boldsymbol{\upsilon}, \boldsymbol{\lambda}_\rho, \boldsymbol{\lambda}_u, \boldsymbol{\lambda}_v\} \quad \in \mathbb{R}^{11+q_\rho+q_u+q_v}. \tag{49}$$

In order to estimate all those parameters, an appropriate nonlinear optimization procedure is needed. For real-time systems, only local ones can generally be applied. An initial estimate $\widehat{\mathbf{x}}$ sufficiently close to the true solution is then required. This estimate is integrated into the proposed model (42) as

$$\boldsymbol{\mathcal{I}}'_{gh}\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}, \mathbf{p}^*\big) = \boldsymbol{\mathcal{S}}\big(\widetilde{\boldsymbol{\Gamma}} \circ \widehat{\boldsymbol{\Gamma}}, \mathbf{p}^*\big) \bullet \boldsymbol{\mathcal{I}}\big(\mathbf{w}(\mathbf{g}(\widetilde{\mathbf{z}}_g) \circ \widehat{\mathbf{g}}, \mathbf{p}^*)\big) + \widetilde{\boldsymbol{\beta}} \circ \widehat{\boldsymbol{\beta}} \quad \geq \mathbf{0}, \tag{50}$$

where $\widetilde{\mathbf{x}} = \mathbf{x}(\widetilde{\mathbf{z}})$ represents incremental values, and the composition operator '$\circ$' depends on the involved Lie group. For example, if a matrix Lie group is involved then the product operation to be performed is the matrix multiplication. If real-valued (resp. nonzero) vectors are considered, then the respective product operation may be defined, for example, as the (resp. elementwise multiplication) addition (Warner, 1987; Varadarajan, 1974).

Instead of using a plane-based warping model $\mathcal{I}'_g$ (5) in (6), a general direct image registration system can be devised by applying the general photogeometric transformation model $\boldsymbol{\mathcal{I}}'_{gh}$ (50) in (6). In this way, a general system can be cast as the following nonlinear optimization problem:

$$\min_{\widetilde{\mathbf{z}}=\{\widetilde{\mathbf{z}}_g, \widetilde{\mathbf{z}}_h\}} \frac{1}{2} \sum_i \Big[\underbrace{\boldsymbol{\mathcal{I}}'_{gh}\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}, \mathbf{p}_i^*\big) - \boldsymbol{\mathcal{I}}^*(\mathbf{p}_i^*)}_{d_i(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}})} \Big]^2, \tag{51}$$

which seeks to minimize the set of intensity differences $\mathbf{d} = \big\{d_i\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}\big)\big\}$. Differently from (Baker and Matthews, 2004; Bartoli, 2008), no transformation is applied on the reference image $\boldsymbol{\mathcal{I}}^*$. This allows for encompassing various classes of image warps, and for modeling both local and global illumination changes. Further, via a suitable adaptation, this also allows for simultaneously estimating the 3D camera pose and the scene structure (Silveira et al., 2008). Another benefit is that the object of interest does not have to be fully visible in the images. Finally, larger domain and rate of convergence for the optimization are obtained in this way. We find that these reasons largely overcompensate the marginal increase in the computational cost of calculating the involved Jacobians at each iteration. Indeed, we describe below a computationally efficient procedure to solve that optimization problem (51) with nice convergence properties.

4.2 The Optimization Procedure

Given the real-time requirements of robotic applications, only minimization methods that have limited convergence domain can be applied. Global methods such as Simulated Annealing (Horst and Pardalos, 1995) are too time consuming to be considered in a real-time setting. In this section, we propose an algorithm that is both computationally efficient and has a relatively large domain of convergence.

Suppose that an estimate $\widehat{\mathbf{x}}$ sufficiently close to the true parameters $\overline{\mathbf{x}}$ is available (this initialization issue will be discussed later). Further, consider that the underlying functions are (at least piecewise) smooth so that they can be expanded in Taylor series. The nonlinear optimization problem (51) can be concisely rewritten as

$$\min_{\widetilde{\mathbf{z}}} \frac{1}{2} \big\|\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}\big)\big\|^2, \tag{52}$$

where the objective consists in finding the optimal $\mathbf{x}(\widetilde{\mathbf{z}}^\circ)$ such that its composition with the estimate $\widehat{\mathbf{x}}$ yields the true values $\overline{\mathbf{x}}$, i.e.

$$\overline{\mathbf{x}} = \mathbf{x}(\widetilde{\mathbf{z}}^\circ) \circ \widehat{\mathbf{x}}. \tag{53}$$

In this case, the image alignment is perfectly achieved: $\boldsymbol{\mathcal{I}}'_{gh}(\overline{\mathbf{x}}, \mathbf{p}^*) = \boldsymbol{\mathcal{I}}^*(\mathbf{p}^*), \forall \mathbf{p}^*$. A standard technique to solve this problem consists in first performing an expansion of the function in Taylor series and applying a necessary condition for optimality. From an initial estimate $\widehat{\mathbf{x}}_0$, the solution is obtained by finding an incremental displacement $\widetilde{\mathbf{x}} = \mathbf{x}(\widetilde{\mathbf{z}}_k)$ and updating it iteratively:

$$\widehat{\mathbf{x}}_{k+1} = \mathbf{x}(\widetilde{\mathbf{z}}_k) \circ \widehat{\mathbf{x}}_k \tag{54}$$

such that $\lim_{k\to\infty} \widehat{\mathbf{x}}_k = \overline{\mathbf{x}}$, where k indexes the iterations. In practice, the convergence to the optimal solution can be

[3] By definition, the Lie algebra $\mathfrak{sa}(3)$ is of dimension 11, since the (3×3) matrix $\mathbf{G}$ is an element of the Lie group $\mathbb{SL}(3)$. Given that elements of this group have determinant one, a degree-of-freedom is already constrained.

established when $\mathbf{x}(\widetilde{\mathbf{z}}_k)$ is arbitrarily close to the identity element of the involved group, i.e., when $\|\widetilde{\mathbf{z}}_k\| < \epsilon$.

With respect to the Taylor expansion, a key technique to achieve nice convergence properties is to perform an efficient second order approximation of $\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}})\circ\widehat{\mathbf{x}}\big)$. Indeed, its second order approximation in Taylor series about the current estimate $\widehat{\mathbf{x}}$ (i.e., about $\widetilde{\mathbf{z}}=\mathbf{0}$) is

$$\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}})\circ\widehat{\mathbf{x}}\big) = \mathbf{d}(\widehat{\mathbf{x}}) + \nabla_{\widetilde{\mathbf{z}}}\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}})\circ\widehat{\mathbf{x}}\big)\Big|_{\widetilde{\mathbf{z}}=\mathbf{0}}\widetilde{\mathbf{z}} + \frac{1}{2}\nabla_{\widetilde{\mathbf{z}}}\Big(\nabla_{\widetilde{\mathbf{z}}}\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}})\circ\widehat{\mathbf{x}}\big)\Big|_{\widetilde{\mathbf{z}}=\mathbf{0}}\widetilde{\mathbf{z}}\Big)\widetilde{\mathbf{z}} + o\big(\|\widetilde{\mathbf{z}}\|^3\big), \quad (55)$$

or more compactly,

$$\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}})\circ\widehat{\mathbf{x}}\big) = \mathbf{d}(\widehat{\mathbf{x}}) + \mathbf{J}(\widehat{\mathbf{x}})\,\widetilde{\mathbf{z}} + \frac{1}{2}\mathbf{S}(\widehat{\mathbf{x}},\widetilde{\mathbf{z}})\,\widetilde{\mathbf{z}} + o\big(\|\widetilde{\mathbf{z}}\|^3\big), \quad (56)$$

where the rectangular matrix $\mathbf{S}(\widehat{\mathbf{x}},\widetilde{\mathbf{z}})$ also encompasses the square Hessian matrices, and $o\big(\|\widetilde{\mathbf{z}}\|^3\big)$ is the third order Lagrange remainder. In turn, the first order Taylor expansion of $\mathbf{J}(\mathbf{x}(\widetilde{\mathbf{z}})\circ\widehat{\mathbf{x}})$ again about the current estimate $\widehat{\mathbf{x}}$ (i.e., about $\widetilde{\mathbf{z}}=\mathbf{0}$) is given by

$$\mathbf{J}\big(\mathbf{x}(\widetilde{\mathbf{z}})\circ\widehat{\mathbf{x}}\big) = \mathbf{J}(\widehat{\mathbf{x}}) + \mathbf{S}(\widehat{\mathbf{x}},\widetilde{\mathbf{z}}) + o\big(\|\widetilde{\mathbf{z}}\|^2\big), \quad (57)$$

with the second order remainder $o\big(\|\widetilde{\mathbf{z}}\|^2\big)$. By injecting $\mathbf{S}(\widehat{\mathbf{x}},\widetilde{\mathbf{z}})$ from (57) in (56) and neglecting the third order terms, an efficient second order approximation (i.e., using only first order derivatives) of $\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}})\circ\widehat{\mathbf{x}}\big)$ is obtained:

$$\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}})\circ\widehat{\mathbf{x}}\big) = \mathbf{d}(\widehat{\mathbf{x}}) + \frac{1}{2}\Big(\mathbf{J}(\widehat{\mathbf{x}}) + \mathbf{J}\big(\mathbf{x}(\widetilde{\mathbf{z}})\circ\widehat{\mathbf{x}}\big)\Big)\,\widetilde{\mathbf{z}}. \quad (58)$$

We can then apply a necessary condition for optimality. A necessary condition for $\widetilde{\mathbf{z}}=\widetilde{\mathbf{z}}^\circ$ to be a stationary point of our cost function in (52) is

$$\mathbf{0} = \nabla_{\widetilde{\mathbf{z}}}\Big(\frac{1}{2}\,\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}})\circ\widehat{\mathbf{x}}\big)^\top\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}})\circ\widehat{\mathbf{x}}\big)\Big)\Big|_{\widetilde{\mathbf{z}}=\widetilde{\mathbf{z}}^\circ} \quad (59)$$

$$= \nabla_{\widetilde{\mathbf{z}}}\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}})\circ\widehat{\mathbf{x}}\big)\Big|_{\widetilde{\mathbf{z}}=\widetilde{\mathbf{z}}^\circ}^{\top}\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}}^\circ)\circ\widehat{\mathbf{x}}\big), \quad (60)$$

or more compactly,

$$\mathbf{J}(\overline{\mathbf{x}})^\top\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}}^\circ)\circ\widehat{\mathbf{x}}\big) = \mathbf{0}, \quad (61)$$

using (53). Provided that $\mathbf{J}(\overline{\mathbf{x}})$ is full rank, we have

$$\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}}^\circ)\circ\widehat{\mathbf{x}}\big) = \mathbf{0}. \quad (62)$$

The roots of this system of nonlinear equations (62) is generally difficult to obtain in closed form. However, using the Taylor approximation (58) about $\widetilde{\mathbf{z}}=\widetilde{\mathbf{z}}^\circ$ along with (53) yield the following system of equations

$$\frac{1}{2}\big(\mathbf{J}(\widehat{\mathbf{x}}) + \mathbf{J}(\overline{\mathbf{x}})\big)\,\widetilde{\mathbf{z}}^\circ = -\mathbf{d}(\widehat{\mathbf{x}}), \quad (63)$$

where $\mathbf{d}(\widehat{\mathbf{x}})$ and the Jacobian $\mathbf{J}(\widehat{\mathbf{x}})$ are completely computed using current information. On the other hand, the entire Jacobian $\mathbf{J}(\overline{\mathbf{x}})$ at the reference (true) values cannot because some of them are unknowns. Only a part of the latter can always be computed (by applying the chain rule), since the reference image is given. The remaining part must be approximated using, for example, the current estimate so that (63) can be a rectangular linear system. Let $\widehat{\mathbf{J}} = \mathbf{J}(\overline{\mathbf{x}})$ represent this approximated Jacobian at the reference values. Nevertheless, in some particular cases where the warping function (14) is a group action on $\mathbb{P}^2$ (e.g., in the planar case[4]), a rectangular linear system is obtained from (63) without any approximation. Independently (either approximately or exactly) of how a rectangular linear system is obtained from (63), its solution is found in the least-squares sense via

$$\widetilde{\mathbf{z}}^\circ = -2\,\big(\mathbf{J}(\widehat{\mathbf{x}}) + \widehat{\mathbf{J}}\,\big)^+\,\mathbf{d}(\widehat{\mathbf{x}}), \quad (64)$$

where $(\cdot)^+$ denotes the pseudoinverse of a matrix and $2\,(\mathbf{J}(\widehat{\mathbf{x}}) + \widehat{\mathbf{J}}\,)^+\mathbf{d}(\widehat{\mathbf{x}})$ represents our proposed descent direction. The Gauss-Newton method does not consider the important contribution of $\widehat{\mathbf{J}}$, which includes the gradient of the reference image. The analytical expressions of these Jacobians can be found in (Malis, 2007; Silveira and Malis, 2007b). It can be noted that the obtained $\widetilde{\mathbf{z}}^\circ$ may not align the images using (53) at the first iteration, especially because a Taylor approximation of the true nonlinear equations (62) is performed. Thus, the solution $\widetilde{\mathbf{z}}^\circ$ from (64) represents an incremental displacement that must be iterated via (54) until convergence.

Therefore, we provide a second order approximation method which leads to a computationally efficient optimization procedure because only first order derivatives are involved. In other terms, differently from second order minimization techniques (e.g., Newton), the Hessians are never computed explicitly. This also contributes to obtain nicer convergence properties. We remark that the computational cost of the proposed second order approximation is equivalent to the cost of the Gauss-Newton method. Indeed, the cost of the addition of $\mathbf{J}(\widehat{\mathbf{x}})$ with $\widehat{\mathbf{J}}$ in (64), within iterations, is truly negligible compared to the cost of the pseudoinverse operation required by both methods. The gradient of the reference image is an one-off computation.

4.3 Initialization Issue

A common limitation of efficient nonlinear optimization procedures regards its domain of convergence. Although the parameters are obtained by a second order approximation method with nice convergence properties, it does not ensure that the global minimum will be reached. As previously stated, global minimization procedures are too computationally intensive to be performed in a real-time setting. Here,

[4] In this particular case, if the homography $\mathbf{G}$ is parametrized as an element of $\mathbb{SL}(3)$, the corresponding warping operator (3) is a group action of $\mathbb{SL}(3)$ on $\mathbb{P}^2$, i.e., $\mathbf{w}(\mathbf{G}_1\mathbf{G}_2,\mathbf{p}^*) = \mathbf{w}(\mathbf{G}_1,\mathbf{w}(\mathbf{G}_2,\mathbf{p}^*))$, $\forall\mathbf{G}_1,\mathbf{G}_2\in\mathbb{SL}(3)$.

we suppose that the image acquisition rate is sufficiently high so as to observe small displacements of the objects in successive images. This is generally true in robotic applications, where smooth camera motions are performed. In other terms, the parameters estimated in the registration of $\mathcal{I}^*$ with $\mathcal{I}^{(t)}$, where t indexes the images, are used here as a starting point for the alignment of $\mathcal{I}^*$ with $\mathcal{I}^{(t+1)}$. Nevertheless, we discuss in the sequel possible solutions if very large interframe displacements are present. We remark that none of the possibilities below are applied in this article.

A possible solution to avoid getting wedged in local minima within direct registration methods consists in using, for example, feature-based techniques as a bootstrap. In addition to augmenting the domain of convergence, this approach may also augment the rate of convergence. If the related parameters are closer to the true ones than those by using the minimization approach, they will act in this case as a prediction for aligning a new image of a video sequence. In any case however, feature-based techniques do not ensure that the global minimum will be attained, since they are not fully invariant to all possible photogeometric changes. Thus, one may also rely on other predictors to improve the convergence properties of direct methods. In fact, the coupling between the image registration method with a filtering technique can be performed at this stage. In the case of a sequential image registration task (i.e., visual tracking), a Kalman filter can be used to provide another estimate of the optimization variables. The input (i.e., observations) to the filtering are the recovered parameters from the minimization process. To initialize the system (i.e., when a new image is available), the best set of parameters among all predictors is chosen by comparing their resulting cost value. Nevertheless, filtering approaches also have limitations in providing sufficiently good predictions. The assumptions on the type of noise (e.g., Gaussian noise) and/or on the model of motion (e.g., constant velocity) may not be realistic in many scenarios.

5 Comparison Results

Let the photometric error be defined as the Root-Mean-Square (RMS) of the difference image between the photogeometrically transformed image $\mathcal{I}'_{gh}$ and the reference $\mathcal{I}^*$.

5.1 Affine Illumination Changes

Existing efficient direct image alignment techniques essentially tackle affine lighting variations. To show the generality of the proposed method, we compared it with an existing algorithm (Bartoli, 2008) that is designed for that particular context. A nonoptimized implementation of our method in C code runs at about 2.4 ms/iteration for a template size of 100 $\times$ 100 pixels and for this affine case (10 parameters to be estimated) on a monocore Pentium 3.2 GHz. See Fig. 6 for the processing times when varying those parameters. Comparison results of a particular image registration task is shown in Fig. 7. The image to be aligned presents relatively large geometric and photometric displacements with respect to the fixed image, and is thus adequate to illustrate the improvements gained by the method. Two conclusions can be drawn directly. First, the error obtained by our technique is always smaller through iterations. Second, the existing algorithm got stuck in a local minimum and thus, obtained a higher error at the convergence. We remark that the difference in the final photometric error is significant as it also reveals that the existing method is prone to fall into irrelevant minima. This means that for a different situation that error may be higher, as well as it may accumulate drifts (thus leading to a failure earlier) within a visual tracking task. With respect to other existing strategies for this particular context, it has been shown (Bartoli, 2008) that, albeit more computationally efficient, that existing algorithm yields exactly the same photometric error of the method proposed by Baker and Matthews (2004). The strategy presented by Jin et al. (2003) did not converge after 100 iterations and was not included in the figure.

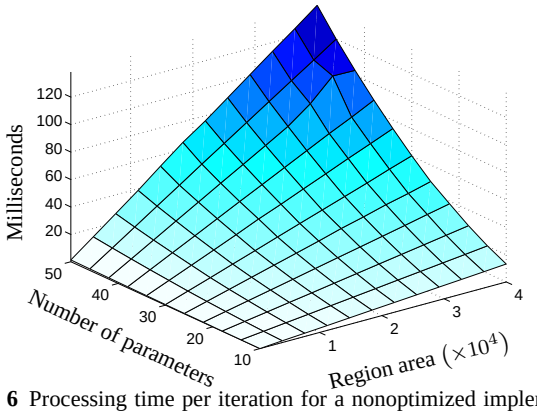
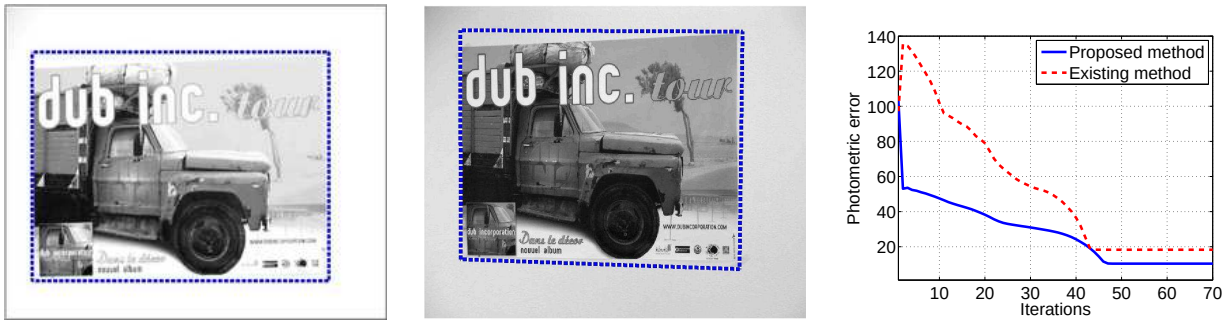


Fig. 6 Processing time per iteration for a nonoptimized implementation of our uncalibrated registration method in C on a Pentium 3.2 GHz.

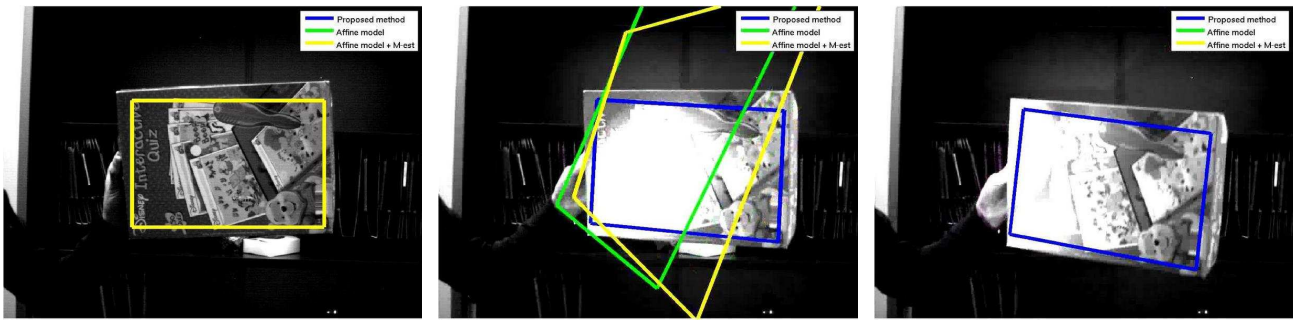
5.2 Generic Illumination Changes

BEAR *sequence* We have also applied the algorithm on a sequence under severe changes in ambient, diffuse and specular reflections. The unknown light sources are varied in power, type, number and moved in space. No existing efficient direct techniques are able to cope with this challenging scenario, especially when the object is not near-Lambertian and/or relatively large displacements are carried out. In all case, we have tried the above-mentioned strategies (Bartoli, 2008; Baker and Matthews, 2004; Jin et al., 2003), but they have failed. This includes their variants, for example, by performing a photometric normalization with/or using a ro-



(a) (b) (c)

Fig. 7 Comparison results of an image alignment task where relatively large displacements are present. As a means to compare with an existing method, the lighting variations between (a) the original image and (b) the synthetically transformed one comprise only affine changes. (c) The proposed method obtains smaller errors and does not get trapped into irrelevant minima.



(a) (b) (c)

Fig. 8 (Color online) BEAR sequence: Comparison results for the general case, using existing direct registration methods with and without a robust function. They are outlined in yellow and in green, respectively. Whereas both of them have failed, the proposed method (outlined in blue) successfully registers (a) the reference image to all other images of the sequence. Some excerpts are shown in (b) and (c).

bust error function (a M-estimator with Tukey's function). In fact, the experiments showed that, when the robust function leads to a convergence for a given image, it takes an average of 2 times more iterations. See Fig. 8 for some excerpts and Online Resource 1 for the entire sequence. The proposed method successfully registers all images with a median photometric error of 15.7 levels of grayscale (over 255), executing a median of 6 iterations per image, for the requested accuracy. The surface related to the illumination changes are approximated by discretization and has not been further interpolated. Each block has a fixed size of 50×50 pixels.

5.3 Optimization Methods

The proposed approach is also tested with a sequence of known ground truth. Indeed, a video sequence has been synthetically created by warping a textured sphere. We then compare the proposed second order minimization method with Gauss-Newton one. The same transformation model is applied to both cases. A region of size 400×400 pixels is selected in the first image as the reference template (see top row of Fig. 9). The centers for the surface approximation are placed on a regular (5×5) grid. To simulate a real-time experiment (we have 31 parameters to estimate), we fixed the number of iterations per image of each algorithm to 5.

Despite the simple spherical structure of the surface, the standard Gauss-Newton fails to register the images since it does not have enough iterations to converge (with 10 iterations/image it works fine). The middle row of Fig. 9 shows the corresponding registration results after 40 images. Observe that the regular grid is not transformed accordingly to a spherical surface, and the area of interest is not correctly registered with respect to the reference template. On the other hand, by using the efficient second order minimization the images are correctly registered (see last row of Fig. 9). The average RMS error for the registration of the 40 images is of 4.9 levels of grayscale (over 255).

6 Experimental Results

The generality and robustness of the proposed direct image registration technique are verified in this section through tracking rigid and deformable objects, with and without severe lighting variations, using both grayscale and color images. To this end, we select a template in the reference (i.e., first) image $\mathcal{I}^*$, which is then optimally aligned to succes-

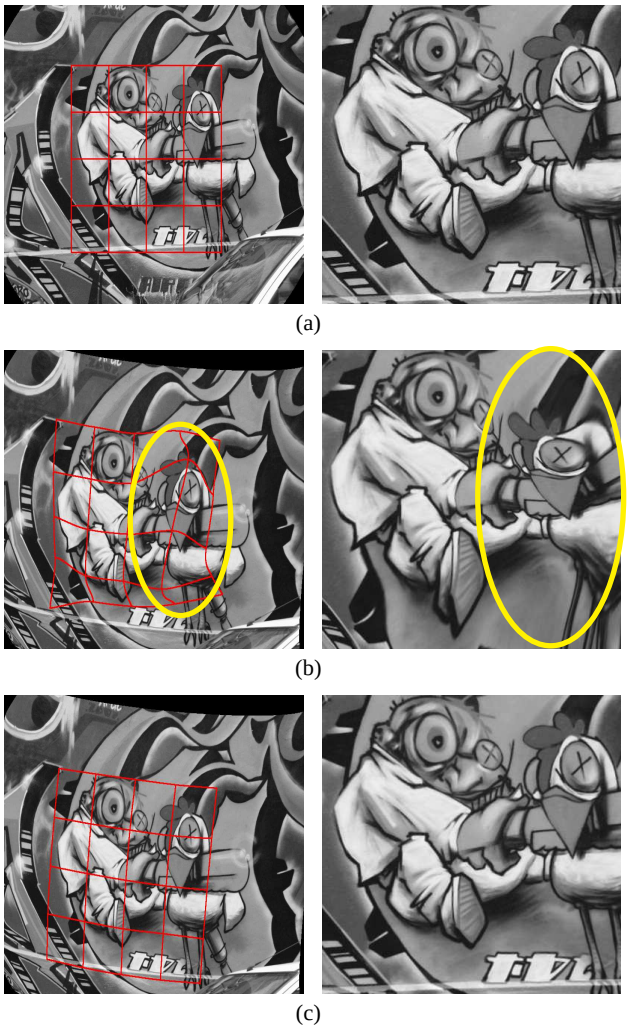


Fig. 9 Comparison results between optimization methods for sequentially registering a 40-image sequence. (a) Left frame shows the reference image, whereas the selected reference template is shown at the right. (b) Results for the Gauss-Newton method. The registration fails due to real-time constraints. This is clearly visible in the right image where the warped area of interest in the reference frame is not equal to the reference template. (c) Successful results for the proposed optimization method.

sive images of the sequence. The hierarchical approach described in Section 3.1 is used, where the observed surface is initially supposed to be a 3D plane parallel to the image plane. We emphasize that the proposed algorithm does not require any off-line training step, that any prediction technique (e.g., Kalman filter) is applied in this article, and also that noncausal estimation is not performed in any case. The parameters estimated in the registration of $\mathcal{I}^*$ with $\mathcal{I}^{(t)}$, where t indexes the images, are used here as a starting point for the alignment of $\mathcal{I}^*$ with $\mathcal{I}^{(t+1)}$. Table 1 describes the supplemental material and summarizes the details of all experiments.

6.1 Purely Geometric Direct Visual Tracking

6.1.1 Rigid Surfaces

VAULT *sequence* In this experiment we track a smooth vault which is painted in still-life deception. Some excerpts from the tracked sequence are given in the top row of Fig. 10 (see Online Resource 2 for the entire sequence). A regularly spaced (11 $\times$ 9) grid is used, leading to a total of 110 parameters to be estimated. In the bottom row, the first image shows the reference template. The other images correspond to the current images warped, to the first frame, with the estimated parameters. The images are correctly registered, with an average photometric error of approximately 6.6 levels of grayscale (over 255), and an average of 6 iterations of the algorithm per image. Observe that the template partly goes out of the image (see the upper corner of last image), without any perturbation on the registration. In other terms, the object must not be fully visible in the images.

BALL *sequence* In this experiment we track a basketball in a sequence of images acquired with an uncalibrated camera. Some excerpts from the tracked sequence are given in the top row of Fig. 11 (see Online Resource 3 for the entire sequence). The bottom row shows that the area of interest is successfully registered with respect to the template. A regularly spaced (3 $\times$ 3) grid is used. The average photometric error is of approximately 13.6 levels of grayscale (over 255), and the algorithm performed an average of 7 iterations per image for the required accuracy.

6.1.2 Deformable Surfaces

PAPER *sequence* A deforming sheet of paper is tracked in this experiment, using a regularly spaced (6 $\times$ 5) grid of size 251 $\times$ 201 pixels. All possible deformations are estimated using the hierarchical approach, leading to a total of 98 parameters to be recovered. Some excerpts of the results are shown in Fig. 12 (see Online Resource 4 for the entire sequence). The bottom row shows that all images of the sequence have been correctly aligned with the reference template, with an average photometric error of approximately 14.5 levels of grayscale (over 255).

6.2 Photogeometric Direct Visual Tracking

We have also applied the algorithm on several real-world sequences under generic illumination changes. They present severe variations in ambient, diffuse and specular reflections as well as shadows, interreflections and glints. In addition, they comprise relatively large geometric displacements and objects with unknown reflectance properties. The surfaces

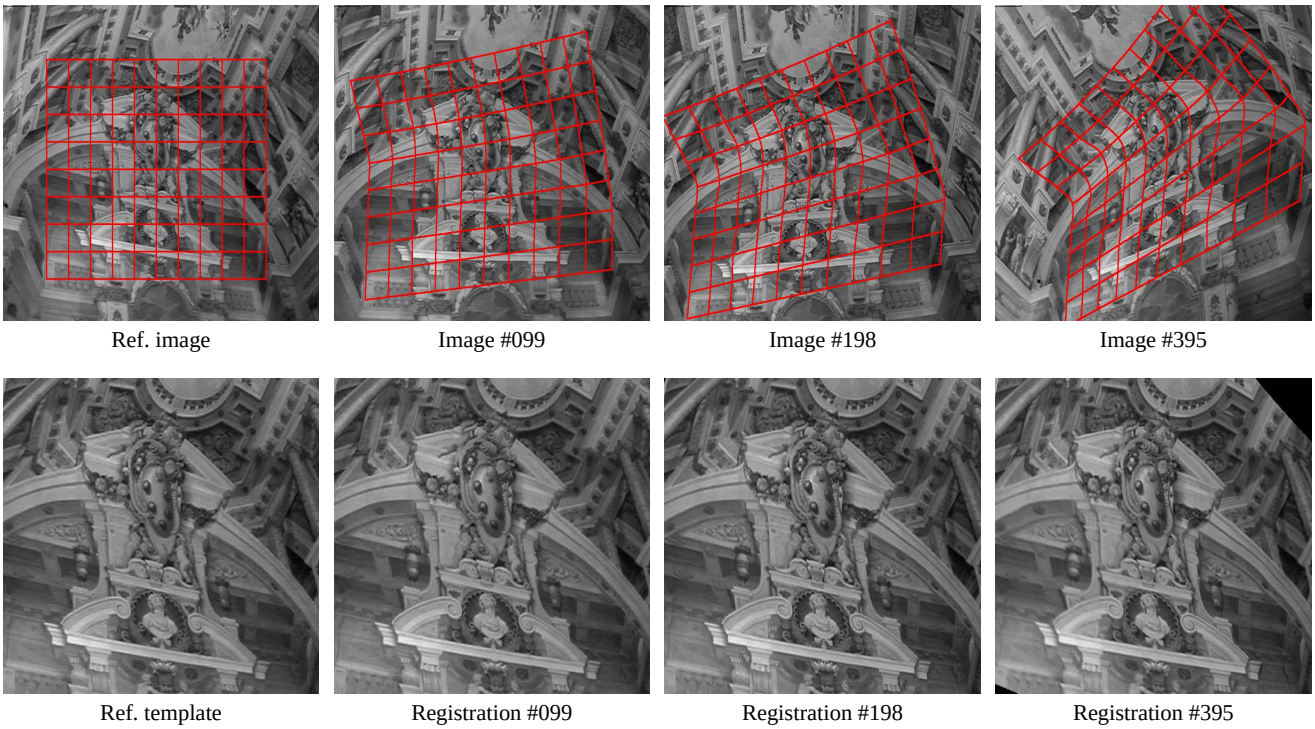


Fig. 10 VAULT sequence: Visual tracking of a reference template in a sequence of images acquired with an uncalibrated camera. (Top) Warped grid is superimposed on the tracked area. (Bottom) The area of interest is registered with respect to the template. Last image shows that the template can partly go out of the image without perturbing the task.

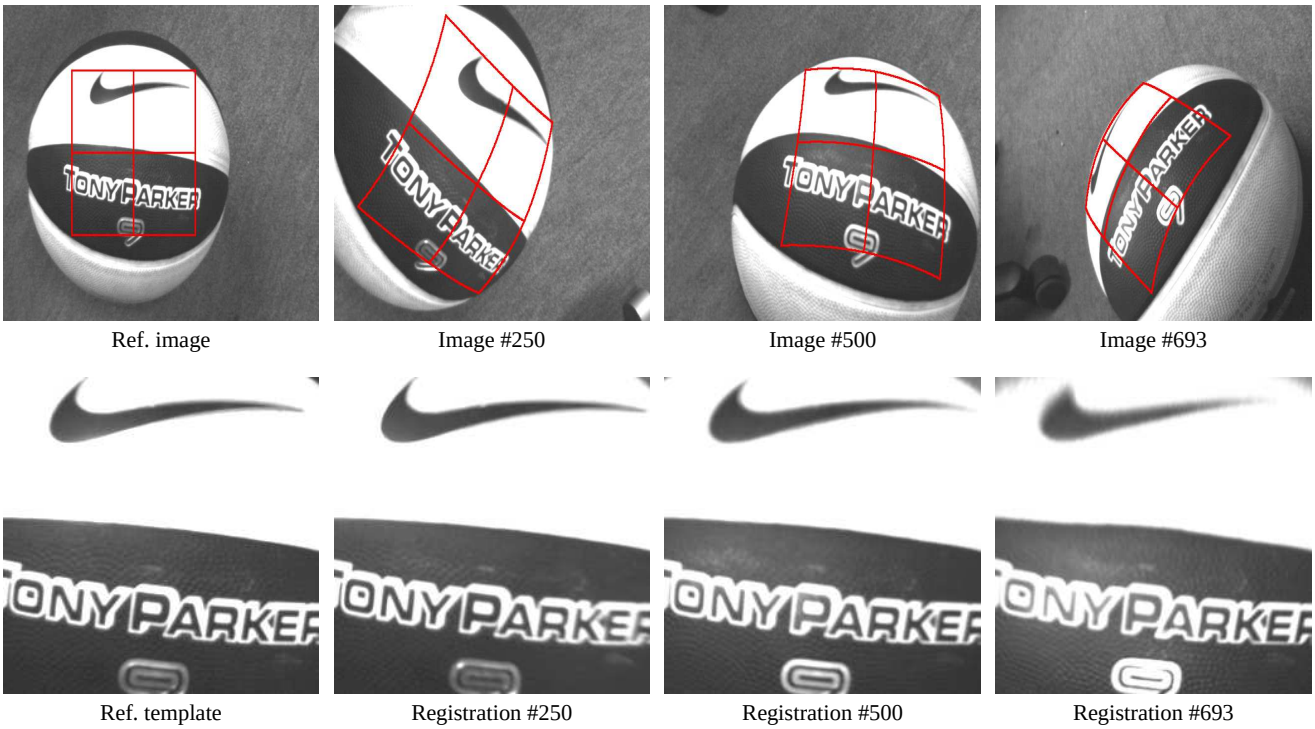


Fig. 11 BALL sequence: Visual tracking of a basketball in a sequence of images acquired with an uncalibrated camera. (Top) Warped grid is superimposed on the tracked area. (Bottom) Registered images with respect to the template. They demonstrate the stability of the tracker.

Table 1 Index to multimedia Online Resources and their implementation details. Legend: Template refers to the size in pixels (width × height) of the reference region; #Iter/Img stands for the median number of Iterations per Image; GM is the applied Geometric Model, where PO is for Planar Object, RO is for Rigid Object, IDEF for Invariant Deformation, and GDEF is for General Deformation; GS denotes the type of Surface used for the Geometric model, where RBF stands for Radial Basis Function, D for Discretization and I for cubic Interpolation; #GP is the number of Geometric Parameters estimated; PM is the applied Photometric Model, where S signifies that a Surface is estimated (so as to model local variations), and G means Global parameters; PS denotes the type of Surface for the Photometric model; #PP is the number of Photometric Parameters estimated; and RMS is the median photometric error obtained (over 255 levels of grayscale) along the sequence.

Resource	Type	Description	#Images	Template	#Iter/Img	GM	GS	#GP	PM	PS	#PP	RMS
1	Video	Comparison result: BEAR seq.	953	367 × 244	6	PO	-	8	S+G	D	41	15.7
2	Video	Tracking result: VAULT seq.	396	500 × 400	6	RO	RBF	110	-	-	-	6.6
3	Video	Tracking result: BALL seq.	694	250 × 250	7	RO	RBF	20	-	-	-	13.6
4	Video	Tracking result: PAPER seq.	1365	251 × 201	6	GDEF	RBF	98	-	-	-	14.5
5	Video	Tracking result: BOOK seq.	183	251 × 201	7	PO	-	8	S+G	RBF	31	5.4
6	Video	Tracking result: BEAR-II seq.	1783	427 × 318	4	PO	-	8	S+G	D	64	14.3
7	Video	Tracking result: BALLOON seq.	1083	262 × 262	5	IDEF	D+I	27	G	-	1	5.6
8	Video	Tracking result: *color* CAT seq.	898	250 × 250	9	PO	-	8	3S+3G	D	78	15.7
9	Video	Tracking result: *color* CAT-II seq.	175	150 × 225	7	RO	D+I	23	3S+3G	D	165	16.8

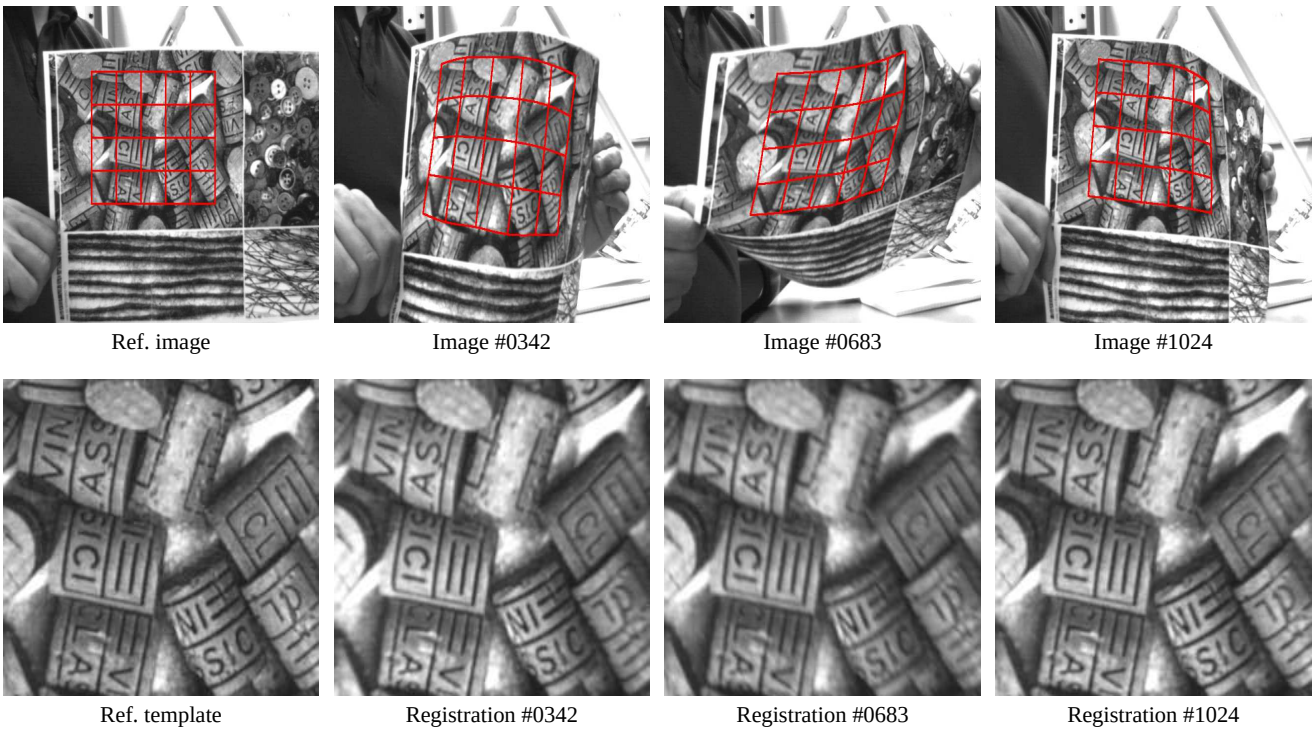


Fig. 12 PAPER sequence: Visual tracking of a deformable surface with an uncalibrated camera. (Top) Warped grid is superimposed on the tracked area. (Bottom) The area of interest is registered with respect to the template. The registered images show the stability of the tracker.

ranged from smooth to rough, and including metal and dielectric objects. The unknown light sources are varied in power, type, number and moved in space.

6.2.1 Rigid Surfaces

BOOK *sequence* In this experiment we track a planar object under variable reflections. The specular component is primarily produced by a line source, albeit no assumptions of its characteristics are made. Some excerpts from the tracked sequence are shown in Fig. 13 (see Online Resource 5). The third row shows the estimated illumination changes relatively to the reference image. They are shown as an evolving surface so as to emphasize how these changes are viewed in this article. The error images between the reference template and photogeometrically transformed images are given in the bottom row. They are nearly all black for all sequences. The centers are apart from each other by 50 pixels.

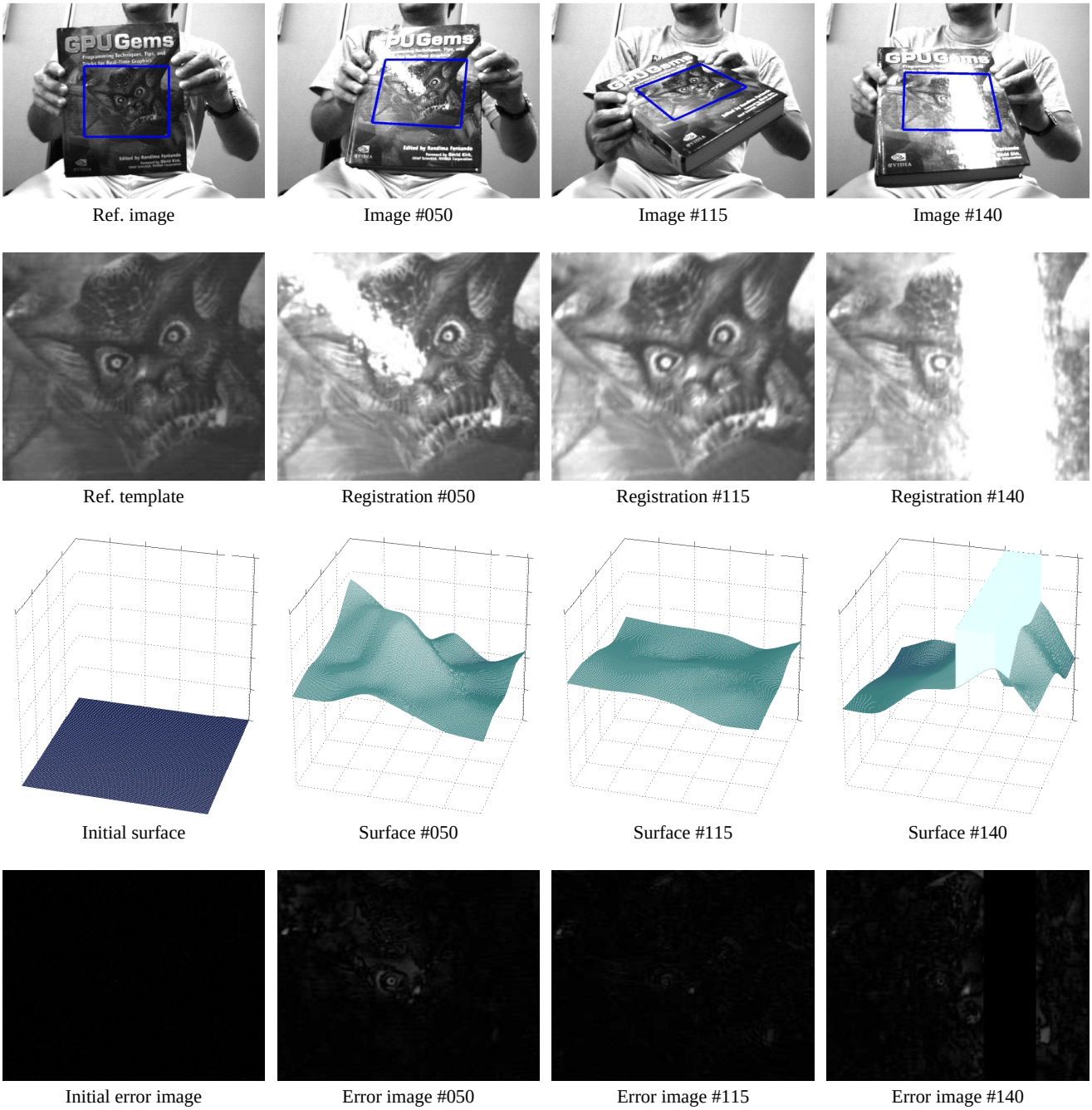


Fig. 13 BOOK sequence: Direct image registration of a reference template to successive frames of a video sequence. The sequence contains severe changes in the specular, diffuse and ambient reflections. (Third row) The estimated surface represents the illumination changes with respect to the reference template. (Bottom) Error images after the registration and photometric transformation (using the estimated surface) of the current image with respect to the reference template.

BEAR-II *sequence* Some results obtained for another grayscale sequence are shown in Fig. 14 (see Online Resource 6 for the entire sequence). For the requested accuracy, the approach performed along these sequences a median of 4 iterations per image, and returned a median photometric error of 14.3 levels of grayscale (over 255). The surface related to the illumination changes are approximated by discretization and has not been further interpolated.

6.2.2 Deformable Surfaces

BALLOON *sequence* In this experiment we track a deforming balloon with an uncalibrated camera. We selected in the first image a template of size 262 $\times$ 262 pixels, and placed the centers on a regularly spaced (4 $\times$ 4) grid. Here we do not use RBFs for surface approximations. We use bicubic interpolation to approximate the surface given the

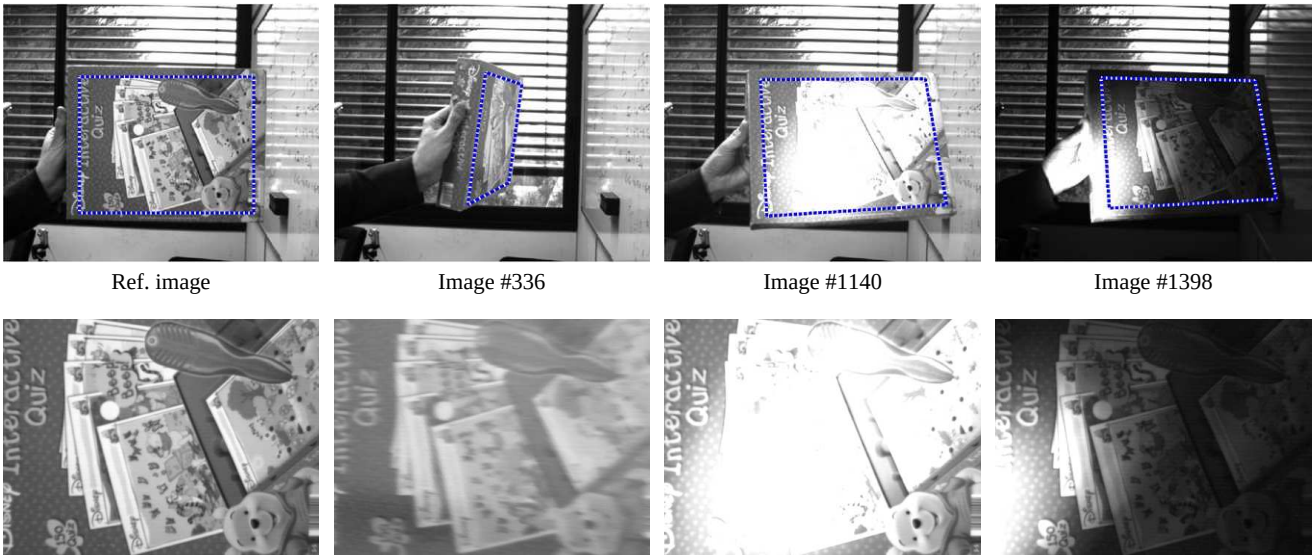


Fig. 14 BEAR-II sequence: Sequence with large surface obliquity and instantaneous changes in lighting. During the tracking, a large part of the region has been occluded by the highlight. (Bottom) Registered images demonstrate the stability of the proposed visual tracking technique.

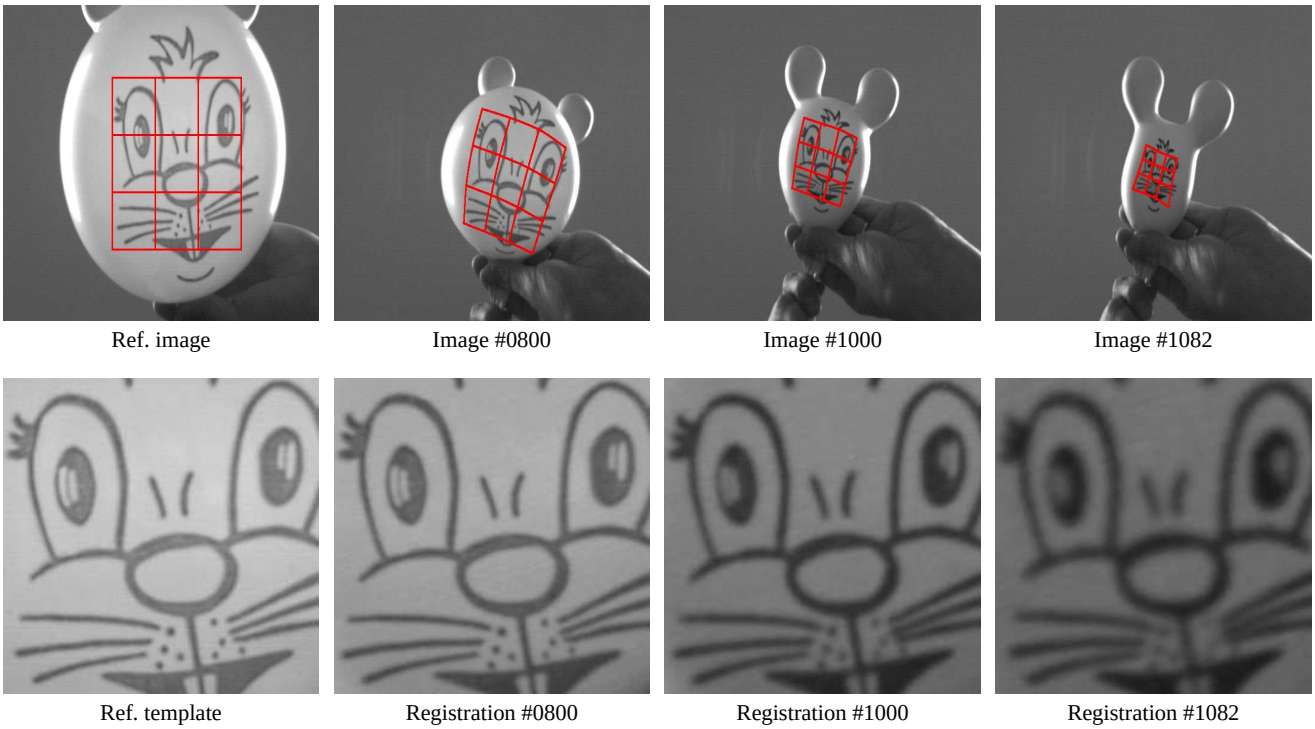


Fig. 15 BALLOON sequence: Visual tracking of a deformable surface in a sequence of images acquired with an uncalibrated camera. (Top) The regular grid used to track the area of interest in the sequence. (Bottom) the area of interest registered with respect to the template.

centers. In the hierarchical approach to describe the surfaces, it was sufficient to estimate up to an invariant deformation of the object (i.e., $\boldsymbol{\delta}^* = \mathbf{0}$) and only the compensation of ambient illumination changes has been needed. The bottom row of Fig. 15 (see Online Resource 7 for the entire tracked sequence) shows that all images of the sequence have been correctly aligned with the reference template, despite the lighting variations and a large change in the balloon's size and its deformations. The average photometric error for this sequence is around 5.6 levels of grayscale (over 255).

6.3 Photogeometric Direct Visual Tracking in Color Images

It is shown here some tracking results for color images using different objects, including that of a nonplanar rigid object. No prior knowledge of the object's attributes (e.g., shape, albedos) is exploited.

6.3.1 Rigid Surfaces

CAT *sequence* Some excerpts from this experiment are given in Fig. 16 (see Online Resource 9 for the entire tracked sequence). In spite of severe specularities, shadows and instantaneous changes in diffuse and ambient reflections, the bottom row shows that the images are successfully registered with respect to the template. Last image also shows that the tracked object partly goes out of the image without problems. For the requested accuracy, the approach performed along these sequences a median of 9 iterations per image, and returned a median photometric error of 15.73 levels of grayscale (over 255).

CAT-II *sequence* In this experiment we have used the same color pattern as in the previous sequence, although with an object of different shape. Of course, this prior knowledge is not provided to the algorithm. Some excerpts of the tracking results are shown in Fig. 17 (see Online Resource 10 for the entire tracked sequence). Once again, a challenging scenario is set up with very disparate types of lighting variations, and the images are successfully aligned. For the requested accuracy, the approach performed along these sequences a median of 7 iterations per image, and returned a median photometric error of 16.76 levels of grayscale (over 255).

7 Conclusions

We have proposed a general and robust direct image registration technique for tracking various classes of objects despite generic lighting changes, even in color images. Indeed, a concise unified warping model is proposed so that rigid and deformable surfaces can be successfully tracked. Severe lighting changes are handled via a new model of illumination changes. We propose to view these changes as an evolving surface. In this way, even local variations can be adequately modeled. Of course, the cost of processing is dependent on the number of parameters to be estimated, which increases with increasing complexity of the surfaces.

All parameters of the proposed photogeometric transformation model are simultaneously estimated by an efficient second-order optimization procedure. It is computationally efficient because the Hessians are never calculated explicitly. In addition, the proposed procedure allows the object to undergo relatively large interframe displacements without getting trapped in irrelevant minima, and to partly go out of the image. Nevertheless, as with any direct image registration method, the object must still be sufficiently textured. Furthermore, the surfaces must be at least piecewise smooth so that the cost function can be expanded in Taylor series. Finally, a promising research direction consists in overcoming the main limitation of real-time direct registration methods, i.e., its limited domain of convergence. Some of possible solutions to overcome this limitation are briefly discussed in the article.

Comparisons results with existing direct techniques show significant improvements in the tracking performance. Extensive experiments using rigid and deformable objects, with and without severe lighting variations, and using both grayscale and color images, confirm the generality and robustness of our method.

Acknowledgments

This work is also partially supported by the Brazilian CAPES Foundation under grant no. 1886/03-7, and by the international agreement FAPESP-INRIA under grant no. 04/13467-5.

References

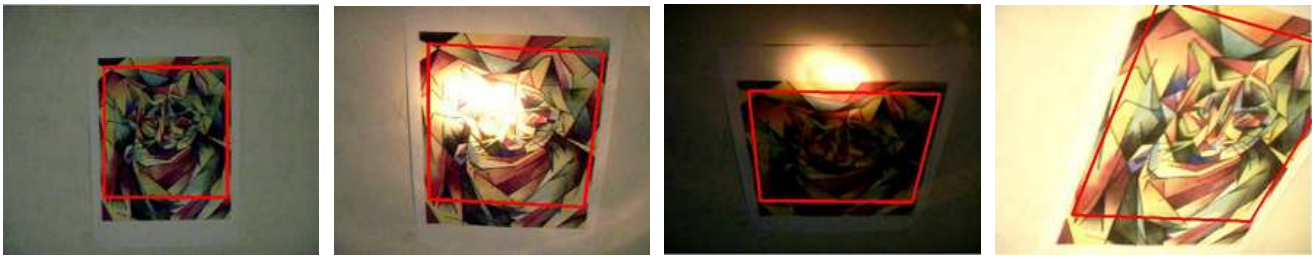
Baker, S. and Matthews, I. (2001). Equivalence and efficiency of image alignment algorithms. In *IEEE Computer Vision and Pattern Recognition*, pages 1090–1097.

Baker, S. and Matthews, I. (2004). Lucas-Kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255.

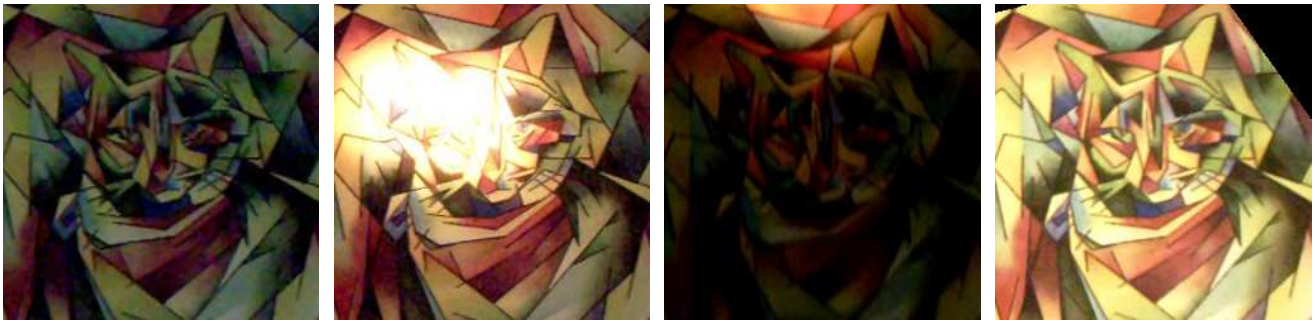
Bartoli, A. (2008). Groupwise geometric and photometric direct image registration. *IEEE Trans. Pattern Anal. Machine Intell.*, 30(12):2098–2108.

Bartoli, A. and Zisserman, A. (2004). Direct estimation of non-rigid registration. In *Proc. of the British Machine Vision Conference*, pages 899–908.

Benhimane, S. and Malis, E. (2007). Homography-based 2D visual tracking and servoing. *Int. J. Rob. Res.*, 26(7):661–

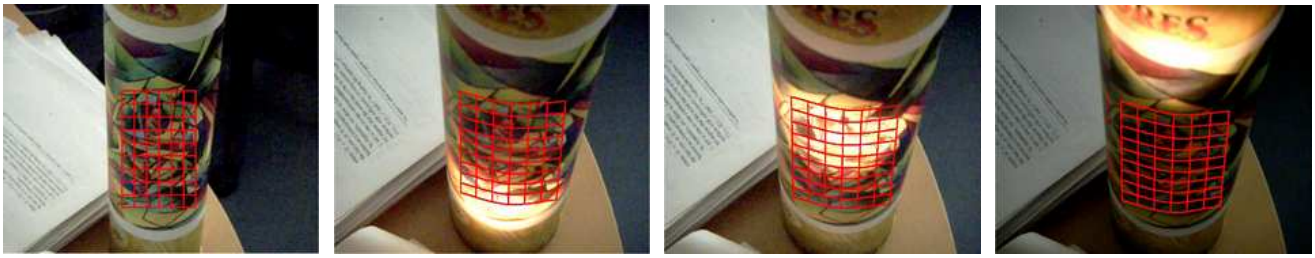


Ref. Image | Image #102 | Image #224 | Image #624



Ref. template | Registration #102 | Registration #224 | Registration #624

Fig. 16 CAT sequence: Direct image registration of a reference image to successive color frames of a video sequence. The sequence contains severe changes in the specular, diffuse and ambient reflections. (Bottom) The registered images demonstrate the stability of the tracker. Last image shows that the template can partly go out of the image without problems.



Ref. Image | Image #147 | Image #154 | Image #162

Ref. template | Registration #147 | Registration #154 | Registration #162

Fig. 17 CAT-II sequence: Direct visual tracking of a reference color image. The unknown light source and camera perform unknown motions in space. No prior knowledge of the object is exploited. (Bottom) Registered images demonstrate the stability of the proposed visual tracker.

676. Special Issue on Vision and Robotics joint with the *International Journal of Computer Vision*.

Black, M. J., Fleet, D. J., and Yacoob, Y. (2000). Robustly estimating changes in image appearance. *Computer Vision and Image Understanding*, 78:8–31.

Blinn, J. F. (1977). Models of light reflection for computer synthesized pictures. In *SIGGRAPH*, pages 192–198.

Brown, L. G. (1992). A survey of image registration techniques. *ACM Computing Surveys*, 24:325–376.

Carr, J., Fright, W., and Beatson, R. (1997). Surface interpolation with Radial Basis Functions for medical imaging. *IEEE Transactions on Medical Imaging*, 16(1).

Comaniciu, D., Ramesh, V., and Meer, P. (2000). Real-time tracking of non-rigid objects using mean-shift. In *IEEE Computer Vision and Pattern Recognition*.

Cook, R. and Torrance, K. (1982). A reflectance model for computer graphics. *ACM Trans. Graphics 1*, pages 7–24.

Faugeras, O., Luong, Q.-T., and Papadopoulo, T. (2001). *The geometry of multiple images*. The MIT Press.

Finlayson, G., Drew, M., and Funt, B. (1994). Color constancy: Generalized diagonal transforms suffice. *J. Opt. Soc. Am. A*, 11(11):3011–3020.

Gouiffès, M., Collewet, C., Fernandez-Maloigne, C., and Trémeau, A. (2006). Feature points tracking using photometric model and colorimetric invariants. In *Proc. Eur. Conf. on Colour in Graph., Imag., and Vis.*, pages 18–23.

Hager, G. and Belhumeur, P. (1998). Efficient region tracking with parametric models of geometry and illumination. *IEEE Trans. on Patt. Analysis and Machine Intell.*, 20(10):1025–1039.

Hartley, R. and Zisserman, A. (2000). *Multiple View Geometry in Computer Vision*. Cambridge Univ. Press.

Haussecker, H. W. and Fleet, D. J. (2001). Computing optical flow with physical models of brightness variation. *IEEE Trans. on Patt. Analysis and Machine Intell.*, 23(6).

Horst, R. and Pardalos, P. M., editors (1995). *Handbook of Global Optimization*. Kluwer.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons.

Irani, M. and Anandan, P. (1999). All about direct methods. In *Proc. Workshop on Vision Alg.: Theory and practice*.

Jin, H., Favaro, P., and Soatto, S. (2001). Real-time feature tracking and outlier rejection with changes in illumination. In *Proc. of the IEEE International Conference on Computer Vision*, pages 684–689.

Jin, H., Favaro, P., and Soatto, S. (2003). A semi-direct approach to structure from motion. *The Visual Computer*, 6:377–394.

Jurie, F. and Dhome, M. (2002). Real time robust template matching. In *Proc. of the British Machine Vision Conference*, pages 123–131.

Klinker, G. J., Shafer, S. A., and Kanade, T. (1990). The measurement of highlights in color images. *International Journal of Computer Vision*, 2:7–32.

La Cascia, M., Sclaroff, S., and Athitsos, V. (2000). Fast, reliable head tracking under varying illumination: An approach based on robust registration of texture-mapped 3d models. *IEEE Trans. on Patt. Analysis and Machine Intell.*, 22:322–336.

Lai, S.-H. and Fang, M. (1999). Robust and efficient image alignment with spatially varying illumination models. In *IEEE Computer Vision and Pattern Recognition*.

Lucas, B. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proc. of the Int. Joint Conf. on Art. Intell.*, pages 674–679.

Maintz, J. B. and Viergever, M. A. (1998). A survey of medical image registration. *Med. Image Anal.*, 2(1):1–36.

Malis, E. (2004). Improving vision-based control using Efficient Second-order Minimization techniques. In *Proc. of the IEEE Int. Conf. on Robotics and Automation*, USA.

Malis, E. (2007). An efficient unified approach to direct visual tracking of rigid and deformable surfaces. In *Proc. of the IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, USA.

Montesinos, P., Gouet, V., Deriche, R., and Pele, D. (1999). Matching color uncalibrated images using differential invariants. *Image and Vision Computing*, 18(9):659–671.

Mégret, R., Authesserre, J.-B., and Berthoumieu, Y. (2008). The bi-directional framework for unifying parametric image alignment approaches. In *Proc. of the European Conference on Computer Vision*.

Nastar, C., Moghaddam, B., and Pentland, A. (1996). Generalized image matching: Statistical learning of physically-based deformations. In *Proc. Eur. Conf. on Comp. Vision*.

Negahdaripour, S. (1998). Revised definition of optical flow: Integration of radiometric and geometric cues for dynamic scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(9):961–979.

Shum, H. Y. and Szeliski, R. (2000). Construction of panoramic image mosaics with global and local alignment. *Int. Journal of Computer Vision*, 16(1):63–84.

Silveira, G. and Malis, E. (2007a). Direct visual servoing with respect to rigid objects. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, USA.

Silveira, G. and Malis, E. (2007b). Real-time visual tracking under arbitrary illumination changes. In *IEEE Computer Vision and Pattern Recognition*, USA.

Silveira, G., Malis, E., and Rives, P. (2008). An efficient direct approach to visual SLAM. *IEEE Transactions on Robotics*, 24:969–979.

Szeliski, R. (2005). Image alignment and stitching. In Paragios, N., Chen, Y., and Faugeras, O., editors, *Handbook of Math. Models in Comp. Vision*, pages 273–292. Springer.

Tan, R. and Ikeuchi, K. (2005). Separating reflection components of textured surfaces using a single image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(2):178–193.

Varadarajan, V. (1974). *Lie groups, Lie algebras, and their representations*. Prentice-Hall.

Warner, F. W. (1987). *Foundations of differential manifolds and Lie groups*. Springer Verlag.