MINES PARIS
ParisTech

THÈSE

préparée à

L'INRIA Sophia Antipolis – Méditerranée

et présentée à

L'ÉCOLE NATIONALE SUPÉRIEURE DES MINES DE PARIS

pour obtenir le grade de

DOCTEUR EN SCIENCES

spécialité

Informatique Temps Réel, Robotique et Automatique

soutenue publiquement le 29 octobre 2008 par

Geraldo SILVEIRA Filho

sur le thème

Contributions aux méthodes directes d'estimation et de commande basées sur la vision

selon la traduction en anglais intitulée

Contributions to direct methods of estimation and control from visual data

devant le jury composé de :

Dr.	Yves	ROUCHALEAU	Président
Dr.	Marie-Odile	BERGER	Rapporteur
Dr.	François	CHAUMETTE	Rapporteur
Dr.	Samuel	BUENO	Examinateur
Dr.	Ezio	MALIS	Directeur de Thèse
Dr.	Bruno	SICILIANO	Examinateur

To my first professors, my parents.

Contents

Avant-propos

Cette thèse concerne, à différents niveaux, les domaines fondamentaux de recherche liés à la navigation autonome des robots : localisation, cartographie, planification et commande. Une synthèse récente de ces disciplines (et beaucoup d'autres) sont rassemblées dans (Siciliano and Khatib, 2008). En particulier, on considère ici le cas où toutes les observations sur l'environnement, et sur les états internes du système, sont fournies par une seule caméra. Dans la communauté internationale de robotique, ce cadre monoculaire est également appelé *bearing-only*. Ce terme traduit le fait qu'une simple mesure est insuffisante pour fournir une indication de profondeur. En fait, une des difficultés majeures de cette modalité de détection est liée à ce problème d'inobservabilité. Notons que, selon leur positionnements globaux, le fait d'ajouter plusieurs caméras au système pour lever ce problème d'inobservabilité peut ne pas suffire. En effet, si l'environnement est trop éloigné par rapport aux distances entre les caméras, seule l'information de *bearing* sera fournie de façon fiable. La maîtrise du cadre monoculaire est donc une problématique fondamentale.

Pour l'accomplissement de la plupart des tâches robotiques, un problème central concerne le développement de techniques efficaces, précises et robustes d'estimation monoculaire. En effet, afin de naviger de façon autonome, un robot doit pouvoir créer une représentation de son environnement et estimer sa pose relative par rapport à celui-ci. L'estimation basée sur la vision concerne le développement de modèles et méthodes pour inférer et traiter, à partir des images, l'information nécessaire à une tâche donnée. Dans cette thèse, l'efficacité computationnelle désigne la capacité d'exploiter toute l'information visuelle possible, tout en respectant les exigences de calcul temps réel. L'exigence de robustesse est également essentielle, afin de pouvoir obtenir des estimations précises malgré les erreurs inévitables de modélisation et de mesure.

Ainsi, une partie importante de cette thèse concerne la conception de modèles paramétriques appropriés, ainsi que de méthodes d'estimation, pour récupérer des paramètres relatifs, tout en respectant les exigences de performance. Nous nous concentrons sur des méthodes de recalage d'image pour obtenir les paramètres. Le recalage d'image consiste à estimer les transformations qui permettent de recaler au mieux une image de référence (fixe) par rapport à une deuxième image (en mouvement). En outre, une considération particulière est donnée ici aux méthodes directes de recalage (Irani and Anandan, 1999). Dans cette classe de méthodes, les valeurs d'intensité des pixels sont directement exploitées pour obtenir les paramètres, en s'affranchissant des étapes d'extraction et de mise en correspondance de primitives géométriques (e.g. points, lignes). Dans la communauté de vision par ordinateur, ces méthodes directes sont souvent désignés par les termes “méthodes basées sur l’intensité”, “sur l’apparence”, voir “sur la texture”. Nous discuterons en détail les avantages et les limitations de ces techniques. En effet, nous cherchons les modèles et les méthodes directes, simples, précis, et génériques pour effectuer diverses tâches basées sur la vision, telles que le suivi visuel et le SLAM (pour *Simultaneous Localization And Mapping*) visuel, c’est-à-dire, la localisation et cartographie simultanées de l’environnement, par rapport à un système de coordonnées donné. D’autres applications concernent l’imagerie médicale (Maintz and Viergever, 1998) et la réalité augmentée (Berger and Simon, 1998).

Une autre partie importante de cette thèse porte sur la commande basée sur la vision (Chaumette and Hutchinson, 2006). Cette branche de l’automatique a suscité beaucoup d’attention pendant presque deux décennies. En effet, elle constitue également un domaine fondamental de recherche lié à la navigation de robots autonomes, et elle se situe au carrefour de nombreuses disciplines, comme l’automatique et l’estimation basée sur la vision. Cette thèse propose différentes manières d’augmenter la flexibilité et la fiabilité des techniques existantes, en considérant le cadre d’objets inconnus dans des conditions inconnues de formation d’image. Ceci sera montré par deux approches différentes de commande.

Contributions de la Thèse. Les contributions majeures de cette thèse peuvent être énumérés comme suit :

1. Nouveau modèle de changement d’illumination pour des méthodes directes de recalage d’image. L’application de ce modèle à des objets planaires de divers types sous conditions inconnues de formation d’image est présentée dans (Silveira and Malis, 2007c; Silveira and Malis, 2007d). On montre que ce modèle permet de traiter effectivement des objets de forme inconnue sous des conditions inconnues de formation d’image ;
2. Généralisation du modèle de changement d’illumination proposé à toute type d’image couleur. Ce modèle photométrique générique permet de coupler tous les canaux d’image afin de faire face aux objets inconnus sous des conditions inconnues d’illumination, dont les images formées à partir de caméras de caractéristiques inconnues. Notre technique non-calibrée robuste et générique de recalage d’image est établie sur ce modèle ;
3. Nouvelle formulation du problème de SLAM visuel comme une tâche de recalage directe d’image. Un cas particulier de cette approche est présenté dans (Silveira et al., 2007; Silveira et al., 2008c). La version étendue de ces documents a été publiée dans (Silveira et al., 2008a). Cette solution correspond à notre technique calibrée robuste et efficace de recalage d’image ;
4. Nouvelle technique d’asservissement visuel directe qui ni nécessite ni n’estime aucune information métrique sur la forme de l’objet et/ou sur le mouvement de le caméra (Silveira and Malis, 2007a; Silveira and Malis, 2008). Une version étendue de ces articles a été publiée dans (Silveira and Malis, 2007b). Ceci correspond à notre technique de commande référencée vision robuste et générique où une image de référence est fournie ;

5. Démonstration qu'une planification de trajectoire simple est suffisante pour rassurer un domaine de convergence très grand pour l'asservissement visuel, si la technique directe proposée est appliquée. Cette propriété est montrée dans (Silveira and Malis, 2007a; Silveira and Malis, 2008; Silveira and Malis, 2007b) ;

6. Nouveau schéma d'asservissement visuel où ni le modèle de la scène ni l'image de référence sont disponibles a priori. Un cas particulier de ce schéma est décrit dans (Silveira et al., 2006b). Une version étendue de ce papier a été publiée dans (Silveira et al., 2008b). Ceci correspond à notre technique de commande basée sur la vision robuste et efficace où une pose de référence est donnée ;

7. Nouveau détecteur de régions planaires dans une paire d'images non-calibrée (Silveira et al., 2006a). Ce détecteur est un composant du schéma proposé de commande basée sur la vision efficace où la pose de référence est donnée.

Organisation de la Thèse. Cette thèse est organisée comme suit :

La Partie I est consacrée principalement à un bref rappel sur le background nécessaire. Des modèles de base et des méthodes utiles sont présentés ici. En outre, les problèmes de recherche abordés dans cette thèse sont informellement formulés dans cette première partie. Une remarque importante est que l'état de l'art des techniques liées à ces problèmes ne sont pas présentées ici. Nous avons préféré les décrire dans le cours du document.

La Partie II présente les contributions liées à l'estimation directe. Elle comporte les contributions à partir de l'item 1 au 3 au-dessus, et est articulée en trois chapitres. Cette partie propose un cadre unifié pour le recalage directe d'images qui rassemble les configurations non-calibré et calibré. Des modèles et méthodes communs aux deux configurations sont présentés dans le premier chapitre de cette partie. Les spécificités de chaque cas sont alors prises en considération dans différents chapitres, ainsi que les techniques de l'état de l'art.

La Partie III décrit les contributions liées à la commande directe. Elle comporte les contributions à partir de l'item 4 au 7 au-dessus, et est articulée en deux chapitres. Le premier chapitre aborde le problème de commande basé sur la vision où les signaux désirés (i.e. à atteindre) sont donnés par une image de référence. Dans le deuxième chapitre de cette dernière partie, les signaux désirés sont directement définis dans l'espace Cartésien.

Les conclusions générales et des directions possibles pour les travaux futurs sont ensuite discutées, en termes d'estimation et de commande à partir d'informations visuelles. Les annexes incluent, notamment, toutes les démonstrations théoriques.

Le parti pris a été fait d'écrire une thèse concise. Au lieu de fournir un compendium approfondi de tous les concepts relatifs et formulations existantes, seulement ceux que nous pensons utiles et essentiels à la compréhension des contributions sont décrits. Des références soigneusement choisies sont naturellement données pour plus de détails.

Remerciements. Je tiens à remercier l'aide financière de la Fondation brésilienne Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) processus numéro 1886/03-7, l'accord international entre la Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)–Institut National de Recherche en Informatique et en Automatique (INRIA) processus numéro 04/13467-5, et du Centro de Tecnologia da Informação Renato Archer (CTI), ancien CenPRA.

Preface

Although at different levels, this thesis concerns the fundamental research domains related to autonomous navigation of robots: localization, mapping, planning and control. A recent review of these topics (and many others) are collected in (Siciliano and Khatib, 2008). In particular, it is considered here the framework where all observations of the surrounding environment, and of the internal states of the system, are provided by a single camera. In the robotics community, this monocular framework is also referred to as bearing-only. This term is due to the fact that a single measurement is insufficient to provide an indication of range. As a matter of fact, one of the major difficulties of this sensing modality is related to this observability issue. It can be noted that, depending on the overall setting, the fact of adding more cameras to the system may not help. Indeed, if the environment is sufficiently distant with respect to the baselines, then only bearing information will be provided anyway. Hence, expertise in monocular frameworks is desired.

In order to accomplish those usually non-trivial yet fundamental robotic tasks, a central issue concerns the development of efficient, accurate and robust techniques of monocular estimation. Indeed, in order to autonomously navigate, a robot must be able to build a representation of the environment as well as to recover its relative location. Vision-based estimation refers to the models and methods required to infer information from images that is useful to a given task. Throughout this thesis, computational efficiency is referred to as the ability of exploiting all possible visual information whilst satisfying real-time requirements. Moreover, the requirement of robustness is also essential so that accurate estimates can be obtained in spite of unavoidable modeling and measurement errors.

Thus, an important part of this thesis focuses on devising appropriate parametric models, as well as estimation methods for recovering the related parameters, such that those performance requirements are verified. We concentrate on image registration methods for obtaining the parameters. Image registration consists in estimating the transformations that best align a reference (fixed) image to a second (moving) one. Furthermore, a special emphasis is given here to direct methods of registration (Irani and Anandan, 1999). In this class of methods, the intensity value of the pixels are directly exploited to obtain the parameters, without having to first extract and match some image features (e.g. points, lines). In the computer vision community, direct methods are also called intensity-based, appearance-based, template-based, or even texture-based. We shall discuss in further detail the advantages and limitations of those techniques. Indeed, we seek simple, accurate, and generic models and methods for directly performing various vision-based tasks, such as visual tracking and visual Simultaneous Localization And Mapping (SLAM) of the environment, with respect to some coordinate system. Other applications comprise medical imagery (Maintz and Viergever, 1998) and augmented reality (Berger and Simon, 1998).

Another relevant part of this thesis focuses on vision-based control (Chaumette and Hutchinson, 2006). This branch of automatic control has received much attention for nearly two decades. Indeed, it also constitutes a fundamental research domain related to autonomous robot navigation, and is at the crossroad of many disciplines. Evidently, this includes automatic control and vision-based estimation. This thesis investigates ways to increase the flexibility and reliability of existing techniques by considering a framework of unknown objects under unknown imaging conditions. This will be shown for two different control approaches.

Thesis contributions. The major contributions of this thesis can be enumerated as follows:

1. A new model of illumination changes for direct image registration methods. For comparisons purposes, planar objects of various types under unknown imaging conditions are used in (Silveira and Malis, 2007c; Silveira and Malis, 2007d). However, it is shown here that it effectively deals with objects of unknown shape under unknown imaging conditions;

2. Generalization of the proposed model of illumination changes to any color image. This generic photometric model is able to fully couple all image channels so as to cope with unknown objects under unknown illumination conditions being imaged by cameras of unknown characteristics. Our robust and generic uncalibrated registration technique is built on this model;

3. A new formulation of the visual SLAM problem as a direct image registration task. A particular case of this approach is presented in (Silveira et al., 2007; Silveira et al., 2008c), whereas an extended version of these papers appeared in (Silveira et al., 2008a). This proposed formulation corresponds to our robust and efficient calibrated registration technique;

4. A new direct visual servoing technique which does not either require or estimate any metric knowledge about the object's shape and/or the camera's motion (Silveira and Malis, 2007a; Silveira and Malis, 2008). An extended version of these papers was published in (Silveira and Malis, 2007b). This corresponds to our robust and generic vision-based control technique where a reference image is given;

5. We show that a straightforward path planning is sufficient to ensure a very large domain of convergence for visual servoing tasks, if the proposed direct technique is applied. This property also appeared in (Silveira and Malis, 2007a; Silveira and Malis, 2008; Silveira and Malis, 2007b);

6. A new vision-based control scheme where neither the scene model nor the reference image are available a priori. A particular case of this scheme is described in (Silveira et al., 2006b). An extended version of this paper appeared in (Silveira et al., 2008b). It corresponds to our robust and efficient visual servoing technique where a reference pose is given;
7. A new detector of planar regions in a pair of uncalibrated images (Silveira et al., 2006a). This detector is a component of the proposed efficient vision-based control scheme where a reference pose is given.

Thesis organization. This thesis is organized as follows.

Part I is mainly devoted to a brief recall on the needed background. Basic models and useful methods are introduced here. Furthermore, the research problems tackled in this thesis are informally formulated in this first part. An important remark is that the state-of-the-art techniques related to these problems are not presented here. We have preferred to describe them throughout the thesis.

Part II presents the contributions related to direct estimation. It comprises the contributions from 1 to 3 above, and is articulated in three chapters. This part proposes a unified framework for directly registering images, either in the uncalibrated setting or in the calibrated one. Common models and methods to both settings are presented in the first chapter of this part. The specificities of each case are then taken into account in different chapters, together with the related state-of-the-art techniques.

Part III describes the contributions to direct control. It comprises the items from 4 to 7 above, and is articulated in two chapters. The first one tackles the vision-based control problem where the desired signals (i.e. to be reached) are given by a reference image. In the second chapter of this last part, the desired signals are directly defined in the Cartesian space.

General conclusions and directions for future work are then discussed, in terms of both estimation and control from visual data. The appendices especially include all theoretical demonstrations.

Every effort has been made to write this thesis concisely. Instead of providing an exhaustive compendium of all related concepts and existing formulations, only those we fell that are both used and essential to understanding the contributions are described. Carefully chosen references are naturally provided for further details.

Acknowledgments. I gratefully acknowledge the financial support from the Brazilian Foundation Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) under Grant 1886/03-7, the international agreement between Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)–Institut National de Recherche en Informatique et en Automatique (INRIA) under Grant 04/13467-5, and from the Centro de Tecnologia da Informação Renato Archer (CTI), former CenPRA.

Notations

Unless otherwise stated, scalars are denoted either in italics or in lowercase Greek letters, e.g. v and λ, vectors in lowercase bold fonts, e.g. $\mathbf{v}$, whereas matrices are represented in uppercase bold fonts, e.g. $\mathbf{V}$. Column vectors are adopted throughout this thesis. Row vectors are obtained with the transpose operation applied to them, e.g. $\mathbf{v}^\top$. Groups are written in uppercase double-struck (i.e. blackboard bold) fonts, e.g. the n-dimensional group of real numbers $\mathbb{R}^n$, and $\{\mathbf{v}_i\}_{i=1}^n$ corresponds to the set $\{v_1, v_2, \ldots, v_n\}$. Besides, $(\mathbf{V}^{-1})^\top = (\mathbf{V}^\top)^{-1}$ is abbreviated by $\mathbf{V}^{-\top}$, $\mathbf{0}$ (resp. $\mathbf{1}$) denotes a matrix of zeros (resp. ones) of appropriate dimensions, and $\mathbf{I}_n = \operatorname{diag}(\mathbf{1})$ represents the $(n \times n)$ identity matrix.

We also 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 $\mathbf{v}$. For an n-dimensional homogeneous vector $\mathbf{v}$, its $(n-1)$-dimensional non-homogeneous version is written $\underline{\mathbf{v}}$. Here, a superscripted asterisk, e.g. $\mathbf{v}^*$, or a subscripted r, e.g. $\mathbf{v}_r$, are interchangeably used to characterize a variable 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 $\mathbf{v}$, whilst $[\mathbf{v}]_\times$ denotes the skew symmetric matrix associated to vector $\mathbf{v}$.

Finally, the gradient operator applied to a vector-valued function $\mathbf{d}(\mathbf{v})$ with respect to the variable $\mathbf{v}$ is denoted $\nabla_{\mathbf{v}}\mathbf{d}(\mathbf{v})$, or simply $\nabla\mathbf{d}(\mathbf{v})$ if it is clear from the context. In line with standard notations, this matrix of first-order partial derivatives is also referred to as the Jacobian matrix $\mathbf{J}(\mathbf{v})$.

Part I

Introduction

Chapter 1

Basic geometric and photometric models

This chapter briefly recalls the basic concepts and models used in subsequent chapters. In particular, it introduces the representation of both the pose and the structure of rigid objects, their kinematics, as well as their interactions with cameras and illuminants for multiple image formation. Further details and theoretical proofs of the statements can be obtained in the referred bibliography.

1.1 Change-of-frame formulae

A frame $\mathcal{F}$ is a right-hand coordinate system centered at the origin $\mathcal{O}$ with the orthonormal basis $\{\vec{x}, \vec{y}, \vec{z}\}$ in the vector space $\mathbb{R}^3$. In the sequel, let us denote the reference (fixed) frame by either $\mathcal{F}_r$ or $\mathcal{F}^*$, and the current (moving) frame by either $\mathcal{F}_c$ or $\mathcal{F}$.

1.1.1 Coordinate transformations

Consider a 3D point $\underline{\mathbf{m}} = [x, y, z]^\top \in \mathbb{R}^3$ (in non-homogeneous coordinates). If this point is defined with respect to the reference frame, let it be represented by $\underline{\mathbf{m}}_r = [x_r, y_r, z_r]^\top$. If this same point $\underline{\mathbf{m}}$ is defined with respect to the current frame, let it also be denoted by $\underline{\mathbf{m}}_c = [x_c, y_c, z_c]^\top$.

The action of a rigid-body displacement on the coordinates of a point is given by

$$\underline{\mathbf{m}}_c = {}^c\mathbf{R}_r \, \underline{\mathbf{m}}_r + {}^c\mathbf{t}_r, \tag{1.1}$$

where ${}^c\mathbf{R}_r \in \mathbb{SO}(3)$ and ${}^c\mathbf{t}_r \in \mathbb{R}^3$ (or simply $\mathbf{R}$ and $\mathbf{t}$, if it is clear from the context) respectively denotes the rotation matrix and the translation vector between the origin of those two frames. Using homogeneous coordinates, this transformation of coordinates can be written in compact form as

$$\mathbf{m}_c = {}^c\mathbf{T}_r \, \mathbf{m}_r, \tag{1.2}$$

where $\mathbf{m}_c = [\underline{\mathbf{m}}_c, 1]^\top$, $\mathbf{m}_r = [\underline{\mathbf{m}}_r, 1]^\top$ and

$$ {}^c\mathbf{T}_r = \begin{bmatrix} {}^c\mathbf{R}_r & {}^c\mathbf{t}_r \\ \mathbf{0} & 1 \end{bmatrix} \quad \in \mathbb{SE}(3). \tag{1.3}$$

The Lie group $\mathbb{SE}(3)$ (i.e. the special Euclidean group) is homeomorphic to $\mathbb{SO}(3) \times \mathbb{R}^3$ and represents the group of rigid displacements (Warner, 1987; Varadarajan, 1974). Hence, the element ${}^c\mathbf{T}_r = {}^r\mathbf{T}_c^{-1}$ of this group encodes the pose (position and orientation) of a rigid body described by $\mathcal{F}_r$ in the basis of $\mathcal{F}_c$.

1.1.2 Velocity transformations

The tangent space of the Lie group $\mathbb{SE}(3)$ at the identity element is the Lie algebra $\mathfrak{se}(3)$ (Warner, 1987; Varadarajan, 1974; Hall, 2003). The coordinates $\mathbf{v}_c = [\boldsymbol{\nu}_c, \boldsymbol{\omega}_c]^\top \in \mathbb{R}^6$ of this tangent space correspond to the translational and rotational velocities, respectively. An element of this space can be written as the (4×4) matrix

$$\mathbf{A}(\mathbf{v}_c) = \begin{bmatrix} [\boldsymbol{\omega}_c]_\times & \boldsymbol{\nu}_c \\ \mathbf{0} & 0 \end{bmatrix} \quad \in \mathfrak{se}(3). \tag{1.4}$$

The velocity of the reference frame moving relative to the current frame ${}^c\dot{\mathbf{T}}_r$, and expressed in the coordinate system of the current frame, can then be obtained by deriving (1.3) as

$${}^c\dot{\mathbf{T}}_r = -\mathbf{A}(\mathbf{v}_c)\, {}^c\mathbf{T}_r. \tag{1.5}$$

Let us assume a constant velocity $\mathbf{v}_c$. The solution of this linear ordinary differential equation (1.5),

$${}^c\mathbf{T}_r(t) = \exp\big(-t\mathbf{A}(\mathbf{v}_c)\big)\, {}^c\mathbf{T}_r(0), \tag{1.6}$$

shows that both spaces $\mathfrak{se}(3)$ and $\mathbb{SE}(3)$ are related[1] through the exponential map

$$\begin{aligned} \exp\colon \mathfrak{se}(3) &\to \mathbb{SE}(3) \qquad (1.7)\\ \mathbf{A}(\mathbf{v}_c) &\mapsto \exp\big(\mathbf{A}(\mathbf{v}_c)\big). \qquad (1.8) \end{aligned}$$

The instantaneous velocity of a 3D point can be found by deriving (1.2) and using (1.5)

$$\dot{\mathbf{m}}_c = -\mathbf{A}(\mathbf{v}_c)\, \mathbf{m}_c. \tag{1.9}$$

The velocity of the reference frame moving relative to the current frame ${}^c\dot{\mathbf{T}}_r$ can also be expressed in another coordinate system. Let the (4×4) matrix

$$\mathbf{A}(\mathbf{v}_r) = \begin{bmatrix} [\boldsymbol{\omega}_r]_\times & \boldsymbol{\nu}_r \\ \mathbf{0} & 0 \end{bmatrix} \quad \in \mathfrak{se}(3) \tag{1.10}$$

[1] In fact, every Lie algebra is related to its Lie group through the exponential map, and not only these particular spaces.

with coordinates $\mathbf{v}_r = [\boldsymbol{\nu}_r, \boldsymbol{\omega}_r]^\top \in \mathbb{R}^6$ express that velocity in the reference frame. The velocity ${}^c\dot{\mathbf{T}}_r$ is then viewed in the reference frame as

$$ {}^c\dot{\mathbf{T}}_r = -{}^c\mathbf{T}_r\, \mathbf{A}(\mathbf{v}_r). \tag{1.11}$$

Manipulating Eqs. (1.5) and (1.11) provides the transformation between instantaneous velocities, commonly known as the adjoint map on the $\mathfrak{se}(3)$:

$$\begin{aligned} \mathrm{ad}_{\mathbf{T}} \colon \mathfrak{se}(3) &\to \mathfrak{se}(3) \qquad &(1.12)\\ \mathbf{A}(\mathbf{v}_c) &\mapsto \mathrm{ad}_{{}^r\mathbf{T}_c}\big(\mathbf{A}(\mathbf{v}_c)\big) = \mathbf{A}(\mathbf{v}_r) = {}^r\mathbf{T}_c\, \mathbf{A}(\mathbf{v}_c)\, {}^r\mathbf{T}_c^{-1}. \qquad &(1.13)\end{aligned}$$

The equation in (1.13) can be easily rewritten as

$$\begin{bmatrix} \boldsymbol{\nu}_r \\ \boldsymbol{\omega}_r \end{bmatrix} = \begin{bmatrix} {}^r\mathbf{R}_c & [{}^r\mathbf{t}_c]_\times {}^r\mathbf{R}_c \\ \mathbf{0} & {}^r\mathbf{R}_c \end{bmatrix} \begin{bmatrix} \boldsymbol{\nu}_c \\ \boldsymbol{\omega}_c \end{bmatrix}. \tag{1.14}$$

1.2 Image formation

The formation of an image depends on highly complex interactions amongst the camera, the scene and the illuminants. These interactions can be modeled both geometrically and photometrically. Models of interest are concisely described in this section.

1.2.1 Geometric modeling

This thesis deals with central cameras. Moreover, let us concentrate on the pinhole camera model and on rigid objects for simplicity (generic deformable objects will be encompassed in Chapter 4).

In this case, according to the Thales' theorem we have

$$ z\, \mathbf{m}' = \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix} \mathbf{m}, \tag{1.15}$$

where $\mathbf{m}' = [x', y', 1]^\top$ denotes normalized pixel coordinates. Homogeneous pixel coordinates $\mathbf{p} = [u, v, 1]^\top$ are obtained in the image (retinal) plane through

$$ \mathbf{p} = \mathbf{K}\, \mathbf{m}', \tag{1.16}$$

where

$$ \mathbf{K} = \begin{bmatrix} \alpha_u & \alpha_{uv} & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1.17}$$

gathers the camera's intrinsic parameters: the scale factors $\alpha_u, \alpha_v > 0$, the skew factor α_{uv}, and the principal point $\mathbf{p}_0 = [u_0, v_0, 1]^\top$. Numerous methods exist to estimate these parameters from images, see for example (Tsai, 1987; Zhang, 2000). Thus, by injecting (1.15) in (1.16) gives

$$ z_c\, \mathbf{p}_c = \mathbf{K} \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix} \mathbf{m}_c. \tag{1.18}$$

If this same 3D point is now defined with respect to the reference frame $\mathcal{F}_r$ instead of $\mathcal{F}_c$, then by injecting (1.2) in (1.18) we have

$$z_c\,\mathbf{p}_c = \mathbf{K}\left[{}^c\mathbf{R}_r \ {}^c\mathbf{t}_r\right]\mathbf{m}_r. \tag{1.19}$$

Since all projective entities are defined up to a non-zero scale factor, let us henceforth write the equations that define them using the symbol '$\propto$'. For example, Equation (1.19) is then rewritten as

$$\mathbf{p}_c \propto \mathbf{K}\left[{}^c\mathbf{R}_r \ {}^c\mathbf{t}_r\right]\mathbf{m}_r. \tag{1.20}$$

1.2.2 Photometric modeling

Let us now focus on modeling the formation of the intensity value of a pixel. According to major illumination models, both experimental (Blinn, 1977) and physically-based ones (Cook and Torrance, 1982), the intensity at a particular pixel $\mathbf{p} = [u, v, 1]^\top$ is due to specular, diffuse and ambient reflections:

$$\mathcal{I}(\mathbf{h}, \mathbf{p}) = \mathcal{I}_s(\mathbf{h}_s, \mathbf{p}) + \mathcal{I}_d(\mathbf{h}_d, \mathbf{p}) + \mathcal{I}_a(\mathbf{h}_a) \tag{1.21}$$

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

Particular case (Lambertian surfaces). These particular surfaces, also called ideal diffuse surfaces, do not change appearance depending on the viewing direction. The specular term is thus null: $\mathcal{I}_s(\mathbf{h}_s, \mathbf{p}) = 0, \ \forall \mathbf{p} \in \mathcal{I}$. Therefore, major illumination models describe these materials as

$$\mathcal{I}(\mathbf{h}', \mathbf{p}) = \mathcal{I}_d(\mathbf{h}_d, \mathbf{p}) + \mathcal{I}_a(\mathbf{h}_a), \tag{1.22}$$

with $\mathbf{h}' = \{\mathbf{h}_d, \mathbf{h}_a\}$.

1.3 Two-view geometry

This section presents the geometric relations between corresponding image points in a pair of images. For a more thorough treatment, the reader is referred to for example (Faugeras et al., 2001; Hartley and Zisserman, 2000; Ma et al., 2003). In the sequel, consider that the two images are acquired with the same intrinsics camera parameters. Also, let these two images be defined by the reference frame $\mathcal{F}^*$ and the current frame $\mathcal{F}$.

1.3.1 Uncalibrated camera

The uncalibrated case corresponds to the setting where the camera's intrinsic parameters are neither known a priori nor estimated on-line. All involved entities are directly defined in the projective space.

In this case, the generic relation between corresponding image points is given by

$$\mathbf{p} \propto \mathbf{G}\,\mathbf{p}^* + \rho^*\,\mathbf{e}, \tag{1.23}$$

where $\mathbf{G} \in \mathbb{SL}(3)$ (the Lie group $\mathbb{SL}(3)$ is the special linear group of (3×3) matrices having determinant one) is a homography relative to an arbitrary plane Π not going through $\mathcal{O}^*$, $\mathbf{e} \in \mathbb{R}^3$ denotes the epipole, and $\rho^* \in \mathbb{R}$ is the parallax (relative to Π) of the 3D point projected in the image as $\mathbf{p}^*$. This projective parallax also encodes the inverse of the depth of this 3D point. Amongst various possibilities, the projective homography $\mathbf{G}$ can be characterized relatively to the homography at infinity $\mathbf{G}_\infty \in \mathbb{SL}(3)$:

$$\mathbf{G} \propto \mathbf{G}_\infty + \mathbf{e}\,\mathbf{q}^{*\top}, \tag{1.24}$$

where in this case the 3-vector $\mathbf{q}^*$ is a representation of the image of the line at infinity of Π.

An useful relation can be derived from (1.23) by algebraic manipulation, though degenerate configurations exist. Multiplying both sides of (1.23) on the left by $\mathbf{p}^\top[\mathbf{e}]_\times$ yields the Luong-Faugeras constraint

$$0 = \mathbf{p}^\top[\mathbf{e}]_\times\mathbf{G}\,\mathbf{p}^* = \mathbf{p}^\top\mathbf{F}\,\mathbf{p}^*, \tag{1.25}$$

where $\mathbf{F} \propto [\mathbf{e}]_\times\mathbf{G}$ is the so-called Fundamental matrix. It can be noted that $\mathbf{F}$ is obtained for any planar homography, not only $\mathbf{G}_\infty$, since (1.24) is only a possible characterization of $\mathbf{G}$. A degenerate configuration arises when at least two distinct (i.e. linearly independent $\mathbf{F}$) satisfy (1.25), e.g. when the imaged object corresponds to a critical surface.

Particular case (Simplified relation). Two special cases are of particular importance to subsequent discussions within this thesis. Both cases lead to a simplified relation of (1.23). The first one corresponds to a critical surface: the planar case, which provides $\rho^* = 0$. The other special case concerns a particular displacement between the two views: a pure rotation motion, which provides $\mathbf{e} = \mathbf{0}$. In any of these cases, the generic relation (1.23) between corresponding points is fully defined by a homography:

$$\mathbf{p} \propto \mathbf{G}\,\mathbf{p}^*. \tag{1.26}$$

1.3.2 Calibrated camera

The calibrated case is referred here to the setting where the involved entities are defined in the Euclidean space. This requires the knowledge of the camera's intrinsic parameters $\mathbf{K}$.

The generic relation between corresponding image points in calibrated images can be obtained by writing (1.18) for $\mathbf{p}^*$ and plugging the result into (1.20). Dividing this last outcome by $z^* > 0$ yields:

$$\mathbf{p} \propto \mathbf{K}\,\mathbf{R}\,\mathbf{K}^{-1}\mathbf{p}^* + (z^*)^{-1}\,\mathbf{K}\,\mathbf{t}, \tag{1.27}$$

where the Euclidean parallax is directly the inverse of the depth $(z^*)^{-1}$ and

$$\mathbf{K}\,\mathbf{t} \propto \mathbf{e}. \tag{1.28}$$

As for the uncalibrated case, an useful relation can be derived from (1.27) by algebraic manipulation, though degenerate configurations still exist. Multiplying both sides of (1.27) on the left by $\mathbf{m'}^\top[\mathbf{t}]_\times\mathbf{K}^{-1}$ (which is proportional to $\mathbf{p}^\top[\mathbf{e}]_\times$) and using (1.16), the Longuet-Higgins constraint is obtained

$$0 = \mathbf{m'}^\top[\mathbf{t}]_\times\mathbf{R}\,\mathbf{m'}^* = \mathbf{m'}^\top\mathbf{E}\,\mathbf{m'}^*, \tag{1.29}$$

where $\mathbf{E} = [\mathbf{t}]_\times\mathbf{R}$ is the so-called Essential matrix.

Particular case (Simplified relation). The same special cases to the uncalibrated setting are also relevant to the part of this thesis related to the calibrated domain. The first particular case of interest occurs when the imaged object is planar. Describing this plane by its unit normal vector $\mathbf{n}^*$ and its signed distance $-d^*$, a 3D point $\underline{\mathbf{m}}^*$ lying on this plane verifies $\mathbf{n}^{*\top}\underline{\mathbf{m}}^* = d^*$. The Euclidean parallax can be written for this particular case as

$$(z^*)^{-1} = (d^*)^{-1}\mathbf{n}^{*\top}\,\mathbf{K}^{-1}\mathbf{p}^*, \tag{1.30}$$

using (1.15) along with (1.16). By injecting (1.30) in (1.27), Equation (1.26) still holds with

$$\mathbf{G} \propto \mathbf{K}\left(\mathbf{R} + (d^*)^{-1}\,\mathbf{t}\,\mathbf{n}^{*\top}\right)\mathbf{K}^{-1}. \tag{1.31}$$

Another particular case of interest is that of a pure rotation motion. Since in this case $\mathbf{t} = \mathbf{0}$, then Eq. (1.26) also holds using (1.27) but with

$$\mathbf{G} \propto \mathbf{K}\,\mathbf{R}\,\mathbf{K}^{-1} \propto \mathbf{G}_\infty. \tag{1.32}$$

As a matter of fact, the homography at infinity $\mathbf{G}_\infty \in \mathbb{SL}(3)$ establishes the duality between distant scenes (i.e. $z^* \to \infty$) and pure rotation motions (i.e. $\mathbf{t} = \mathbf{0}$): Equation (1.26) holds with (1.32) for any scene structure if the ratio $\|\mathbf{t}\|/z^* \to 0$ (or equivalently $\|\mathbf{t}\|/d^* \to 0$, for the planar case).

Chapter 2

Problems statement

This chapter aims to informally formulate the main research problems tackled in this thesis, i.e. vision-based estimation and control. It also presents the major approaches to estimation from visual information, as well as some design considerations to vision-based control. The class of direct approaches to these problems will be formally discussed in next chapters, including the related state-of-the-art techniques.

2.1 Parametric estimation

Vision sensors can provide an enormous quantity of information: the geometric and photometric properties of the scene, of the illuminants, and of the camera, both intrinsic and extrinsic parameters. We remark that photometric properties also include their spectral response characteristics. All those properties can be defined by a set of parameters. Thus, parametric estimation from visual data aims at finding the best set of parameters that fits a particular model, given a set of images. This model is devised so as to appropriately describe that amount of information, whereas the best set is defined relative to a given similarity measure (more generically, to a given cost function). Then, as with any proposed method for solving a given problem, one should analyze its properties such as robustness and observability. This is important in order to establish the working conditions of an algorithm.

Therefore, a first step consists in defining a suitable parametric model to the task at hand. The searched parameters to a given task may comprise only a subset of the above-mentioned ones, given prior knowledge or assumptions of the others. In addition, they may be encoded in a few entities combining many of the parameters. In all case, the designed model should be simple yet accurate enough to attain a given objective. It can be noted that simplicity is fundamental to real-time systems, e.g. visually-servoed systems. This is not only due to the reduced computational cost in finding the parameters, but also to avoid spurious local minima. Local minima frequently arise since vision-based estimation methods generally involve a model described by non-linear equations.

Another issue consists in solving the related data association problem. The outcome from this subtask is a set of one-to-one correspondences between 3D points (defined in some coordinate system) and their pixel intensities or, similarly, a set of one-to-one correspondences of their projections in different images.

Then, from the underlying system of non-linear equations in the unknowns, these parameters should be estimated using efficient and robust numerical methods. Efficiency is important to satisfy real-time requirements, whilst robustness is essential so that accurate estimates can be obtained in spite of unavoidable modeling and measurement errors.

Existing approaches to performing this overall parametric estimation, given a suitable model and a set of images, can be classified into two major classes: feature-based and direct methods. They are both briefly described below.

2.1.1 Feature-based methods

In this class of methods of parametric estimation, the data association problem is separated from the resolution of the system of non-linear equations. This subsection discusses the use of image features, i.e. geometric primitives such as points, lines, contours, so as to perform the estimation.

Within feature-based methods, the parametric estimation process is divided into three main steps, as follows:

1. the data extraction, i.e. feature detection;
2. the data association, i.e. feature matching;
3. the parameter estimation, i.e. seek of the parameters that optimally and robustly explain that association, given a model.

If a dense mapping of the scene is desired, then other post-processing steps are necessary. This strategy is appealing because a difficult problem is broken down into smaller, potentially solvable ones. However, some considerations are imperative.

First of all, the feature detector should ideally extract the same features in all images, if they are visible. To achieve this, the ideal detector must be fully invariant to all possible changes in all those geometric and photometric parameters, as well as be robust to sensor noise. Obviously, this detector does not exist. Thus, two measures are commonly adopted to evaluate their performance: accuracy and repeatability. Accuracy is important because the error committed in the extraction process will never be corrected in the subsequent steps. Pre-processing steps performed on the images (e.g. smoothing) to achieve the ideal detector affect this measure. Repeatability is also central and reveals the degree of invariance in detecting the same feature when varying the imaging conditions. Since fully invariance is not possible, detectors in general rely on a threshold to decide whether a feature is present or not in an image.

Furthermore, data association is usually performed through evaluating a similarity measure between some descriptors of each feature. The cost of exhaustively comparing features is prohibitively high to be considered in a real-time setting. In that case, feature descriptors are not fully invariant to all those geometric and photometric parameters. Generally, they can be made robust only up to affine image transformations, including affine illuminations changes. This type of feature has been proposed in (Lowe, 2004; Tuytelaars and Van Gool, 2004). Moreover, the similarity measure cannot tolerate gross errors on the descriptors.

Finally, a great care must be taken in the last step of the estimation process since the data association procedure is highly error-prone. Indeed, attempts to eliminate the mismatched features are usually performed by enforcing some geometric constraints within a robust estimation technique, such as RANSAC (Fischler and Bolles, 1981) and M-estimators (Huber, 1981). As a closing remark, that enforcement is hence made after establishing the full set of correspondences, i.e. a posteriori.

Nevertheless, under the assumption that the images are different only by affine transformations, or a sufficiently small departure from this class (e.g. small perspective deformations), feature-based methods possess a relatively large domain of convergence. Indeed, the vast majority of existing solutions rely on these methods. Numerous successful applications are available in the literature, such as localization of the camera, and structure-from-motion techniques. See standard textbooks, for example (Faugeras et al., 2001; Hartley and Zisserman, 2000).

2.1.2 Direct methods

First of all, in this class of methods of parametric estimation there is no step of feature extraction. These methods are so called because the intensity value of the pixels is directly exploited to recover the related parameters (Irani and Anandan, 1999; Stein and Shashua, 2000). Another important characteristic is that they simultaneously solve the data association and the parameter estimation problems, given the parametric model.

Indeed, these estimation methods are usually formulated as a single nonlinear optimization problem. Given a model to deform (i.e. transform, generate) images and an initial estimate of the related parameters, the optimal ones are said to be found when the data association is said to be obtained, according to a similarity measure. We can then identify the following iterative steps (performed until convergence) within the resolution of this optimization problem:

1. the transformation of the image, given the model and current parameters;

2. the computation of the similarity measure, i.e. the cost function;

3. the computation of the increment on the parameters that decreases the cost function.

In this way, there are no post-processing steps to obtain a dense mapping of the scene, since the entire image can be exploited. As a matter of fact, this constitutes an important strength of these methods: all possible image information can be used, even from areas where gradient information is weak and no distinctive feature exists. Another strength concerns the simultaneous enforcement of structural constraints within the procedure, instead of a posteriori as in feature-based methods. Therefore, more accurate algorithms can be devised.

However, as it can be noted, an initial estimate of the parameters sufficiently close to the true ones is needed. The relatively smaller domain of convergence represents one of the main limitations of direct methods of estimation.

We shall formally and thoroughly present this framework in next chapters (in Part II), since this thesis makes contributions to this class of parametric estimation only.

2.2 Robot control

This section informally discusses how the estimated parameters from visual data can be effectively used in feedback control loops. No distinctions will be made here concerning the estimation method, which can be performed using either the previously described feature-based or direct methods.

2.2.1 Design of vision-based control schemes

The design of vision-based control schemes is primarily dependent on the task at hand. For example, the desired signal to be stabilized within a given task can be a reference pose or a reference velocity. Without loss of generality, let us focus in this thesis on the former, which is widely referred to as a positioning task in the robotics community.

Naturally, the design of vision-based control schemes is also in function of the available prior knowledge of the overall system. Considering that positioning task, the reference pose can indeed be provided:

- in the Cartesian space;
- in the sensor space, i.e. by means of a reference image.

Each one has its own set of advantages and drawbacks (Chaumette and Hutchinson, 2006).

In particular, if the reference pose is directly defined in the Cartesian space, then standard feedback control approaches can be applied. The problem becomes that of controlling the pose reconstructed from images (i.e. localization of the camera). On the other hand, the main limitation in this case is that any measurement or modeling error may produce a discrepancy in the final pose with respect to the desired one. Furthermore, since no control in the image is performed, the issue of object visibility becomes crucial to the stability of the system.

The second case of defining the reference pose by means of a reference image corresponds to a strategy well-known in the visual servoing community as "teach-by-showing" (Weiss and Anderson, 1987). The idea behind this scheme is to construct a control error whose space is diffeomorphic to the Cartesian space. This case is motivated by an increase in robustness with respect to modeling errors. However, the design of this control scheme can be tricky. Indeed, many existing strategies start by assuming that the imaging conditions (e.g. illumination, camera's internal parameters) do not vary between the time of acquiring the reference image and whilst executing the positioning task. Furthermore, the convergence properties of the overall system can dramatically vary from a control scheme to another. The design of the most appropriate control error and control law is not straightforward in many situations where the teach-by-showing strategy is applied.

In both cases, important design challenges regard to building efficient and accurate solutions. It can be noted that in order to achieve accuracy, no matter the control scheme, it should take into consideration all available sensory information, should enforce all structural constraints, and should have a domain of convergence as large as possible. If system flexibility is also sought, then one should not rely on prior knowledge.

We shall formally and thoroughly present our contributions to those both cases in future chapters (in Part III).

2.2.2 Design of vision-based control laws

The design of vision-based control laws is greatly dependent on the robot type. Several classifications of robots are possible depending on the involved variables. Let us classify them here in function of the stabilizability properties of the linear systems which approximate them around equilibrium points. From this perspective, we can classify most of them as (Morin, 2004):

- non-critical non-linear systems;
- critical non-linear systems.

An autonomous non-linear system is called critical when the corresponding linearized systems are not asymptotically stabilizable. Whereas local stabilizers for non-critical systems can often be derived from their linear approximations, one has to rely on truly non-linear methods in the case of critical systems.

The contributions of this thesis to vision-based control consider a robot of the first type. In particular, let us assume that the camera-mounted system is asymptotically stabilizable in all six degrees-of-freedom. This case corresponds, for instance, to a camera-mounted classical manipulator robot. Vision-based control of critical non-linear systems, for example underactuated or non-holonomic robots, constitutes a subject for future research. For example, the vision-based control of airships (Bueno et al., 2002) for stationary flights or of ground mobile robots (Maya-Mendez et al., 2006).

In all cases, the design challenge is to build control laws that are robust to measurement and modeling errors, and that ensure stability (in some sense) of the robotic system. The latter issue must be addressed taking into consideration the specificity of the robot type. Furthermore, if system flexibility is desired, then one should not rely on prior knowledge.

Part II

Direct estimation from visual data

Chapter 3

Direct image registration

Image registration (also called image alignment) refers to the process of estimating the appropriate parameters that optimally overlays two or more images of the same scene, taken at different imaging conditions (e.g. illumination, motion) and possibly by different imaging modalities (Maintz and Viergever, 1998). Direct methods refer to those that exploit the intensity value of the pixels in order to recover the related parameters (Irani and Anandan, 1999; Szeliski, 2005).

This chapter presents the proposed models and methods to directly register images, which are common to all subsequent chapters. As we will demonstrate throughout this thesis, they represent an important tool for a wide spectrum of applications. The specificities of each case are taken into account in different chapters, together with the related state-of-the-art techniques.

3.1 Central issues

Let $\mathcal{I}^* \subset \mathbb{R}^2$ represent a reference image of an unknown scene. Strictly speaking, $\mathcal{I}^*$ usually corresponds to a region (also called template) within the entire captured image. After changing the imaging conditions, another image $\mathcal{I}$ of the same scene is acquired. In line with standard conventions (although with abuse of notation), let the intensity value of a particular pixel $\mathbf{p} \in \mathbb{P}^2$ be denoted by $\mathcal{I}(\mathbf{p}) \geq 0$.

The problem of direct image registration consists in searching for the best set of parameters $\mathbf{x}$ (of a given model) to transform the current $\mathcal{I}$ such that all its intensity values match as closely as possible to the corresponding ones in the reference $\mathcal{I}^*$. More formally, a typical direct image registration system solves non-linear optimization problems of the type

$$\min_{\mathbf{x}} \; \frac{1}{2} \sum_i \Big[\underbrace{\mathcal{I}'(\mathbf{x}, \mathbf{p}_i^*) - \mathcal{I}^*(\mathbf{p}_i^*)}_{d_i(\mathbf{x})} \Big]^2, \tag{3.1}$$

where $\mathcal{I}'$ denotes the transformed image from $\mathcal{I}$ and is used to compute the set of intensity discrepancies $\mathbf{d}(\mathbf{x}) = \{d_i(\mathbf{x})\}$. See Fig. 3.1 for an illustrative example.

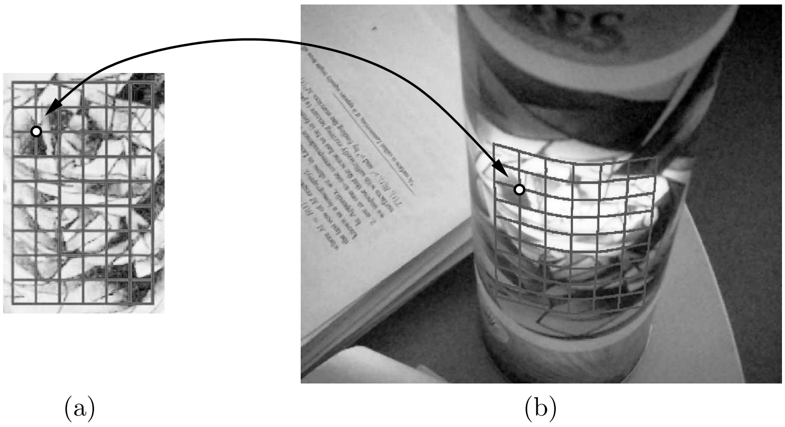


Figure 3.1. (a) Reference image (also called reference template) superimposed by a grid. (b) Current image superimposed by the aligned grid. Image registration consists in estimating the appropriate parameters to optimally align all pixels within a reference template to another image of the same object, taken at different imaging conditions.

Therefore, two central issues can be identified immediately: the choice of the appropriate parametric transformation model and of the optimization method. Both of them are discussed in the sequel. Of course, the cost function can also be a design parameter, but the sum-of-square-differences in (3.1) is the most widely used one for registering images of the same modality without aberrant measures. Let us focus here 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 included in (3.1). Finally, the initialization issue of such an estimation system is briefly discussed at the end of this chapter.

3.2 Parametric transformation models

As discussed in past chapters, appropriate transformation models have to be chosen to accomplish a particular task. They may comprise both geometric and photometric models of the involved interactions between the scene, the sensor and the illuminants.

3.2.1 Generic warping model

The warping model describes a transformation between pixel coordinates. Hence, it encodes geometric variations between views. More formally,

$$\mathbf{w}\colon G \times \mathbb{P}^2 \to \mathbb{P}^2 \tag{3.2}$$

$$(\mathbf{g}, \mathbf{p}^*) \mapsto \mathbf{p} = \mathbf{w}(\mathbf{g}, \mathbf{p}^*) \tag{3.3}$$

where G is a Lie group, and $\mathbf{g} \in G$ encodes a geometric description of the scene structure, of the camera itself and of its motion. The geometric modeling (3.3) is generic in the sense that its description is independent on the space of the parameters. Depending on the specific task, this set of parameters can be defined either in the projective or in the Euclidean space. The former case is treated in Chapter 4, whereas the latter case is applied in Chapter 5.

Particular case (Planar surface). A very simple example consists in considering a planar object imaged by an uncalibrated pinhole camera. In this case, $\mathbf{g} = \mathbf{G} \in \mathbb{SL}(3)$ using (1.26) and therefore, Equation (3.3) above can be explicitly written as

$$\begin{aligned} \mathbf{p} &= \mathbf{w}(\mathbf{G}, \mathbf{p}^*) && (3.4) \\ &= \left[\frac{g_{11}u^* + g_{12}v^* + g_{13}}{g_{31}u^* + g_{32}v^* + g_{33}}, \frac{g_{21}u^* + g_{22}v^* + g_{23}}{g_{31}u^* + g_{32}v^* + g_{33}}, 1 \right]^\top, && (3.5) \end{aligned}$$

where $\{g_{ij}\}$ denotes the elements of the homography $\mathbf{G}$. In this particular case, the warping (3.4) is a group action of $\mathbb{SL}(3)$ on $\mathbb{P}^2$, i.e.

$$\mathbf{w}(\mathbf{G}_1\mathbf{G}_2, \mathbf{p}^*) = \mathbf{w}\big(\mathbf{G}_1, \mathbf{w}(\mathbf{G}_2, \mathbf{p}^*)\big), \quad \forall \mathbf{G}_1, \mathbf{G}_2 \in \mathbb{SL}(3). \qquad (3.6)$$

This property has been exploited in (Benhimane and Malis, 2004) for efficiently registering images of planar objects under the brightness constancy assumption.

Remark 3.1. Independently of the considered space for $\mathbf{g}$, the resulting pixel coordinates $\mathbf{p} \in \mathbb{P}^2$ in (3.3) can in fact be a non-integer number. Since direct image registration exploits the intensity value $\mathcal{I}(\mathbf{p})$ of $\mathbf{p}$, and the image is a discretized observation of the scene, a suitable interpolation method (e.g. bilinear, bicubic) has to be used to find the required intensities.

Remark 3.2. It can be noted that a typical registration system (3.1) warps the coordinates of reference pixels $\mathbf{p}^*$ into the current $\mathbf{p}$ by using (3.3) (and interpolates the result using $\mathcal{I}$) in order to obtain the image $\mathcal{I}'$. In this way, the geometric parameters $\mathbf{g} \in G$ allow for a mapping from the reference frame to the current frame, in accordance to the conventions used in Chapter 1.

3.2.2 Generic photometric model

For visual tracking purposes, the photometric modeling aims at explaining the lighting changes between two views. In other words, it concerns the recovery of which lighting variations has to be applied to the current image $\mathcal{I}$ (1.21) in order to obtain an image $\mathcal{I}'$ whose illumination conditions are as closely as possible to those at the time of acquiring $\mathcal{I}^*$. This transformation model is written as

$$\mathcal{I}'(\boldsymbol{\alpha}_s, \boldsymbol{\alpha}_d, \beta, \mathbf{p}) = \boldsymbol{\alpha}_s(\mathbf{p})\mathcal{I}(\mathbf{p}) + \boldsymbol{\alpha}_d(\mathbf{p})\mathcal{I}(\mathbf{p}) + \beta, \qquad (3.7)$$

where $\boldsymbol{\alpha}_s(\mathbf{p}), \boldsymbol{\alpha}_d(\mathbf{p}), \beta \in \mathbb{R}$ capture 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 variations depend on the albedos of each point on the surface, as well as its shape, the camera parameters and other imaging conditions.

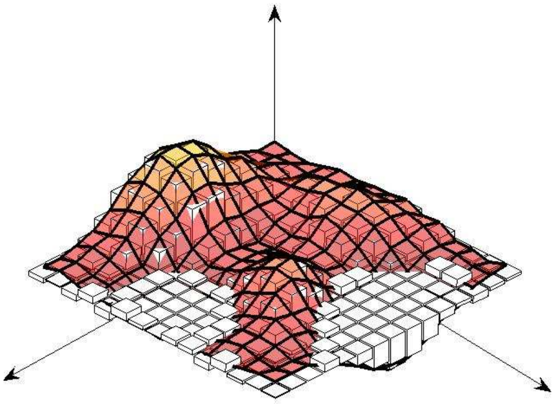


Figure 3.2. The illumination changes are viewed as an evolving three-dimensional surface $\mathcal{S}$ (colored). Thus, local lighting variations are also captured by this model.

This is then a difficult, computationally intensive problem where many images and priors are required to consistently recover those parameters. Indeed, two assumptions are commonly adopted by direct image registration algorithms, e.g. (Baker et al., 2003). The first assumption is to consider that the surface is perfectly Lambertian so that $\boldsymbol{\alpha}_s(\mathbf{p}) = 0, \forall \mathbf{p} \in \mathcal{I}$. Secondly, they assume that the entire surface holds exactly the same reflectance properties so that $\forall \mathbf{p} \in \mathcal{I}$, $\boldsymbol{\alpha}_d(\mathbf{p})$ is a constant. Although suited to some applications, both assumptions are obviously violated in many cases.

Since we do not make assumptions about either the imaging conditions or the materials, we develop a new model of illumination changes. Instead of using (3.7), we seek an elementwise multiplicative lighting variation $\mathcal{S}$ over the current $\mathcal{I}$, and a global $\beta \in \mathbb{R}$, such that $\mathcal{I}'$ matches as closely as possible to $\mathcal{I}^*$. That is, we propose the following generic (in the case of gray-scale images) photometric model:

$$\mathcal{I}' = \mathcal{S} \cdot \mathcal{I} + \beta, \tag{3.8}$$

where the dot operator '$\cdot$' denotes here the elementwise multiplication. Hence, the lighting variation $\mathcal{S}$ is viewed as a surface that evolves with time. Notice that, whilst the offset β captures only global variations, the surface $\mathcal{S}$ also models local illumination changes (e.g. produced by specular reflections). See Fig. 3.2. Very importantly, this model allows the registration to be performed without prior knowledge of either the object's attributes (e.g. albedos, shape) or the characteristics of the illuminants (e.g. number, power, pose).

The model (3.8) is also different from the one presented in (Negahdaripour, 1998), where the offset is also as a function of the pixels. This existing model is over-parametrized, but is shown in that work to give satisfactory results in the case of optical flow computation. This computation is not our primary objective, though registration methods also recover the optical flow simultaneously. A strategy to reduce the problems related to that over-parametrized model (e.g. convergence issues) is presented in (Lai and Fang, 1999).

Particular case (Affine model). It is easy to verify that the affine case corresponds to a particular model from the generic photometric one (3.8). In this case, the surface is described by a simple constant:

$$\mathcal{S} = \gamma \mathbf{1}, \tag{3.9}$$

with $\gamma \in \mathbb{R}$ and $\dim(\mathbf{1}) = \dim(\mathcal{I})$. This model is appropriate if those previously mentioned prior knowledge of the imaging conditions and the object is available.

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

$$\mathcal{S} \approx f_h(\boldsymbol{\gamma}, \mathbf{p}), \quad \forall \mathbf{p} \in \mathcal{I}, \tag{3.10}$$

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

A widely used technique to regularize a surface is via Radial Basis Functions (RBF) (Carr et al., 1997). In this case, $f_h : \mathbb{R}^{q+3} \times \mathbb{P}^2 \to \mathbb{R}$ is approximated using, for example, the thin-plate spline $\varphi(x) = x^2 \log(x), \ \forall x \in \mathbb{R}_+$, along with a first-degree polynomial:

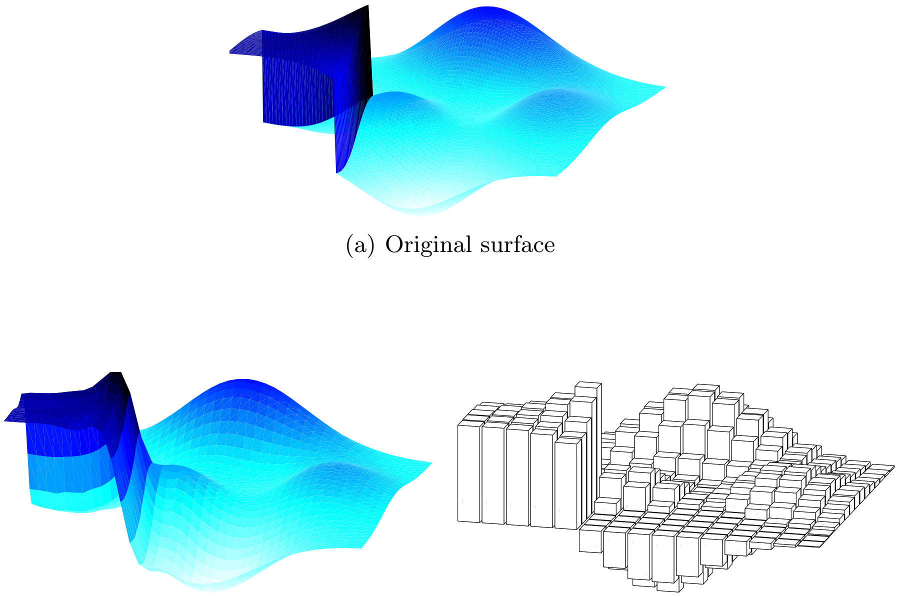
$$f_h(\boldsymbol{\gamma}, \mathbf{p}) \; = \; [\gamma_{q+1}, \, \gamma_{q+2}, \, \gamma_{q+3}]^\top \mathbf{p} \; + \; \sum_{i=1}^{q} \gamma_i \, \varphi(\|\mathbf{p} - \mathbf{q}_i\|), \tag{3.11}$$

where $\left\{\mathbf{q}_i \in \mathbb{P}^2\right\}_{i=1}^{q}$ are the image points (also called centers) that can be selected, for example, on a regular grid or correspond to interest points of the image. The side conditions can be easily imposed by solving a linear system whilst the interpolation conditions is indirectly imposed by minimizing a similarity measure (e.g. the sum-of-square-differences). The use of RBFs allows to regularize the surface but they may fail to accurately capture discontinuities, since the function (3.11) has a global support.

A suitable strategy for dealing with discontinuous surfaces is to discretize it into q sufficiently small $(\Delta u \times \Delta v)$ regions such that

$$\iint_{\mathcal{I}} \mathcal{S}(\mathbf{p}) \; du \, dv \; \approx \; \sum_{i=1}^{q} \; f_h(\boldsymbol{\gamma}_i, \mathbf{p}) \; \Delta u \, \Delta v. \tag{3.12}$$

This discretization leads to a computationally efficient solution since a linear system is obtained, along with the fact that the parameters are estimated independently of each other. Nevertheless, the appropriateness of a particular



(a) Original surface

(b) Approx. by a RBF

(c) Approx. by discretizing

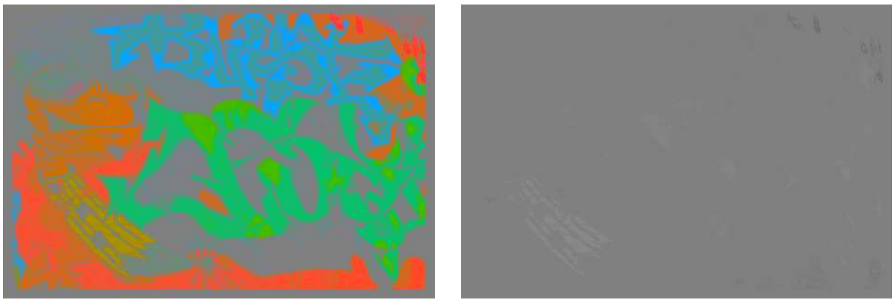
Figure 3.3. Some possibilities to approximate a surface. (b) Radial Basis Functions (RBF) regularize it but do not capture discontinuities. (c) Discretization deals with discontinuities and yields a computationally efficient system, but ignores smoothness.

approximation depends on various factors, such as the assumptions concerning the object and the required system's performance (compromise between computational efficiency, accuracy and robustness). See Fig. 3.3 for illustrative examples.

Particular case (Saturations due to highlights and shadows). These particular effects are here interpreted as well-structured types of occluders. This characterization is well-justified since, wherever they are present, all information which are useful for registration purposes are hidden. Moreover, they are well-structured because a saturation pattern is exhibited either to the highest or to the lowest intensity levels. Therefore, they can be filtered suitably: one only needs to check whether or not those homogeneous patterns appear in each warped image region.

3.2.3 Generalization to any color image

Color images can be of particular importance in many scenarios. As a matter of fact, extreme cases exist where all information is completely lost if the objects are observed with gray-scale cameras (see Fig. 3.4). Even if this is an unlikely situation in practice, we can make a conjecture that in many situations color cameras provide much richer information than their gray-scale counterparts.



(a)

(b)

Figure 3.4. (a) Original color image and (b) after its conversion to gray-scale. Almost all information has been lost in this example. Please print in color so as to see how rich the original image is!

Color cameras, like the human eye, are generally (but not always) trichromatic. In this case each pixel of a color image is a three-vector, one component per sensor channel. An active research topic concerns color constancy, 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] the results presented in (Finlayson et al., 1994) claim that a linear transformation matrix is sufficient to support color constancy in practice. This framework has been exploited in color-based point tracking e.g. (Montesinos et al., 1999; Gouiffès et al., 2006), and in color image registration (Bartoli, 2006).

This subsection describes a photometric model to overcome the limitations of both the Mondrian world[1] and of those working conditions,[2] whilst naturally encompassing the gray-level case. In other words, it describes how to extend the photometric model presented in Subsection 3.2.2 to the case of color images. Furthermore, the extension will be made for any color image.

Let us denote a color image by $\boldsymbol{\mathcal{I}}$, obtained by stacking the channels $\mathcal{I}_k$, $k = 1, 2, \ldots, n$. We propose to obtain the transformed color image $\boldsymbol{\mathcal{I}}'$ that best matches the reference $\boldsymbol{\mathcal{I}}^*$ through the model

$$\boldsymbol{\mathcal{I}}' = \boldsymbol{\mathcal{S}} \bullet \boldsymbol{\mathcal{I}} + \boldsymbol{\beta}, \tag{3.13}$$

where

$$\boldsymbol{\mathcal{S}} = \begin{bmatrix} \mathcal{S}_{11} & \mathcal{S}_{12} & \cdots & \mathcal{S}_{1n} \\ \mathcal{S}_{21} & \mathcal{S}_{22} & \cdots & \mathcal{S}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{S}_{n1} & \mathcal{S}_{n2} & \cdots & \mathcal{S}_{nn} \end{bmatrix} \tag{3.14}$$

comprises the surfaces related to the illumination changes, and $\boldsymbol{\beta} \in \mathbb{R}^n$ captures the per-channel shift both in the ambient lighting changes and in the camera bias. The operator '$\bullet$' stands for a linear combination of the color channels,

[1] A Mondrian is a planar surface composed of only 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 inter-reflections, etc.

elementwise multiplied by the corresponding surface. That is, Equation (3.13) can be rewritten using the operator for elementwise multiplication '·' as

$$\mathcal{I}'_k = \sum_{j=1}^{n} \mathcal{S}_{kj} \cdot \mathcal{I}_j + \beta_k, \quad k = 1, 2, \ldots, n. \tag{3.15}$$

The proposed fully coupling photometric model (3.13) 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 generic model if they are available. For example, prior knowledge concerning the spectral response of the camera sensors (e.g. from its data-sheet) allows for suitably uncoupling the lighting variation $\boldsymbol{\mathcal{S}}$, at an eventual expense of robustness. This particular case is described below.

Particular case (Known spectral characteristics). If the color camera's data-sheet specifies that the n sensors are narrow-band, then a fully uncoupled model can be used by adopting

$$\boldsymbol{\mathcal{S}} \approx \operatorname{diag}\big(\mathcal{S}_{11}, \mathcal{S}_{22}, \ldots, \mathcal{S}_{nn}\big). \tag{3.16}$$

If only some of them are narrow-band, it is possible to devise other particular models from the generic one (3.13) 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 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 $\mathcal{S}_{12} = \mathcal{S}_{21}$ and/or $\mathcal{S}_{23} = \mathcal{S}_{32}$.

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

$$\boldsymbol{\mathcal{S}} \approx \boldsymbol{f}_h(\boldsymbol{\Gamma}, \mathbf{p}), \quad \forall \mathbf{p} \in \boldsymbol{\mathcal{I}}, \tag{3.17}$$

and where $\boldsymbol{\Gamma} = \{\gamma_{kj}\}$. One then has to choose an appropriate finite-dimensional approximation $\boldsymbol{f}_h(\boldsymbol{\Gamma}, \mathbf{p})$ of the actual $\boldsymbol{\mathcal{S}}$. Next subsection describes an efficient optimization procedure to estimate all related parameters.

Particular case (Saturations due to highlights and shadows). Similarly to the gray-scale 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 Efficient optimization procedure

Let us turn back to the typical non-linear optimization problem expressed in (3.1). Given the real-time requirements of robotic applications, only minimization methods that have limited convergence domain can be applied. Global optimization methods such as Simulated Annealing (Horst and Pardalos, 1995) are too computationally intensive to be considered in a real-time setting. Thus, the objective here is to design algorithms that are both computationally efficient and have a large domain of convergence. Besides the method itself, another fundamental aspect to achieve these properties concerns the parametrization $\mathbf{z}$ of the involved variables $\mathbf{x}$, i.e. $\mathbf{x} = \mathbf{x}(\mathbf{z})$. This latter issue will be discussed in next chapters. In the sequel, consider that the underlying functions are (at least piecewise) smooth so that they can be expanded in Taylor series.

Hence, given the iterative nature of most existing efficient methods, Equation (3.1) has firstly to be changed into

$$\min_{\widetilde{\mathbf{z}}} \ \frac{1}{2}\sum_i \Big[\underbrace{\boldsymbol{\mathcal{I}}'\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}, \mathbf{p}_i^*\big) - \boldsymbol{\mathcal{I}}^*(\mathbf{p}_i^*)}_{d_i(\mathbf{x}(\widetilde{\mathbf{z}})\circ\widehat{\mathbf{x}})}\Big]^2, \tag{3.18}$$

where an initial estimate $\widehat{\mathbf{x}}$ sufficiently close to the true parameters $\overline{\mathbf{x}}$ is needed. In this case, the optimization problem is solved by iteratively finding an incremental displacement $\widetilde{\mathbf{x}} = \mathbf{x}(\widetilde{\mathbf{z}}_k)$ in order to generate a sequence of values

$$\widehat{\mathbf{x}}_{k+1} = \mathbf{x}(\widetilde{\mathbf{z}}_k) \circ \widehat{\mathbf{x}}_k \tag{3.19}$$

such that

$$\lim_{k\to\infty} \widehat{\mathbf{x}}_k = \overline{\mathbf{x}}, \tag{3.20}$$

where k indexes the iterations and the composition operator '$\circ$' depends on the involved Lie group. For example, if one considers a matrix Lie group then the product operation to be performed is the matrix multiplication. If real-valued (resp. non-zero) vectors are considered, then the respective product operation may be defined, for example, as the (resp. elementwise multiplication) addition. See (Warner, 1987; Varadarajan, 1974) for further information. In practice, the convergence to the optimal $\widehat{\mathbf{x}}$ can be established when the incremental displacement $\widetilde{\mathbf{x}}_k = \mathbf{x}(\widetilde{\mathbf{z}}_k)$ is arbitrarily close to the identity element of the involved group, i.e. when $\|\widetilde{\mathbf{z}}_k\| < \epsilon$.

A major difference amongst existing iterative non-linear optimization procedures concerns how the increment is found. This involves the determination of both the best direction of descent and the optimal step along this direction (Isaacson and Keller, 1966; Dennis and Schnabel, 1983; Luenberger, 1984). For the sake of simplicity, this step was not explicitly considered above.

In the sequel, we describe an efficient second-order approximation method (Malis, 2004) to obtain the direction of descent. This method is of particular interest when the gradient at the (unknown) solution is available, and if the Lie algebra is used to parametrize the variables. Standard line search algorithms can afterward be applied to compute the optimal step. This section ends with a discussion on how to appropriately initialize the optimization procedure.

3.3.1 Derivation

The non-linear optimization problem (3.18) can be concisely rewritten as

$$\min_{\widetilde{\mathbf{z}}} \; \frac{1}{2} \left\| \mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}\big) \right\|^2, \tag{3.21}$$

where the objective consists in finding the optimal $\mathbf{x}(\widetilde{\mathbf{z}}^{\circ})$ such that

$$\mathbf{x}(\widetilde{\mathbf{z}}^{\circ}) \circ \widehat{\mathbf{x}} = \overline{\mathbf{x}}. \tag{3.22}$$

In this case, the image alignment is achieved, i.e. $\boldsymbol{\mathcal{I}}' = \boldsymbol{\mathcal{I}}^*$. A standard technique to iteratively solve this problem consists in performing an expansion of the function in Taylor series and applying a necessary condition of optimality.

In 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, a second-order approximation of $\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}\big)$ in Taylor series about the current estimate $\widehat{\mathbf{x}}$ (i.e. about $\widetilde{\mathbf{z}} = \mathbf{0}$) is

$$\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}\big) = \mathbf{d}(\widehat{\mathbf{x}}) + \nabla_{\widetilde{\mathbf{z}}} \mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}\big)\Big|_{\widetilde{\mathbf{z}}=\mathbf{0}} \widetilde{\mathbf{z}} + \frac{1}{2} \nabla_{\widetilde{\mathbf{z}}} \Big(\nabla_{\widetilde{\mathbf{z}}} \mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}\big)\Big|_{\widetilde{\mathbf{z}}=\mathbf{0}} \widetilde{\mathbf{z}} \Big) \widetilde{\mathbf{z}} + o\big(\|\widetilde{\mathbf{z}}\|^3\big), \tag{3.23}$$

or more compactly,

$$\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}\big) = \mathbf{d}(\widehat{\mathbf{x}}) + \mathbf{J}(\widehat{\mathbf{x}})\, \widetilde{\mathbf{z}} + \frac{1}{2} \mathbf{S}(\widehat{\mathbf{x}}, \widetilde{\mathbf{z}})\, \widetilde{\mathbf{z}} + o\big(\|\widetilde{\mathbf{z}}\|^3\big), \tag{3.24}$$

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 $\mathbf{J}\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}\big)$ again about the current estimate $\widehat{\mathbf{x}}$ (i.e. about $\widetilde{\mathbf{z}} = \mathbf{0}$) is

$$\mathbf{J}\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}\big) = \mathbf{J}(\widehat{\mathbf{x}}) + \mathbf{S}(\widehat{\mathbf{x}}, \widetilde{\mathbf{z}}) + o\big(\|\widetilde{\mathbf{z}}\|^2\big), \tag{3.25}$$

with the second-order remainder $o\big(\|\widetilde{\mathbf{z}}\|^2\big)$. By injecting $\mathbf{S}(\widehat{\mathbf{x}}, \widetilde{\mathbf{z}})$ from (3.25) in (3.24) 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 finally obtained:

$$\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}\big) = \mathbf{d}(\widehat{\mathbf{x}}) + \frac{1}{2} \Big(\mathbf{J}(\widehat{\mathbf{x}}) + \mathbf{J}\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}\big) \Big)\, \widetilde{\mathbf{z}}. \tag{3.26}$$

We can then apply a necessary condition of optimality. A necessary condition so that $\widetilde{\mathbf{z}} = \widetilde{\mathbf{z}}^{\circ}$ is an extremum of our cost function in (3.21) is

$$\mathbf{0} = \nabla_{\widetilde{\mathbf{z}}} \Big(\frac{1}{2} \mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}\big)^{\top} \mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}\big) \Big)\Big|_{\widetilde{\mathbf{z}}=\widetilde{\mathbf{z}}^{\circ}} \tag{3.27}$$

$$= \nabla_{\widetilde{\mathbf{z}}} \mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}}) \circ \widehat{\mathbf{x}}\big)\Big|_{\widetilde{\mathbf{z}}=\widetilde{\mathbf{z}}^{\circ}}^{\top} \mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}}^{\circ}) \circ \widehat{\mathbf{x}}\big), \tag{3.28}$$

or more compactly,

$$\mathbf{0} = \mathbf{J}(\overline{\mathbf{x}})^{\top} \mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}}^{\circ}) \circ \widehat{\mathbf{x}}\big), \tag{3.29}$$

using (3.22). Provided that $\mathbf{J}(\overline{\mathbf{x}})$ is full rank (this condition will be discussed in the next subsection), then one must have

$$\mathbf{d}\big(\mathbf{x}(\widetilde{\mathbf{z}}^{\circ}) \circ \widehat{\mathbf{x}}\big) = \mathbf{0} \tag{3.30}$$

from (3.29). The roots of this system of non-linear equations are generally difficult to obtain in closed form. However, using the Taylor approximation (3.26) about $\widetilde{\mathbf{z}} = \widetilde{\mathbf{z}}^{\circ}$ along with (3.22) yield the following system of equations

$$\frac{1}{2}\big(\mathbf{J}(\widehat{\mathbf{x}}) + \mathbf{J}(\overline{\mathbf{x}})\big)\, \widetilde{\mathbf{z}}^{\circ} = -\mathbf{d}(\widehat{\mathbf{x}}), \tag{3.31}$$

where $\mathbf{d}(\widehat{\mathbf{x}})$ and the Jacobian $\mathbf{J}(\widehat{\mathbf{x}})$ are both completely computed using current information. On the other hand, the entire Jacobian $\mathbf{J}(\overline{\mathbf{x}})$ at the reference (true) values cannot be obtained from current data because some of them are unknowns. Only a part of it can be computed (by applying the chain rule), since the reference image is anyway available. Some parts must then be approximated, e.g. using the current estimate, so that (3.31) is a rectangular linear system. Nevertheless, in some particular cases where the warping function (3.3) is a group action of G on $\mathbb{P}^2$ (e.g. in the planar case of Subsection 3.2.1), a rectangular linear system is obtained from (3.31) without any approximation. Independently (either approximately or exactly) of how a rectangular linear system is obtained from (3.31), i.e.

$$\mathbf{J}'\, \widetilde{\mathbf{z}}^{\circ} = -\mathbf{d}(\widehat{\mathbf{x}}), \tag{3.32}$$

where $\mathbf{J}'$ represents our direction of descent, its solution $\widetilde{\mathbf{z}}^{\circ}$ is found in the least-squares sense by solving its normal equations

$$\mathbf{J}'^{\top}\mathbf{J}'\, \widetilde{\mathbf{z}}^{\circ} = -\mathbf{J}'^{\top}\mathbf{d}(\widehat{\mathbf{x}}), \tag{3.33}$$

obtained through multiplying both sides of (3.32) on the left by $\mathbf{J}'^{\top}$. It can be noted that the obtained $\widetilde{\mathbf{z}}^{\circ}$ may not align the images at the first iteration, especially because a Taylor approximation of the true non-linear equations (3.30) is performed. Thus, the solution $\widetilde{\mathbf{z}}^{\circ}$ represents an incremental displacement that is iteratively used to generate the sequence (3.19) 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 words, differently from second-order minimization techniques (e.g. Newton), the Hessians are never computed explicitly. This also contributes to obtain nicer convergence properties.

3.3.2 Initialization

Non-linear optimization procedures have to be adequately initialized, either because of observability issues or because they are in general only locally convergent. It is discussed below how both issues can be tackled within direct image registration methods.

A hierarchical formulation

Observability in control theory is a measure for how well internal states of a system can be inferred by knowledge of its external outputs. In this thesis, the output is provided by a single camera. Whilst for state-space systems this measure can be obtained by analyzing the rank of the observability Gramian, within image registration methods this measure can be obtained by analyzing the rank of the associated Jacobians. We remark that these Jacobians (motion, structure, illumination changes) can be defined in both calibrated and uncalibrated settings. However, due to various types of system noise, in many cases this analysis is not very useful in practice.

A typical observability problem faced by roboticists occurs when mapping the scene using bearing-only sensors. In this case, a single measurement is insufficient to constraint a landmark location. Rather, it must be sensed from multiple vantage points. Translating to our monocular case, at the beginning of a sequential image registration task (i.e. registration of a reference image to successive frames of a video sequence) the amount of translation may be small relatively to the distance to the scene. If this occurs, the Jacobian related to the structure is ill-conditioned, indicating that the structure parameters are not yet observable. In this situation, the motion parameters together with the illumination ones can explain most of the image differences. This latter reasoning also applies once the optimal structure parameters have already been obtained. In this case, there is no reason to maintain them as optimization variables. Besides that their values may be perturbed, e.g. when the image resolution decreases, less parameters in the minimization signify more available computing resources. Once again, motion parameters and illumination ones can explain most of the image discrepancies. For these reasons, we propose a hierarchical framework in the sense of the number of parameters to explain the image motion.

In other terms, for every new image that is acquired, we initially attempt to align it using only a subset of parameters. The structure parameters are only simultaneously used as optimization variables whenever the difference between the obtained cost value and the resulting one from previous (image) optimization exceeds the image noise. We remark that in any case the structure (plus motion and illumination) parameters are always required to compute the discrepancies $\mathbf{d}(\widehat{\mathbf{x}})$. These parameters can be either the given initial values or the optimal ones from preceding image registrations (in the case of a sequential registration task). In fact, this shows how all past observations in a sequential task effectively contribute to incrementally building and maintaining a coherent description of the map and poses.

Augmenting the domain and the rate of convergence

A common limitation of efficient non-linear optimization procedures regards its domain of convergence. Although the parameters are obtained by a second-order approximation method with nice convergence properties, there is no guarantee that the global minimum will be reached. As stated, global minimization procedures are too computationally intensive to be performed in a real-time setting.

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. Nonetheless, we remark that, even though a recovered set of parameters can represent a local minimum, it may be close to the global one. If this occurs, the regions may have been effectively aligned in the image. A standard pose recovery technique can then be used with all these registered (i.e. corresponding) pixels. Afterward, the scene can be reconstructed by triangulating them (Faugeras et al., 2001). In addition to augmenting the domain of convergence, this approach may also augment the rate of convergence. If these estimated motion and/or structure are closer to the true ones than those obtained by the direct registration, 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 also do not ensure that the global minimum will be attained. As stated in Subsection 2.1.1, these techniques are not fully invariant to all possible changes in geometric and photometric parameters.

Thus, one can also rely on other predictors to improve the convergence properties of direct registration methods. In fact, the coupling between the proposed deterministic image registration with a probabilistic filtering technique can be performed at this stage. In the case of a sequential image registration task, a Kalman filter can be used to provide both another estimate of the optimization variables and the covariances. The input (i.e. observations) to the filtering are the recovered parameters from the optimization process. In order to initialize the system (i.e. when a new image is available), the best set of parameters amongst all predictors is simply 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) and/or on the model of motion (e.g. constant velocity) may not be realistic in many scenarios.

3.4 Summary

This chapter presents appropriate parametric transformation models and optimization methods for directly and robustly aligning images, including color ones. The proposed photometric models ensure robustness to arbitrary illumination changes, do not require prior knowledge (including the spectral ones) of the object, illuminants and camera, and naturally encompass gray-level images. Various design parameters, some limitations of the framework and possible solutions are also discussed here. Next chapters are devoted to the application of such a framework for performing various vision-based tasks.

Chapter 4

Uncalibrated camera

Consider the pinhole camera model (see Chapter 1). The uncalibrated image registration case refers to the setting where the camera's intrinsic parameters are neither known a priori nor estimated on-line. These parameters are not explicitly used since all involved entities to perform the registration are defined in the projective space. This setting is used in many vision-based applications, such as for visual tracking.

This chapter proposes a generic technique for directly and robustly visual tracking unknown objects under unknown imaging conditions. Another application of this technique (visual servoing) will be investigated in Chapter 6. The proposed algorithm makes large use of the proposed models and methods described in Chapter 3. Comparison results with existing direct methods show significant improvements in the tracking performance. Extensive experiments confirm the robustness and reliability of our method.

4.1 Related work on uncalibrated direct image registration

Image registration is a fundamental component for a variety of vision-based applications, e.g. in medical image analysis, augmented reality, and robotics. Given its importance, a huge body of literature has been published (Brown, 1992; Maintz and Viergever, 1998). An exhaustive description of this production is beyond the scope of this chapter. Therefore, let us introduce the context on which this chapter focuses.

First of all, this chapter refers only to uncalibrated direct algorithms that are both robust to (at least a certain degree of) illumination changes and potentially real-time for a robotic system. Thus, techniques which suppose either that the brightness constancy assumption holds (Lucas and Kanade, 1981; Benhimane and Malis, 2004) or that perform a bundle adjustment are not considered here. This latter aims at avoiding both non-causal estimation and computational burden.

Additionally, since only local non-linear optimization techniques can be used in a real-time setting, we suppose that an initial estimate sufficiently close to the true parameters are available. This is the case where either the images present a sufficient amount of overlapping, or a suitable prediction is available (see Subsection 3.3.2). Methods based on optical flow computation (Negahdaripour, 1998; Black et al., 2000; Haussecker and Fleet, 2001) are also not considered because they suppose a too small inter-frame displacement of the objects.

Furthermore, we consider applications where off-line learning steps are not possible to be executed prior to the registration task. Hence, methods such as (Hager and Belhumeur, 1998; La Cascia et al., 2000; Jurie and Dhome, 2002; 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 generic scheme. In particular, this enables the control of all six degrees-of-freedom of a robot. Thus, the visual tracking technique proposed in, e.g., (Comaniciu et al., 2000), though effective, is not sufficient for our purposes since it provides up to a similarity transformation. Moreover, this latter technique only works for color images. We investigate techniques that can work with both gray-scale and color images.

4.2 Visual tracking robust to generic illumination changes

Visual tracking can be formulated as an incremental direct registration. That is, as the problem of estimating the incremental transformations which optimally align a reference image with successive frames of a video sequence (see Fig. 3.1). In this case, the reference image is also called the fixed image, and the current image is referred to as the moving one.

Specifically, we tackle here an important issue to all vision-based algorithms: the robustness to generic lighting changes. Indeed, we address the efficient tracking of either Lambertian or non-Lambertian objects under unknown imaging conditions. To this end, a possible scheme to increase the robustness to variable illumination is through 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 inter-frame displacements (geometric and/or photometric) of the object are large (Baker et al., 2003). Another widely used technique is to model the change in illumination as an affine transformation, e.g. (Baker et al., 2003; Bartoli, 2006; Jin et al., 2003). 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 proved to be inefficient in the case of direct tracking (Baker et al., 2003). The reasons are twofold. Firstly, 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. Secondly, 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). For example, strong differences caused by motion may be discarded though weak differences produced for example by shadows may be considered. On the other hand, those robust functions might be applied to handle unknown occlusions since their realistic modeling is a rather difficult task.

This chapter proposes a new direct visual tracking approach where the robustness to lighting changes is assured by using the proposed model of illumination changes presented in Chapter 3, together with an appropriate geometric model of image motion. The resulting photo-geometric generative model is generic. As for the model of lighting variations, it does not require the attributes of the imaging sensors (e.g. spectral response characteristics), of the light sources (e.g. number, power, pose), or about the properties of the surface (e.g. reflectance, shape). As for the geometric model of image motion, we show here how to encompass both rigid and deformable objects whilst still preserving that robustness property. All used models can be devised such that the real-time constraint are satisfied. Furthermore, we demonstrate that the efficient minimization procedure presented in Chapter 3 can simultaneously obtain the optimal global and local parameters related to all those models. Hence, relatively large rates and domains of convergence are achieved. We remark that the procedure is computationally efficient because the Hessians are never computed explicitly.

Results are reported 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 inter-reflections. Results demonstrate that the proposed approach also accommodates them without making any additional change. For the experiments, representative sample surfaces were chosen, which range from smooth to rough, and 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 large inter-frame displacements of the object are carried out.

4.2.1 From planar rigid objects to generic deformable ones

Subsection 1.3.1 briefly presented the geometric relation between a pair of uncalibrated images of a rigid scene. It is also shown that planar surfaces represent a particular case of that generic law. Here, it is described an extension of that relation in order to encompass generic deformable surfaces as well.

Consider a 3D point $\underline{\mathbf{m}}^* \in \mathbb{R}^3$. A change of its relative position with respect to the reference frame can written as (Malis, 2007):

$$\underline{\boldsymbol{m}}^* = \frac{1}{\kappa^*}\,\underline{\mathbf{m}}^* + \boldsymbol{\eta}^*, \tag{4.1}$$

where $\kappa^* \in \mathbb{R}_+$ takes into consideration only deformations that change the 3D structure of the object in the reference frame but do not change the reference image, and $\boldsymbol{\eta}^* = [\eta_u^*, \eta_v^*, 0]^\top \in \mathbb{R}^3$ captures the remaining deformations. Thus, we can generalize the geometric model (1.23) presented in Chapter 1

$$\mathbf{p} \propto \mathbf{G}\,\mathbf{p}^* + \rho^*\,\mathbf{e}$$

as

$$\mathbf{p} \propto \mathbf{G}\,(\mathbf{p}^* + \boldsymbol{\delta}^*) + \rho^*\,\mathbf{e}, \tag{4.2}$$

where $\boldsymbol{\delta}^* = [\delta_u^*, \delta_v^*, 0]^\top \in \mathbb{R}^3$ is a deformation vector proportional to $\boldsymbol{\eta}^*$ and to the unknown (and possibly time-varying) camera's intrinsic parameters. It can be noted that, in this case, the projective parallax $\rho^* \in \mathbb{R}$ also takes into consideration the deformation $\kappa^* > 0$. As in Eq. (1.23), $\mathbf{G} \in \mathbb{SL}(3)$ is a homography relative to an arbitrary plane Π not going through $\mathcal{O}^*$, and $\mathbf{e} \in \mathbb{R}^3$ denotes the epipole.

Therefore, by using the generic relation expressed in (4.2) a unified hierarchical geometric modeling is achieved. Indeed, easy transition between models is assured as follows.

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

$$\boldsymbol{\delta}^* = \mathbf{0} \quad \text{and} \quad \rho^* = 0. \tag{4.3}$$

Particular case (Rigid objects). The class of non-planar 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:

$$\boldsymbol{\delta}^* = \mathbf{0} \quad \text{and} \quad \dot{\rho}^* = 0. \tag{4.4}$$

Particular case (Objects under "invariant" deformation). Increasing the degree of complexity, the next class comprises the deformable surfaces such that

$$\boldsymbol{\delta}^* = \mathbf{0} \quad \text{and} \quad \dot{\rho}^* \neq 0. \tag{4.5}$$

Finally, in the most generic case (i.e. the one with the highest degree of complexity), the class of generic deformable objects has

$$\boldsymbol{\delta}^* \neq \mathbf{0} \quad \text{and} \quad \dot{\rho}^* \neq 0 \tag{4.6}$$

within the generic relation expressed in (4.2).

4.2.2 Surface modeling

The modeling of surfaces is an important design parameter within estimation methods from visual data. Besides the scene structure, illumination changes are also modeled in this thesis as a surface, as described in Chapter 3. Additionally, we showed that regularization techniques are needed in both cases so as to avoid constructing an under-constrained 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. Typically, they represent a compromise between computational complexity, robustness and accuracy.

Let us discuss first the number of surfaces. Consider an n-channel image, $n \geq 1$. Of course, the case where $n = 1$ corresponds to a gray-level 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 generic deformable object along with a fully coupled model of lighting variations, a total of $n^2 + 3$ surfaces are required to accurately and robustly explain the image motion. They represent:

- the surface related to the projective parallax:

$$\rho^* = f_\rho(\boldsymbol{\lambda}_\rho, \mathbf{p}); \tag{4.7}$$

- the surface related to the generic deformation in the u-direction:

$$\delta_u^* = f_\delta(\boldsymbol{\lambda}_u, \mathbf{p}); \tag{4.8}$$

- the surface related to the generic deformation in the v-direction:

$$\delta_v^* = f_\delta(\boldsymbol{\lambda}_v, \mathbf{p}); \tag{4.9}$$

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

$$\mathcal{S}_{kj} = f_h(\boldsymbol{\gamma}_{kj}, \mathbf{p}), \quad k, j = 1, 2, \ldots, n. \tag{4.10}$$

With respect to the real-valued function itself $f_{(\cdot)}$, the appropriate choice depends on several factors, such as the assumptions concerning the surface (e.g. smoothness) and on the required system's performance. See Subsection 3.2.2 for a brief discussion about this subject. Of course, different choices can be made for each one of the surfaces. Additionally, the number of samples to define a surface, i.e. $\dim(\boldsymbol{\lambda}_\rho), \dim(\boldsymbol{\lambda}_u), \dim(\boldsymbol{\lambda}_v)$ and $\dim(\boldsymbol{\gamma}_{kj})$, has also an impact on the system's performance. Nevertheless, a hierarchical approach can be applied to find the appropriate number, starting from a planar surface to higher-order approximations.

4.2.3 The full system

The full system is composed of the appropriate parametric transformation model and of the optimization method.

As for the modeling, a photo-geometric generative model can be defined from the generic model of illumination changes (3.13), along with the warping function (3.3) defined from the generic relation between uncalibrated views (4.2). More formally, the proposed parametric transformation model is given by

$$\boldsymbol{\mathcal{I}}'(\mathbf{g}^{\mathrm{u}}, \mathbf{h}, \mathbf{p}^*) = \boldsymbol{\mathcal{S}}(\boldsymbol{\Gamma}, \mathbf{p}^*) \bullet \boldsymbol{\mathcal{I}}\big(\mathbf{w}(\mathbf{g}^{\mathrm{u}}, \mathbf{p}^*)\big) + \boldsymbol{\beta}, \tag{4.11}$$

where 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. Further, in this case of uncalibrated framework (explicited by a superscript 'u' in this standard roman font), the geometry between views is described by the set of parameters

$$\mathbf{g}^{\mathrm{u}} = \{\mathbf{G}, \mathbf{e}, \rho^*, \boldsymbol{\delta}^*\}. \tag{4.12}$$

The set of photometric parameters is denoted by

$$\mathbf{h} = \{\boldsymbol{\mathcal{S}}, \boldsymbol{\beta}\}. \tag{4.13}$$

Let us now discuss the important issue of parametrizing these entities. As for the geometric parameters, consider the (4×4) matrix

$$\mathbf{Q} = \begin{bmatrix} \mathbf{G} & \mathbf{e} \\ \mathbf{0} & 1 \end{bmatrix} \quad \in \mathbb{SA}(3). \tag{4.14}$$

The Lie group $\mathbb{SA}(3)$ (the special affine group) is homeomorphic to $\mathbb{SL}(3) \times \mathbb{R}^3$. The Lie group $\mathbb{SE}(3)$ 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 are here denoted by $\boldsymbol{\upsilon} \in \mathbb{R}^{8+3}$, i.e. $\mathbf{Q} = \mathbf{Q}(\boldsymbol{\upsilon})$. As discussed in Subsection 1.1.2, the mechanism for passing information from the Lie algebra to the related Lie group is the exponential mapping. Hence, the set of geometric entities in the uncalibrated case is fully parametrized by

$$\mathbf{z}_g^{\mathrm{u}} = \{\boldsymbol{\upsilon}, \boldsymbol{\lambda}_\rho, \boldsymbol{\lambda}_u, \boldsymbol{\lambda}_v\}. \tag{4.15}$$

As for the photometric parameters $\mathbf{h}$ (4.13), its parametrization is given by

$$\mathbf{z}_h = \{\boldsymbol{\Gamma}, \boldsymbol{\beta}\} \tag{4.16}$$

with $\boldsymbol{\Gamma} = \{\gamma_{kj}\}$.

Then, for real-time systems, only a local non-linear optimization procedure can be applied to estimate all those parameters. To this end, an initial estimate $\widehat{\mathbf{g}}^{\mathrm{u}}$ and $\widehat{\mathbf{h}}$ sufficiently close to the true solution is needed. Thus, the proposed model (4.11) is transformed into:

$$\boldsymbol{\mathcal{I}}'\big(\mathbf{g}^{\mathrm{u}}(\widetilde{\mathbf{z}}_g^{\mathrm{u}}) \circ \widehat{\mathbf{g}}^{\mathrm{u}}, \mathbf{h}(\widetilde{\mathbf{z}}_h) \circ \widehat{\mathbf{h}}, \mathbf{p}^*\big) = \boldsymbol{\mathcal{S}}\big(\widetilde{\boldsymbol{\Gamma}} \circ \widehat{\boldsymbol{\Gamma}}, \mathbf{p}^*\big) \bullet \boldsymbol{\mathcal{I}}\big(\mathbf{w}(\mathbf{g}^{\mathrm{u}}(\widetilde{\mathbf{z}}_g^{\mathrm{u}}) \circ \widehat{\mathbf{g}}^{\mathrm{u}}, \mathbf{p}^*)\big) + \widetilde{\boldsymbol{\beta}} \circ \widehat{\boldsymbol{\beta}}, \tag{4.17}$$

where the symbol '$\circ$' refers to the related composition rule (see Subsection 3.3). Therefore, the uncalibrated visual tracking problem can be cast as the optimization problem described in (3.18):

$$\min_{\widetilde{\mathbf{z}}^{\mathrm{u}} = \{\widetilde{\mathbf{z}}_g^{\mathrm{u}}, \widetilde{\mathbf{z}}_h\}} \; \frac{1}{2} \sum_i \Big[\boldsymbol{\mathcal{I}}'\big(\mathbf{x}^{\mathrm{u}}(\widetilde{\mathbf{z}}^{\mathrm{u}}) \circ \widehat{\mathbf{x}}^{\mathrm{u}}, \mathbf{p}_i^*\big) - \boldsymbol{\mathcal{I}}^*(\mathbf{p}_i^*) \Big]^2, \tag{4.18}$$

with $\mathbf{x}^{\mathrm{u}} = \{\mathbf{g}^{\mathrm{u}}, \mathbf{h}\}$ and its respective parametrization $\mathbf{z}^{\mathrm{u}} = \{\mathbf{z}_g^{\mathrm{u}}, \mathbf{z}_h\}$. The framework presented in Subsections 3.3 and 3.3.2 can then be applied to solve this optimization problem efficiently and with nice convergence properties.

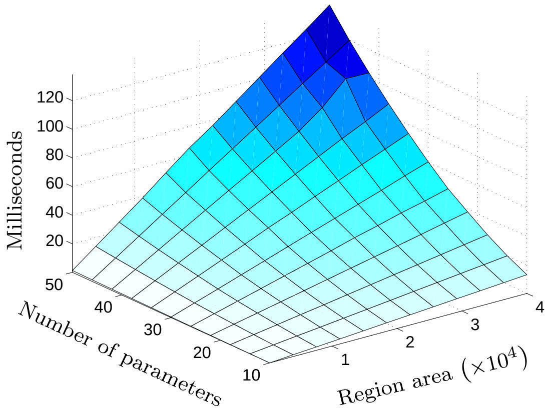


Figure 4.1. Processing time per iteration for a non-optimized implementation of our uncalibrated registration method in C on a Pentium 3.2 GHz.

4.3 Experimental results

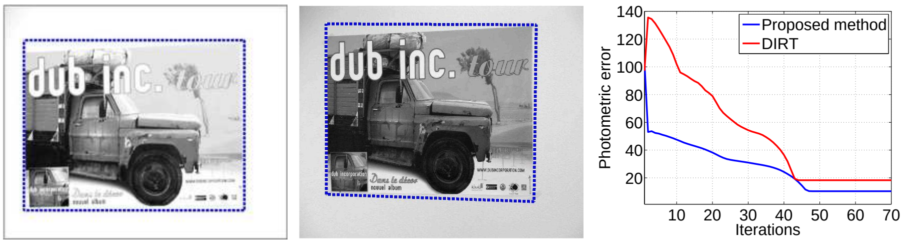
The generality, accuracy and robustness of the proposed direct image registration technique are verified through visually tracking rigid and deformable objects, with and without severe lighting variations, using both gray-level and color images. To this end, we select a template in the reference (i.e. first) image $\mathcal{I}^*$, which is then optimally aligned to successive images of the sequence. The hierarchical approach described in Subsection 4.2.1 is used, where the observed object 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. coarse-to-fine strategy, Kalman filter) is applied here, and also that bundle adjustment 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 $\mathcal{I}(t+1)$. In the sequel, let photometric error be defined as the Root Mean Square (RMS) of the difference image between the transformed image $\mathcal{I}'$ and the fixed one $\mathcal{I}^*$.

4.3.1 Synthetic data

Existing efficient direct image alignment techniques essentially tackle affine lighting variations. In order to show the generality of the proposed method, we compared it with DIRT (Bartoli, 2006) which is designed for that particular context. The non-optimized implementation of our method in C code runs at about 2.4 ms/iteration for an image region of 100 × 100 and for this affine case (10 parameters to be estimated) on a monocore Pentium 3.2 GHz with 2 GB of RAM. See Fig. 4.1 for the processing times when varying those parameters.

Comparison results of a particular image registration task is shown in Fig. 4.2. 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 DIRT 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 much higher, as well as it may accumulate drifts (thus leading to a failure earlier). In regard to other existing strategies, it has already been shown in (Bartoli, 2006) the improvements of DIRT with respect to the well-known SIC (Baker et al., 2003). Also, the strategy presented in (Jin et al., 2003) did not converge after 100 iterations and was not included in the figure.



(a) (b) (c)

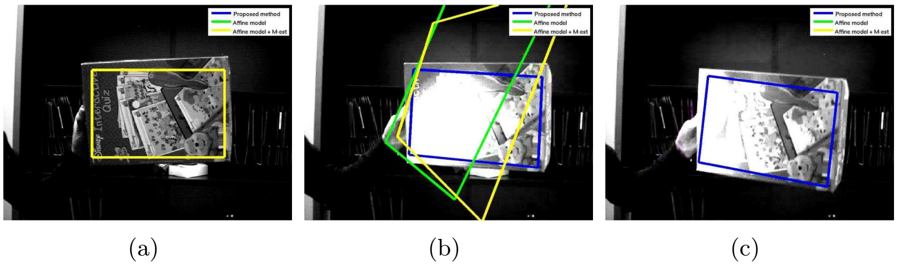
Figure 4.2. Comparison results of an image alignment task where relatively large displacements are present. As a means to compare with an existing method, the lighting variations between (a) the original image and (b) the synthetically transformed one comprise only affine changes. (c) The proposed method obtains smaller errors and does not get trapped into irrelevant minima.

4.3.2 Real data

With respect to generic illumination changes, we have applied the algorithm on several real-world sequences. They present severe changes in ambient, diffuse and specular reflections as well as shadows, inter-reflections and glints. In addition, they comprise relatively large inter-frame displacements and objects with unknown reflectance properties. The objects ranged from smooth to rough, and included metal and dielectrics. The unknown light sources are varied in power, type, number and moved in space. We have tested both the discretization and a Radial Basis Function (RBF) for approximating the surfaces. Albeit the obtained results are similar, the former can be more adequate to real-time systems since it yields a sparse Jacobian matrix.

Other comparison results. No existing efficient direct techniques are able to cope with that challenging scenario, especially when the object is not near-Lambertian and/or large displacements are carried out. In all case, we have tried DIRT, SIC and the method proposed in (Jin et al., 2003), but they have failed. This includes their variants. For example, by performing a photometric normalization with/or a robust error function (e.g. 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. 4.3 for an example, where the proposed method successfully registers all images with a median photometric error of 15.7 levels of gray-scale, executing a median of 6 iterations, 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.



(a) (b) (c)

Figure 4.3. Comparison results for the generic 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).

We have also tested a state-of-the-art robust feature-based technique (SIFT keypoints (Lowe, 2004) with RANSAC (Fischler and Bolles, 1981)) over our sequences. See Fig. 4.4 for a corresponding result. Small perturbations can be observed, such as in the top middle image of Fig. 4.4 (see the corner of the book near the letter ‘s’). If the estimated geometric parameters are to be used in a feedback control loop, then they should be filtered first so as to avoid discontinuities. Furthermore, since these methods have limited robustness to illumination changes, tracking failure (see the top right image of Fig. 4.4) is not surprising in this case of such a challenging scenario.

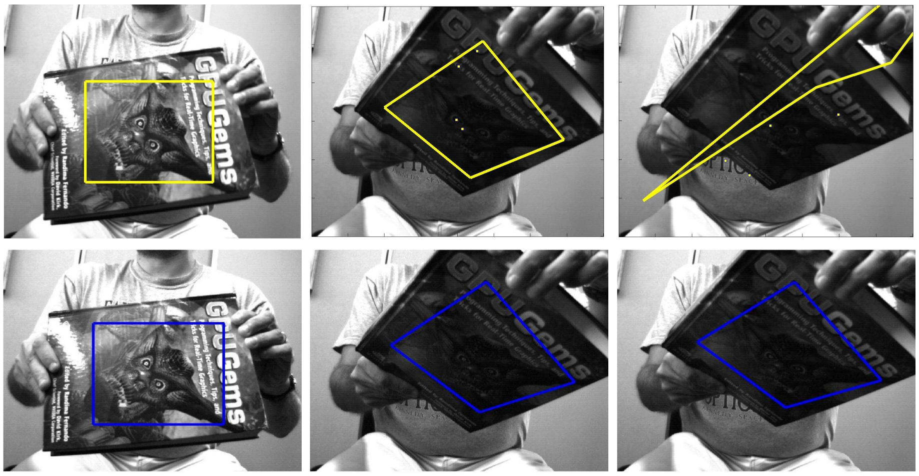
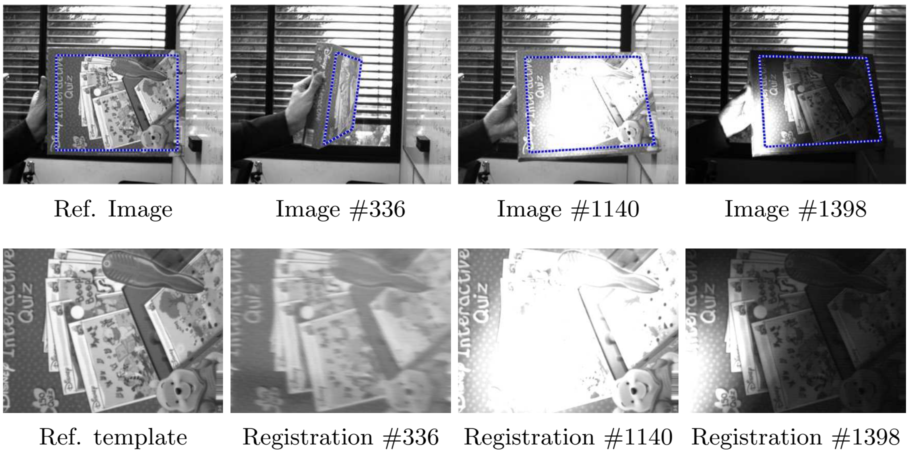


Figure 4.4. (Top) Failure for the generic case using a state-of-the-art robust feature-based technique. (Bottom) The proposed method (outlined in blue) successfully registers the reference image to all other images of the sequence.

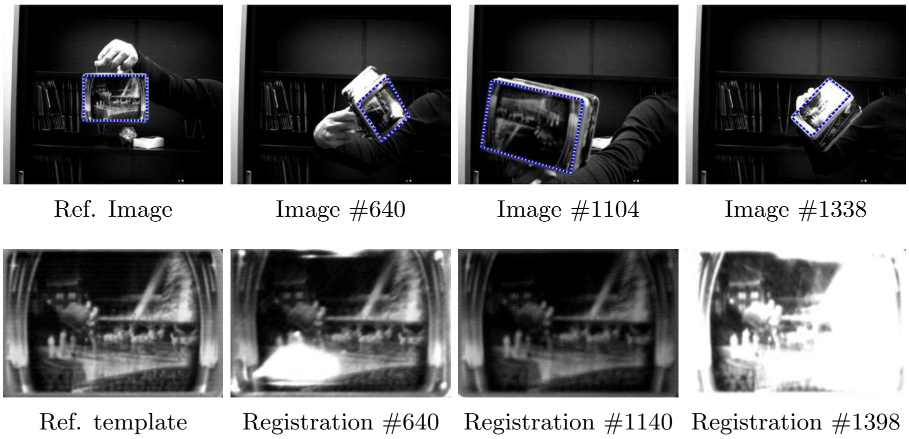
More sequences. Some results obtained for more gray-scale sequences of planar objects under those generic illumination changes are shown in Figs. 4.5 and 4.6. For the requested accuracy, the approach performed along these sequences a median of 4 and 5 iterations, respectively, and returned a median photometric error of 14.29 and 13.84 levels of gray-scale.



Ref. Image Image #336 Image #1140 Image #1398

Ref. template Registration #336 Registration #1140 Registration #1398

Figure 4.5. (Top) 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 tracker.

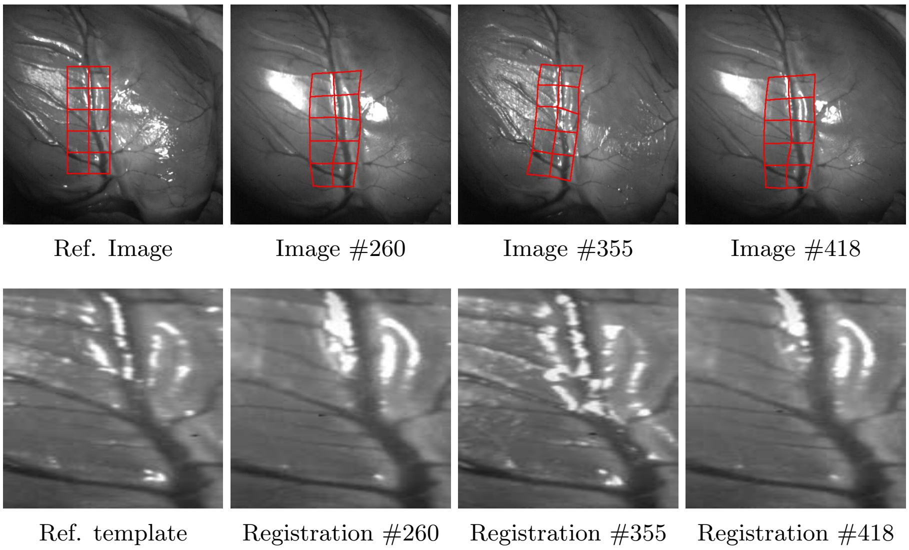


Ref. Image Image #640 Image #1104 Image #1338

Ref. template Registration #640 Registration #1140 Registration #1398

Figure 4.6. (Top) A metallic box is tracked under large changes in rotation and scale whilst experiencing high specular reflections. (Bottom) Registered images demonstrate the robustness of the alignment.

As for a deformable surface under severe illumination changes, the corresponding results are shown in Fig. 4.7. The deformable object is a beating heart. The bottom row shows that the images have been correctly aligned with the reference template, which encompasses an important vein. The stabilized images of the vein can then be used, for example, to improve its analysis and/or aid an intervention. All estimated surfaces are approximated in this case using RBFs, so as to improve accuracy, with centers equally spaced by 50 pixels. The approach performed along this sequence a median of 9 iterations, and returned a median photometric error of 8.32 levels of gray-scale.

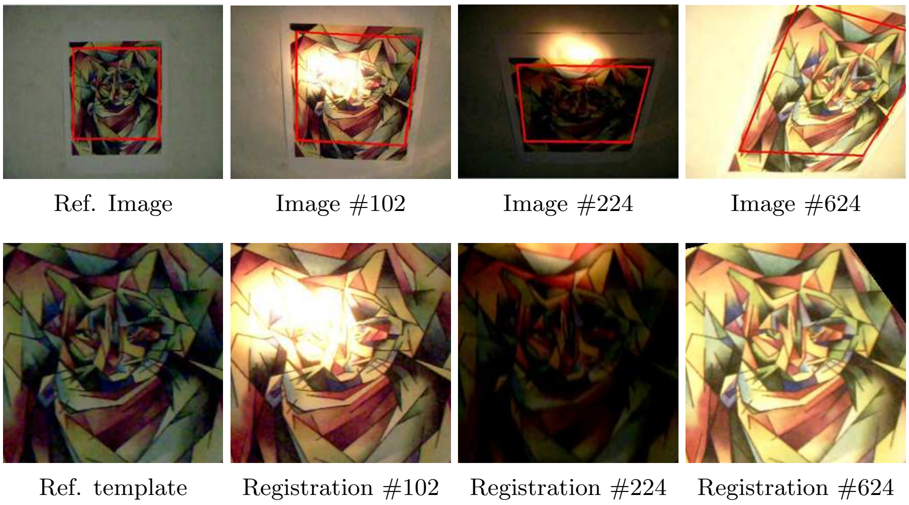

Ref. Image | Image #260 | Image #355 | Image #418

Ref. template | Registration #260 | Registration #355 | Registration #418

Figure 4.7. (Top) Direct image registration of a generic deformable surface. (Bottom) Corresponding templates aligned with respect to the reference one. Stability of the visual tracker is guaranteed despite many variable specularities.

As for color images, some registration results using objects of different shapes, including that of a non-planar object (a cylinder), are given in Figs. 4.8 and 4.9. Other sequences will also be shown in next chapters. No prior knowledge of the object's attributes (e.g. shape, albedos) is exploited. Despite the 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. For the requested accuracy, the approach performed along these sequences a median of 9 and 7 iterations, respectively, and returned a median photometric error of 15.73 and 16.76 levels of gray-scale.

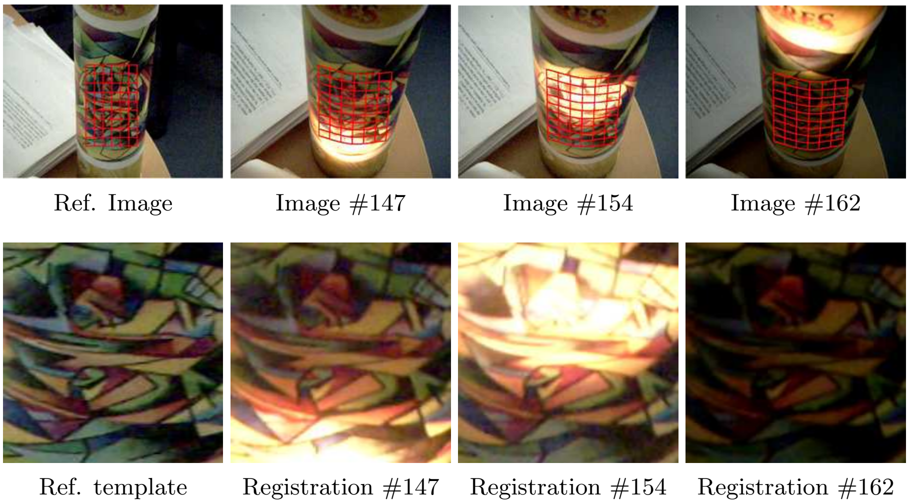
Some applications. Here, other two applications of the proposed registration method are given. The first one is in the field of augmented reality. The objective is to insert in the images geometrically coherent virtual objects. For example, in order to adapt advertisements on TV programs according to the audience's specific interests. See Fig. 4.10 for the corresponding results, where the images present variable specular, diffuse and ambient reflections. The specular component is primarily produced by a line source, albeit no assumptions about its characteristics are made. The last row shows the estimated surface related to the illumination changes, which is also modeled by a RBF with centers equally spaced by 50 pixels.



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

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

Figure 4.8. (Top) 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) Registered images demonstrate the stability of the alignment.

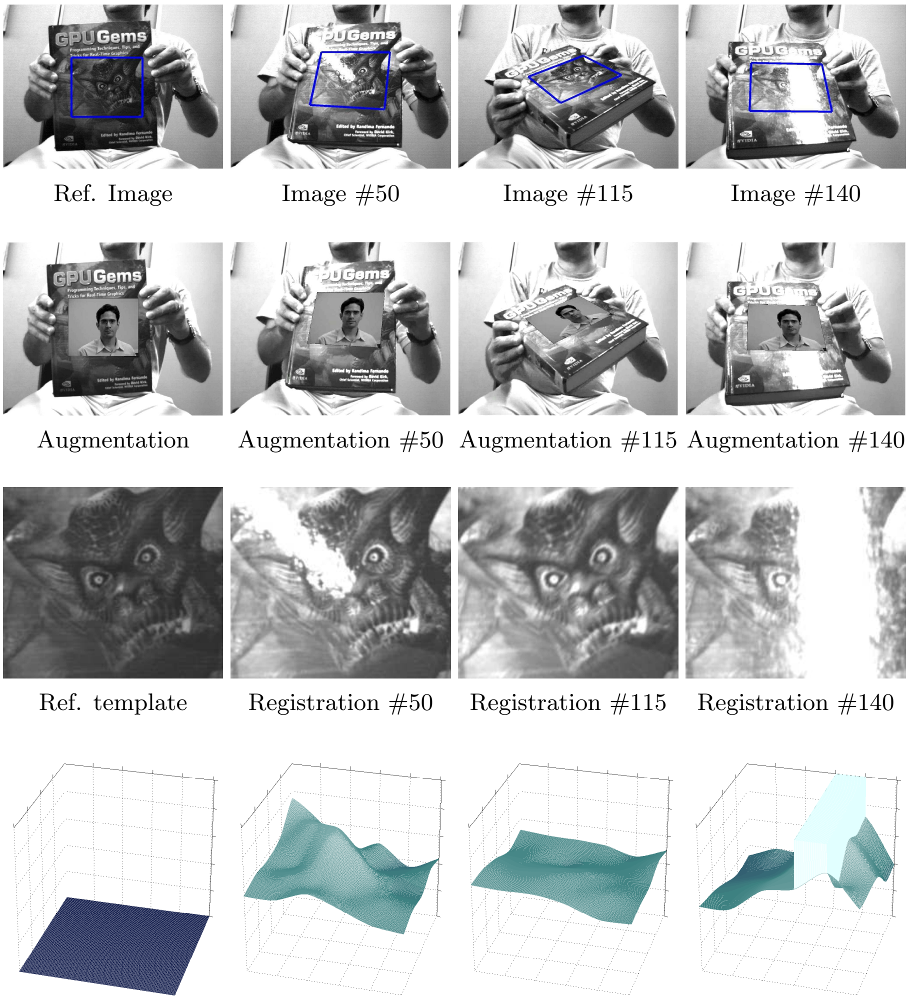


Ref. Image | Image #147 | Image #154 | Image #162

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

Figure 4.9. (Top) Direct color image registration of a cylinder. The unknown light source and camera perform unknown motions in space. No prior knowledge of the object's attributes (e.g. shape, albedos) is exploited. (Bottom) Registered images demonstrate the stability of the proposed tracker.

The second application is in the field of vision-based control. We show that an existing technique for implementing a follower robot (Benhimane et al., 2005) can be made robust to illumination changes. Here, the particular case of affine model for the illumination changes is applied (see Subsection 3.2.2). This robust visual tracking method has been tested and transferred to the company THALES Optronics. See Fig. 4.11 for a real-world experiment.



Ref. Image
Image #50
Image #115
Image #140

Augmentation
Augmentation #50
Augmentation #115
Augmentation #140

Ref. template
Registration #50
Registration #115
Registration #140

Initial surface
Surface #050
Surface #115
Surface #140

Figure 4.10. Application of the robust image registration method to image augmentation. (Second row) A virtual object (in this case, a photo) is automatically and accurately superimposed on the monster's face by geometrically aligning it. (Last row) The estimated surface represents the illumination changes with respect to the reference image.

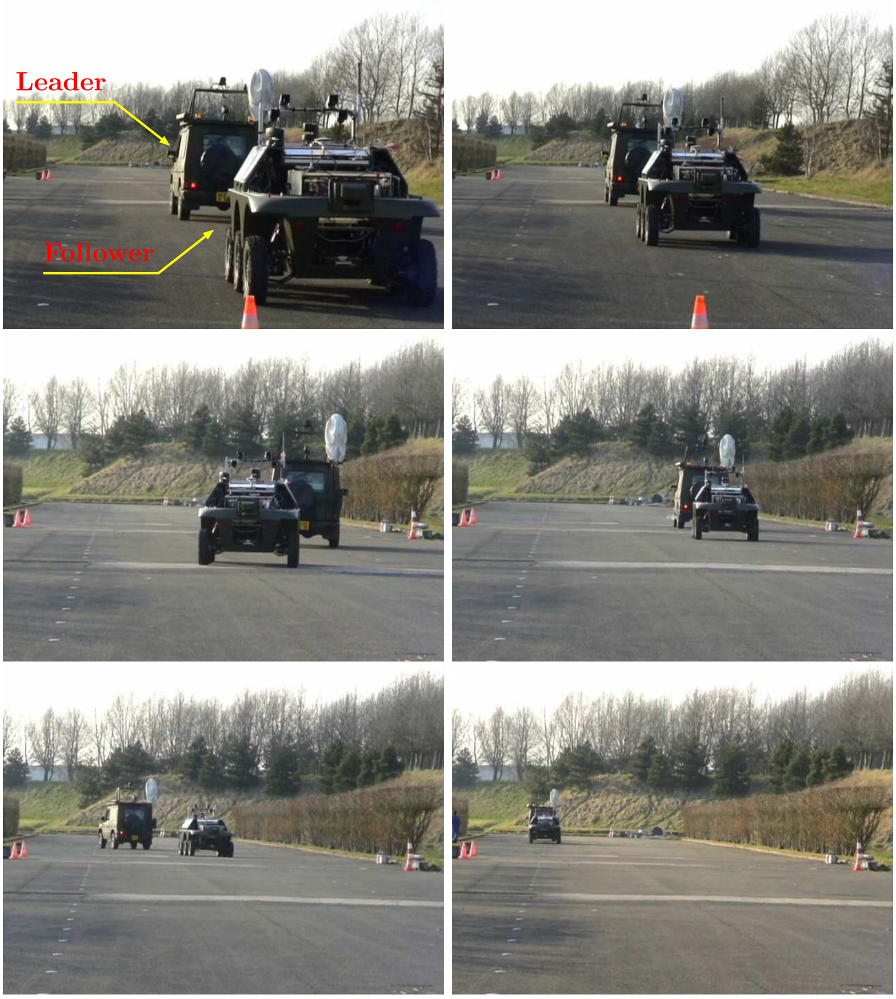


Figure 4.11. (Top to bottom and left to right) Real-world experiment of an automatic vision-based robot following behavior using the robust visual tracker. The tracker is made resilient here only to affine illumination changes.

4.4 Summary

We have presented a robust and generic direct image alignment algorithm for tracking unknown objects under arbitrary illumination changes, even in color images. The technique does not require the characteristics (including the spectral ones) of the light sources, of the surface, of the image sensors, and naturally encompasses gray-level images. Of course, the cost of processing is also dependent on the number of parameters to be estimated, which increases with increasing complexity of the scenario.

The technique uses a new parametric transformation model, and an efficient optimization procedure to simultaneously estimate all related parameters. We have showed that the object may undergo relatively large inter-frame displacements and the algorithm does not get wedged in irrelevant minima. Experimental results confirm the robustness and reliability of the proposed method.

Chapter 5

Calibrated camera

The calibrated registration case is referred here to the setting where the camera's intrinsic parameters are known a priori. They can be obtained by standard calibration techniques. These parameters are necessary to upgrade from the projective space to the Euclidean one. A typical application consists in estimating the camera pose and the scene structure with respect to a given reference frame. This task is central to autonomous robot navigation and hence, a multitude of approaches are available in the literature. However, the vast majority considers feature correspondences as an input to the estimation process.

This chapter formulates this essential task as a calibrated direct image registration problem. The proposed technique simultaneously obtains the correspondences, the camera pose, the scene structure, and the illumination changes, all directly using image intensities as observations. To this end, the models and methods described in Chapter 3 are largely used here. The fact of exploiting all possible image information leads to more accurate estimates, and avoids the inherent difficulties of reliably associating image features. We also show that, in this case, structural constraints can be enforced within the procedure as well (instead of a posteriori), namely the cheirality, the rigidity and those related to the lighting variations. Experimental results are provided for a variety of scenes, including urban and outdoor ones, under general camera motion and different types of perturbations.

5.1 Related work on calibrated direct image registration

In order to autonomously navigate in an unknown environment, a robot must be able to build a representation of the surrounding map and to self-localize with respect to it. Even though it is possible to perform the latter without the former by computer vision using an appropriate tensor (e.g. the Essential matrix), precision may be rapidly lost. This happens because important structural constraints, e.g. the scene rigidity, are not effectively exploited in a long run. Having understood that both estimation processes are intimately tied together, an appealing strategy is then to perform them simultaneously. This is generally referred to as Simultaneous Localization And Mapping (SLAM) in the robotics community. This class of methods focuses on computationally tractable algorithms that incrementally (i.e. causally) integrate information. At the expense of usually accumulating drifts earlier, they are suitable to real-time operation required by robotic platforms. A slightly different class of methods, mainly developed by the computer vision community, refers to Structure From Motion (SFM) techniques. Non-causal schemes fall into this latter class. These algorithms, mostly aimed at high levels of accuracy, are allowed to run in a time consuming batch process. This chapter focuses on the former class. The reader may refer to, e.g., (Tomasi and Kanade, 1992; Torr and Zisserman, 1999) for some well-established SFM methods.

The techniques that simultaneously and causally reconstruct the camera pose and the scene structure can be divided into two classes: feature-based and direct methods, which are briefly discussed below.

Feature-based methods to visual SLAM. A standard scheme to visual SLAM consists in first extracting a sufficiently large set of features (e.g. points, lines), and robustly matching them between successive images. These corresponding features are the input to the joint process of estimating the camera pose and scene structure. The majority of visual SLAM approaches consider feature correspondences as an input to the joint process of estimating the camera pose and the scene structure, e.g., (Broida et al., 1990; Davison, 2003; Eade and Drummond, 2006), independently of the applied filtering technique, e.g. EKF-SLAM (Smith and Cheeseman, 1986), FastSLAM 2.0 (Montemerlo et al., 2003). This represents the discrete case. Another possibility consists in computing the needed correspondences in the form of optical flow (the velocity). This has been exploited in, e.g., (Bruss and Horn, 1983; Hummel and Sundareswaran, 1993). In both cases, since the prior step of data association is highly error-prone, care must be taken in order to avoid propagating them to subsequent steps. On the other hand, these methods may handle relatively large inter-frame displacements of the objects.

Direct methods to visual SLAM. Another class of methods refers to those that directly exploit the intensity value of the pixels to obtain the required parameters. That is, there is no prior step of data association: this is simultaneously solved. An important strength of these methods concerns the level of accuracy that they can attain. This characteristic is mainly due to the exploitation of all possible image information, even from areas where gradient information is weak. The reader may refer to, e.g., (Irani and Anandan, 1999) for a more profound discussion about this subject.

In this spirit, the technique proposed in (Molton et al., 2004) can be assigned to this class. However, it does not consider the strong coupling between motion and structure in their separated estimation processes from pixel intensities. Furthermore, it is sensitive to variable illumination. In that method, new information is initialized with a "best guess". The technique proposed in (Jin et al., 2003), though using a unified framework, relies on the linearity of image gradient. This limits the system to work under very small inter-frame displacements of the objects. This approach is relatively robust to lighting variations, but its model of illumination changes is over-parametrized (which may lead, for example, to convergence problems). New information is initialized in a separate filter, and is inserted into the main filter after a probation period. Also within a unified framework, central catadioptric cameras are adequately dealt with in (Mei et al., 2006). This latter uses the same approximation method we use in this work for obtaining the related optimal parameters. Nevertheless, its set of parameters is different from ours not only because illumination changes are handled here, but also due to the structural constraints we explicitly enforce. Moreover, initialization is not a concern in that work.

5.2 The direct visual SLAM

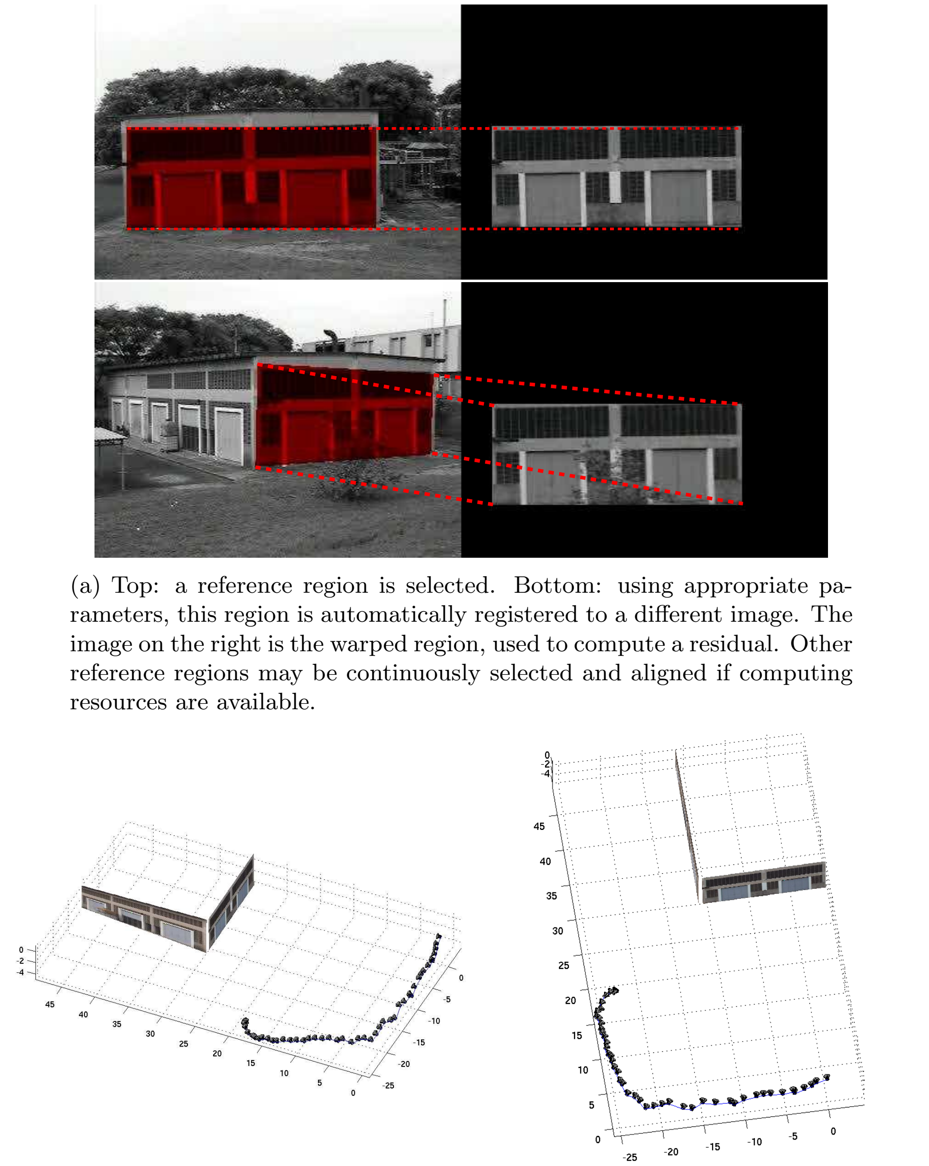
Here, we formulate the visual SLAM problem as a direct image registration task. In other words, we consider visual SLAM as the problem of estimating the appropriate parameters which optimally align a reference image with successive frames of a video sequence. Since the result of direct image alignments is such that each pixel intensity is matched as closely as possible across images, the technique in fact also returns a dense correspondence (see Fig. 5.1).

Indeed, a new approach is proposed for simultaneously obtaining the correspondences, the camera pose, the scene structure and the illumination changes, all directly using image intensities as observations. The fact of exploiting all possible image information leads to more accurate estimates, and avoids the inherent difficulties of reliably associating features. We also show here that, in this case, structural constraints can be enforced within the procedure as well (instead of a posteriori), namely the cheirality, the rigidity and those related to the lighting variations.

5.2.1 Surface modeling

Differently from the previous chapter on uncalibrated registration, the scene is considered here to be rigid so that consistent structure and poses can be obtained using a single camera. Although independently moving objects can make part of the scene, most of information contained in the image should be carried by rigidly attached objects. In this case, those independently moving ones can be viewed as outliers, whose corresponding regions of the image should be detected and rejected from the optimization procedure.

Nevertheless, the issue of surface modeling is still present, both for the scene structure and for the illumination changes. It can be noted that regularization techniques are needed in all cases. Once again, there exist various techniques to perform it and, as discussed in Subsection 3.2.2, the appropriate choice depends on several factors, such as the assumptions concerning the object (e.g. smoothness) and on the compromises to be satisfied. In particular, design choices must be made on the number of surfaces, on the function itself, and on the number of samples.



(a) Top: a reference region is selected. Bottom: using appropriate parameters, this region is automatically registered to a different image. The image on the right is the warped region, used to compute a residual. Other reference regions may be continuously selected and aligned if computing resources are available.

(b) A subset of the parameters recovered by the proposed alignment algorithm is naturally the camera pose and the scene structure. The figures show different viewpoints of the reconstructed scene and pose. Since monocular images are used, the scale factor is set arbitrarily.

Figure 5.1. The 'Hangar' sequence: A 751-frame example of visual SLAM by aligning reference regions to successive images. All pixels within both regions are exploited, leading to a precise result: the recovered angle between walls is of 89.7°. The regions are defined relative to where they were first viewed, and transferred to a common reference frame only for visualization purposes.

Let us discuss the number of surfaces. Consider an n-channel image, $n \geq 1$. Of course, $n = 1$ refers to a gray-level image. As presented in Chapter 1, the parallax in the Euclidean space corresponds to the inverse of the depth $(z^*)^{-1}$. Therefore, since only rigid scenes are considered, in the most general case of a fully coupled model of illumination changes we have a total of $n^2 + 1$ surfaces to be simultaneously estimated:

- the surface related to the Euclidean parallax (i.e. the scene structure):

$$(z^*)^{-1} = f_z(\boldsymbol{\lambda}_z, \mathbf{p}); \tag{5.1}$$

- the surface(s) related to the illumination changes:

$$\mathcal{S}_{kj} = f_h(\boldsymbol{\gamma}_{kj}, \mathbf{p}), \quad k, j = 1, 2, \dots, n. \tag{5.2}$$

5.2.2 The full system

The full system is composed of the appropriate parametric transformation model and of the optimization method.

As for the modeling, a photo-geometric generative model can be defined from our generic model of illumination changes (3.13), along with the warping model (3.3) using the generic relation between calibrated views (1.27). More formally, the parametric transformation model is given by:

$$\boldsymbol{\mathcal{I}}'(\mathbf{g}^{\mathrm{c}}, \mathbf{h}, \mathbf{p}^*) = \boldsymbol{\mathcal{S}}(\boldsymbol{\Gamma}, \mathbf{p}^*) \bullet \boldsymbol{\mathcal{I}}\big(\mathbf{w}(\mathbf{g}^{\mathrm{c}}, \mathbf{p}^*)\big) + \boldsymbol{\beta}, \tag{5.3}$$

where 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. Further, in this case of calibrated framework (explicited by a superscript 'c' in this standard roman font), the geometry between views is described by the set of parameters

$$\mathbf{g}^{\mathrm{c}} = \{\mathbf{R}, \mathbf{t}, (z^*)^{-1}\}. \tag{5.4}$$

The set of photometric parameters is denoted by

$$\mathbf{h} = \{\boldsymbol{\mathcal{S}}, \boldsymbol{\beta}\}. \tag{5.5}$$

Let us now discuss the important issue of parametrizing these entities. The parametrization of the geometry is given as follows. Consider the (4×4) displacement in Eq. (1.3), i.e.

$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} \in \mathbb{SE}(3). \tag{5.6}$$

The natural local parametrization of $\mathbf{T} \in \mathbb{SE}(3)$ is through its related Lie algebra $\mathfrak{se}(3)$, whose coordinates are here denoted by $\mathbf{v} \in \mathbb{R}^6$, i.e. $\mathbf{T} = \mathbf{T}(\mathbf{v})$. As discussed in Subsection 1.1.2, the mechanism for passing information from the Lie algebra to the Lie group is the exponential mapping. An important difference in the calibrated setting relatively to the uncalibrated one concerns the cheirality constraint (the Euclidean parallax has always a strictly positive value), i.e. $(z^*)^{-1} > 0$ for the entire imaged scene. In order to obtain improved results, this constraint should also be enforced within the resolution of the system. Surprisingly, none of existing direct approaches have exploited this constraint. Here, we propose to parametrize the set of geometric parameters by

$$\mathbf{z}_g^{\mathrm{c}} = \{\mathbf{v}, \mathbf{y}\}, \tag{5.7}$$

such that $\boldsymbol{\lambda}_z(\mathbf{y}) = \exp(\mathbf{y})$, using a real-valued $\dim(\boldsymbol{\lambda}_z)$-vector $\mathbf{y}$. The choice of the exponential function is also motivated by the fact that $\exp(x) > 0,\ \forall x \in \mathbb{R}$.

Remark 5.1. By using the proposed parametrization $\boldsymbol{\lambda}_z(\mathbf{y})$ we enforce, within the optimization procedure, that the scene is always in front of the camera, i.e. $(z^*)^{-1} > 0$, $\forall i$.

As for the photometric parameters $\mathbf{h}$ (5.5), its parametrization is given by

$$\mathbf{z}_h = \{\boldsymbol{\Gamma}, \boldsymbol{\beta}\} \tag{5.8}$$

with $\boldsymbol{\Gamma} = \{\gamma_{kj}\}$.

Then, for real-time systems, only a local non-linear optimization procedure can be employed. To this end, an initial estimate $\widehat{\mathbf{g}}^{\mathrm{c}}$ and $\widehat{\mathbf{h}}$ sufficiently close to the true solution is needed. Thus, the model (4.11) is transformed into:

$$\mathcal{I}'\big(\mathbf{g}^{\mathrm{c}}(\widetilde{\mathbf{z}}_g^{\mathrm{c}})\circ\widehat{\mathbf{g}}^{\mathrm{c}}, \mathbf{h}(\widetilde{\mathbf{z}}_h)\circ\widehat{\mathbf{h}}, \mathbf{p}^*\big) = \mathcal{S}\big(\widetilde{\boldsymbol{\Gamma}}\circ\widehat{\boldsymbol{\Gamma}}, \mathbf{p}^*\big) \bullet \mathcal{I}\big(\mathbf{w}(\mathbf{g}^{\mathrm{c}}(\widetilde{\mathbf{z}}_g^{\mathrm{c}})\circ\widehat{\mathbf{g}}^{\mathrm{c}}, \mathbf{p}^*)\big) + \widetilde{\boldsymbol{\beta}}\circ\widehat{\boldsymbol{\beta}}, \tag{5.9}$$

where the symbol '$\circ$' refers to the related composition rule (see Subsection 3.3). Therefore, the visual SLAM problem can indeed be formulated as a calibrated direct image registration problem:

$$\min_{\widetilde{\mathbf{z}}^{\mathrm{c}} = \{\widetilde{\mathbf{z}}_g^{\mathrm{c}}, \widetilde{\mathbf{z}}_h\}} \ \frac{1}{2} \sum_i \ \big[\mathcal{I}'\big(\mathbf{x}^{\mathrm{c}}(\widetilde{\mathbf{z}}^{\mathrm{c}}) \circ \widehat{\mathbf{x}}^{\mathrm{c}}, \mathbf{p}_i^*\big) - \mathcal{I}^*(\mathbf{p}_i^*)\big]^2, \tag{5.10}$$

with $\mathbf{x}^{\mathrm{c}} = \{\mathbf{g}^{\mathrm{c}}, \mathbf{h}\}$ and its respective parametrization $\mathbf{z}^{\mathrm{c}} = \{\mathbf{z}_g^{\mathrm{c}}, \mathbf{z}_h\}$. The framework presented in Subsections 3.3 and 3.3.2 can then be applied to solve this optimization problem efficiently and with nice convergence properties.

Remark 5.2. Another important aspect of this formulation concerns the enforcement of the rigidity constraint of the scene, also within the optimization procedure, since all regions share the same incremental motion parameters.

Particular case (The efficient direct visual SLAM). Besides the use of an efficient optimization method, computational efficiency of the generic visual SLAM approach (5.10) can also be improved by modeling all surfaces with only first-order approximations (Szeliski and Torr, 1998), i.e. as planar surfaces. In this case, the scene is described by a set of patches, where each patch defines a plane. Hence, their structures are estimated independently of each other, leading to a sparse Jacobian matrix. It can be noted that normal vector and Euclidean parallax (i.e. the inverse of scene depths) are related by a simple transformation (1.30):

$$(z^*)^{-1} = \mathbf{n}_d^{*\top}\,\mathbf{K}^{-1}\mathbf{p}^*, \tag{5.11}$$

with

$$\mathbf{n}_d^* = (d^*)^{-1}\mathbf{n}^* = \|\mathbf{n}_d^*\|\,\mathbf{n}^*. \tag{5.12}$$

In other terms, the relation between both representations is simply given by

$$\mathbf{n}_d^* = \mathbf{M}\,\big[\,(z_1^*)^{-1}, (z_2^*)^{-1}, (z_3^*)^{-1}\,\big]^\top, \tag{5.13}$$

with

$$\mathbf{M} = \mathbf{K}^\top\,\big[\,\mathbf{p}_1^*,\,\mathbf{p}_2^*,\,\mathbf{p}_3^*\,\big]^{-\top} \in \mathbb{R}^{3\times3}, \tag{5.14}$$

using the image points of, for example, three vertices of each patch. Therefore, the same previously described direct registration framework can be applied.

In all case, for the purposes of extensive motions, new information has to be adequately inserted into the system. A procedure to perform this subtask is described next.

5.2.3 Selection, insertion and rejection of image regions

Selection of image regions

Despite the impressive computing power to date, in a real-time setting the entire image cannot in general be considered for processing. Therefore, an adequate selection of image regions is performed in this work. Indeed, we select a set of non-overlapping image patches according to an appropriate score. For direct methods, high scores should reflect strong image gradient along different directions.

Let the image region $\mathcal{R}^* \subset \mathcal{I}^*$ be a $(w\times w)$ matrix containing pixel intensities. Then, obtain a suitable gradient-based image $\mathcal{G}^*$ from $\mathcal{I}^*$. Given $\mathcal{G}^*$, a score image $\mathcal{S}^*$ can be defined as the sum of all values of $\mathcal{G}^*$ within a $(w \times w)$ block centered at every pixel. A second criterion to be considered, possibly with a different weight, is based on the quantity of local extrema of $\mathcal{G}^*$ (denoted $\mathcal{E}^*$) within each block. This may prevent the system from assigning high scores on single peaks, which would define patches with the same drawbacks as regions defined around standard interest points (e.g. Harris corners). The neighborhood of an isolated point may not contain enough information to constrain all degrees-of-freedom. Other criteria are also possible, e.g. the degree of spread of the regions around the image, but those two above have experimentally shown to be sufficient.

Hence, all needed block operations to adequately select image regions are efficiently performed by a convolution (denoted by the symbol '$\otimes$') with the $(w\times w)$ kernel $\mathcal{K}_w = \mathbf{1}$:

$$\begin{aligned}\mathcal{S}^* &= \xi_1\,\mathcal{G}^* \otimes \mathcal{K}_w + \xi_2\,\mathcal{E}^* \otimes \mathcal{K}_w \qquad (5.15)\\ &= (\xi_1\,\mathcal{G}^* + \xi_2\,\mathcal{E}^*) \otimes \mathcal{K}_w. \qquad (5.16)\end{aligned}$$

Typical weights are $\xi_1 = \|\mathcal{G}^* \otimes \mathcal{K}_w\|^{-1}$ and $\xi_2 = 1$. The resulting $\mathcal{S}^*$ contains the scores which are sorted, without any absolute thresholds on the strengths to be tuned. The amount of regions (defined around each score) considered for further processing depends only on the available computing resources.

Insertion of image regions

Given that regions may leave the field-of-view due to camera motion, or eventually be rejected from the optimization, the system must be able to insert new regions whenever computing resources are available. The initialization of new regions follows the natural way of specialization: we start by the most generic stratum to the most specialized one. In other words, first we characterize each new region in the projective space. Using this knowledge and of the recovered inter-frame displacement, we can obtain its best possible Euclidean structure until that moment.

This algorithm is detailed as follows. Let the current image be indexed by 'τ'. New regions can be selected in this image according to the procedure described in Subsection 5.2.3. Denote this image $\mathcal{I}^*_\tau$ since it contains the reference template of these particular regions. For the sake of simplicity but without loss of generality, consider that all surfaces are modeled in the sequel using only first-order approximations, i.e. as locally planar surfaces. In other words, let us focus here on the efficient direct visual SLAM formulation previously described. Then:

1. When a new image is available, apply the uncalibrated direct image registration method described in Chapter 4 using that surface approximation. More formally, obtain the projective homography and the lighting variations that best align each j-th newly selected region:

$$\big\{\widehat{\mathbf{G}}_j, \widehat{\gamma}_j, \widehat{\beta}_j\big\} = \operatorname*{arg\,min}_{\substack{\mathbf{G}_j \in \mathbb{SL}(3)\\ \gamma_j, \beta_j \in \mathbb{R}}} \frac{1}{2}\sum_i \big[\gamma_j\, \mathcal{I}\big(\mathbf{w}(\mathbf{G}_j, \mathbf{p}^*_{ij})\big) + \beta_j - \mathcal{I}^*_\tau(\mathbf{p}^*_{ij})\big]^2. \qquad (5.17)$$

   Since each region is treated independently, we have $8 + 2$ parameters to be recovered per region (for the gray-level case). This procedure may be initialized by, for example, a correlation measure;

2. Determine the scaled normal vector relative to the frame where the region was first viewed (i.e. corresponding to $\mathcal{I}^*_\tau$) using the closed-form solution:

$$\widehat{\mathbf{n}}^*_{d\,j} = \frac{\big(\alpha\, \mathbf{K}^{-1}\, \widehat{\mathbf{G}}_j\, \mathbf{K} - \widehat{\mathbf{R}}_\tau\big)^\top \widehat{\mathbf{t}}_\tau}{\big\|\widehat{\mathbf{t}}_\tau\big\|^2}, \qquad (5.18)$$

   with the obtained $\widehat{\mathbf{G}}_j$ in Step 1 and the optimal relative displacement $\widehat{\mathbf{T}}_\tau$ from the running visual SLAM system. See Appendix A.1 for both the necessary and sufficient geometric conditions, as well as for the computation of the normalizing factor $\alpha \in \mathbb{R}$;

3. An iterative refinement may then be conducted using the calibrated direct image registration method described by Eq. (5.10), but using only the structure as optimization variable. That is, with only 3 parameters to be recovered per region. All other parameters are used but are kept constant in Eq. (5.9).

If the j-th new region is not declared as an outlier, it is ready to be exploited from the next image. To this end, the photo-generative model expressed in (5.9) can adequately incorporate each new relative reference frame by multiplying the global $\widehat{\mathbf{T}} = \widehat{\mathbf{T}}_0$ on the right by the inverse of the relative $\mathbf{T}_j = {}^{\tau}\mathbf{T}_0$, i.e. $\widehat{\mathbf{T}}\mathbf{T}_j^{-1}$.

This insertion algorithm is intrinsically different from existing direct ones. For example, aside from being sensitive to variable lighting, the method proposed in (Molton et al., 2004) does not take into account all available knowledge to initialize $\widehat{\mathbf{n}}^*_{d\,j}$ (it uses a “best guess”). This may lead to convergence problems. Furthermore, differently from (Jin et al., 2003) where new regions are back-projected to the global reference frame, we avoid altering the original information by adequately incorporating them in (5.9). This possibility is also an attractive characteristic of the proposed SLAM formulation.

Rejection of image regions

Within direct methods, outliers correspond to regions that do not fit the models. For example, regions related to independently moving objects are considered as outliers within our proposed direct image registration method. Surface discontinuities and occluding boundaries can also be viewed as outliers. Hence, they must be detected and discarded by the algorithm.

To this end, two meaningful metrics are used to evaluate the j-th image region $\mathcal{R}^*_j$: a photometric measure as well as a geometric one. The photometric measure is defined directly from our cost function in (5.10) as

$$\varepsilon_j^2(\widehat{\mathbf{x}}^{\mathrm{c}}) = \frac{1}{\mathrm{card}(\mathcal{R}^*_j)} \sum_i d_{ij}^2(\widehat{\mathbf{x}}^{\mathrm{c}}), \tag{5.19}$$

where card$(\cdot)$ denotes the cardinality of the set. Notice that the illumination variations have already been compensated in such a measure. The geometric measure is the side ratio between the current and the previously warped region. That is, if a template significantly shrinks or elongates in at least one direction, this may signify insufficient content for constraining all the parameters (and thus, can be discarded).

As a remark, whilst (5.19) is evaluated after obtaining the optimal solution to the registration problem, the geometric measure can be evaluated within the iterations, provided that the region has been adequately initialized (see Subsection 3.3.2). This may prevent such regions from perturbing the solution.

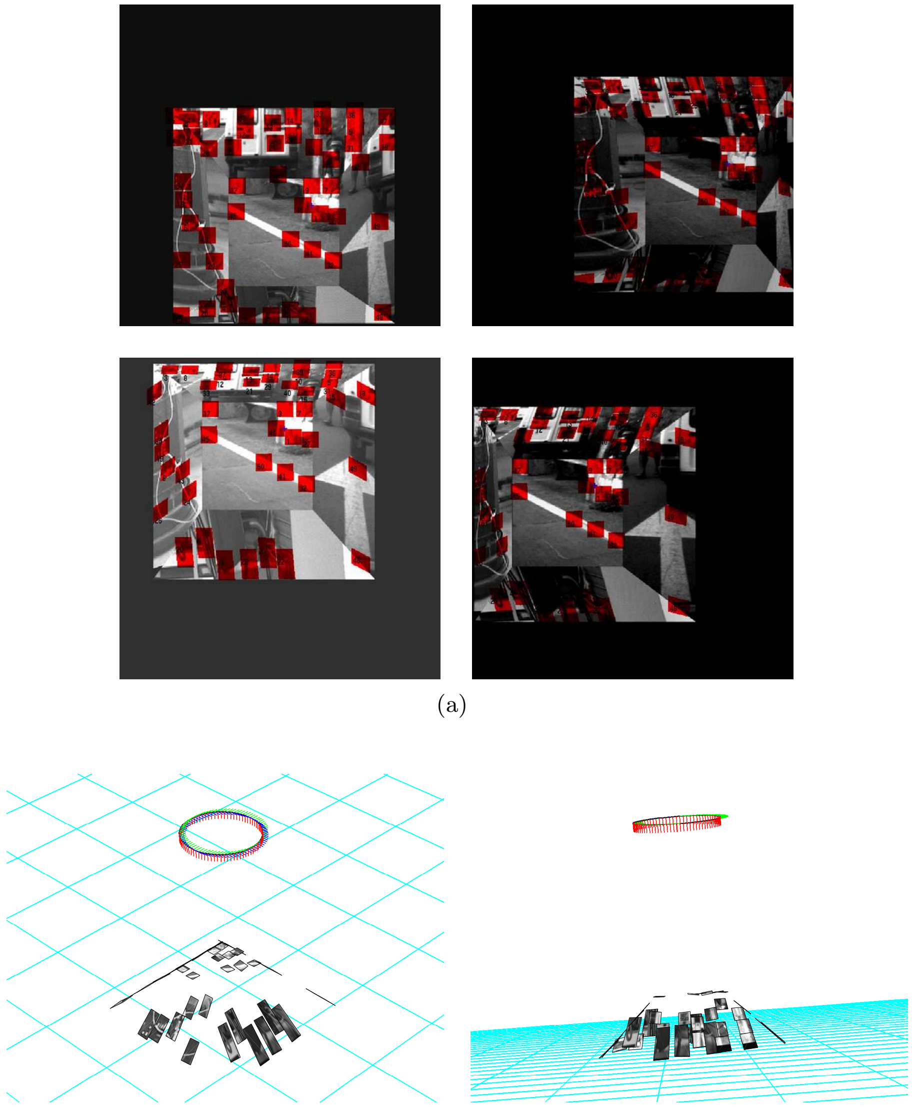
5.3 Experimental results

In order to validate the algorithm and to assess its performance, we have tested it with both synthetic and real-world images. In all cases, trivial initial conditions are used: $\widehat{\mathbf{T}}^{(0)} = \mathbf{I}_4, \widehat{\alpha}_j^{(0)} = 1, \widehat{\beta}_j^{(0)} = 0, \widehat{\mathbf{n}}_{d\,j}^{*(0)} = [0,0,1]^\top, \forall j$. The photometric error is here measured by its RMS (5.19). The j-th region is declared as an outlier if either $\varepsilon_j > 20$ or if its geometric error is over 50%. The RMS of the image noise is considered to be of 0.6 level of gray-scale. Moreover, we emphasize that no other sensory device than a single camera is used.

5.3.1 Synthetic data

The 'Pyramid' sequence. A synthetic scene was constructed so that a ground truth is available. It is composed of four planes disposed in pyramidal form, and cut by another plane on its top. In order to simulate realistic situations as closely as possible, textured images were mapped onto the planes. Then, a sequence of images was generated by displacing the camera while varying the illumination conditions. With respect to the trajectory, the camera performs a circular motion. The objective is twofold. First, returning the camera to the starting pose offers an important benchmark for SLAM algorithms. Second, this aims to show that past observations de facto contribute, within the proposed incremental technique, to build and maintain a coherent description of the structure and motion. With respect to the lighting variations, they are created by applying an $\alpha^{(k)}$ which linearly changes the image intensities up to 50% of its original value, and a $\beta^{(k)}$ which varies sinusoidally with amplitude of 50 levels of gray-scale.

We have then compared our approach (see some SLAM results in Fig. 5.2), which started with 50 regions of size 21×21 pixels, with traditional methods as well as with a direct method. With regard to standard methods, we used SIFT keypoints (1025 matches were initially found), and the sub-pixel Harris detector along with a Zero-mean Normalized Cross-Correlation with mutual consistency check for matching these latter points (235 were initially matched). Other than the initial ones, no features or regions are initialized here. Moreover, there is a relevant difference about how both point correspondences are established along the sequence. While keypoints are matched between the first (reference) and the current images, the latter had to be made between successive images (i.e. had to be tracked). In all cases, corresponding features were fed into a RANSAC procedure (typically 300 trials) with the state-of-the-art 5-point algorithm (Nistér, 2003) for robustly recovering the pose. This corresponds to a standard feature-based framework where a two-image reconstruction is considered and a non-planar scene is assumed (because of the 5-point algorithm). The comparisons are depicted in Fig. 5.3, where those strategies are respectively referred to as S+R+5P and H+ZNCC+R+5P. Since the scale factor is supposed to be unknown, the translation error is measured by the angle between the actual and the recovered translation directions, i.e. $\arccos\big(\mathbf{t}^\top \hat{\mathbf{t}} / (\|\mathbf{t}\| \|\hat{\mathbf{t}}\|)\big)$. Notice that, despite exploiting many more features, the standard techniques obtain relatively larger errors, especially for large displacements (i.e. middle of the loop) and significant lighting changes. In addition, the results show an increas-



(a)

(b)

Figure 5.2. (a) Excerpts from the 81-frame ‘Pyramid’ sequence superimposed with the regions registered (in red) by using the proposed approach. Observe the successful rejection of regions that do not fit the models (notably in the junctions of planes). (b) Reconstructed structure and motion (represented by frames) seen from different viewpoints. Final pose drift is of less than 0.001% of the total amount of translation, and of 0.091° for the rotation.

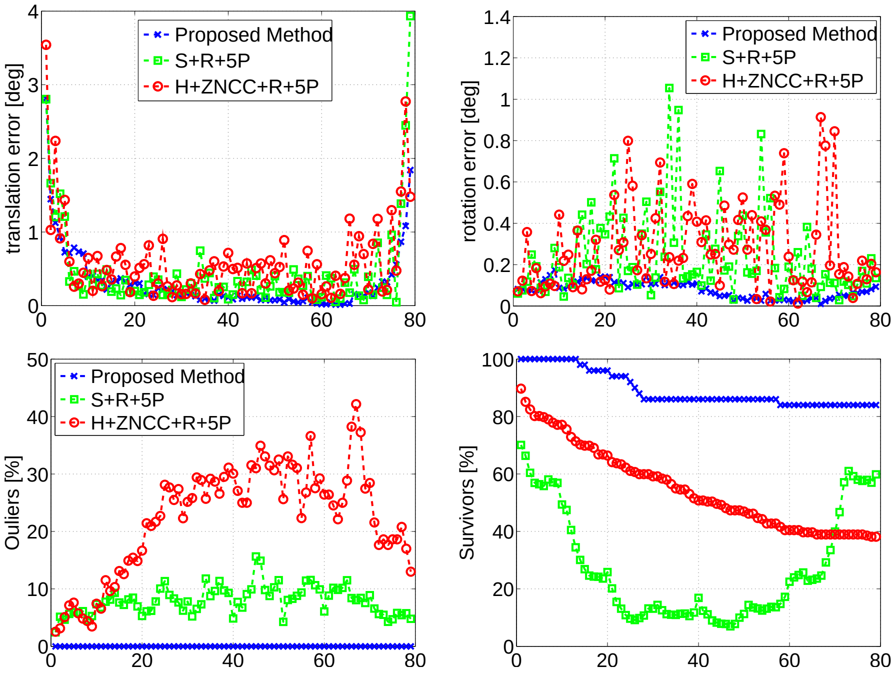


Figure 5.3. Results obtained from the proposed approach and traditional methods for the 'Pyramid' sequence. (Top) Errors in the recovered motion. Relatively larger errors were obtained from traditional methods. (Bottom) Percentage concerning the exploited regions and features. The notion of an outlier is made uniform here by using the same threshold for both features and any pixel of a region.

ing percentage of outliers, and a rapidly decreasing number of corresponding features. Therefore, to avoid an early failure, those methods certainly require a more frequent replacement of features. As a remark, despite their relative inferior accuracy, feature-based methods can have a larger domain of convergence and thus, may be used as a bootstrap to our technique (as discussed in Subsection 3.3.2). For the requested accuracy, the proposed approach performed along the sequence a median of 7 iterations, returned a median photometric error of 9.84 levels of gray-scale, and used a median of 10.4% of each (500×500) image. For this sequence where perfect camera's intrinsic parameters are available, the proposed method realized a drift between the original and final pose (since the camera returns to the starting pose) of less than 0.001% of the total amount of translation, and of 0.091° for the rotation. This shows that precise results with minimal drift are obtained.

With respect to existing direct methods, we have made a comparison with (Jin et al., 2003). Given that the displacements (motion and illumination) were not very small, which violate their assumptions, that algorithm failed at the beginning of the sequence. Our solution is able to deal with larger inter-frame displacements of the objects. The method proposed in (Molton et al., 2004) could not be applied since the scene is supposed to be unknown, and it is not possible to alter the environment (it needs a known target for the initialization).

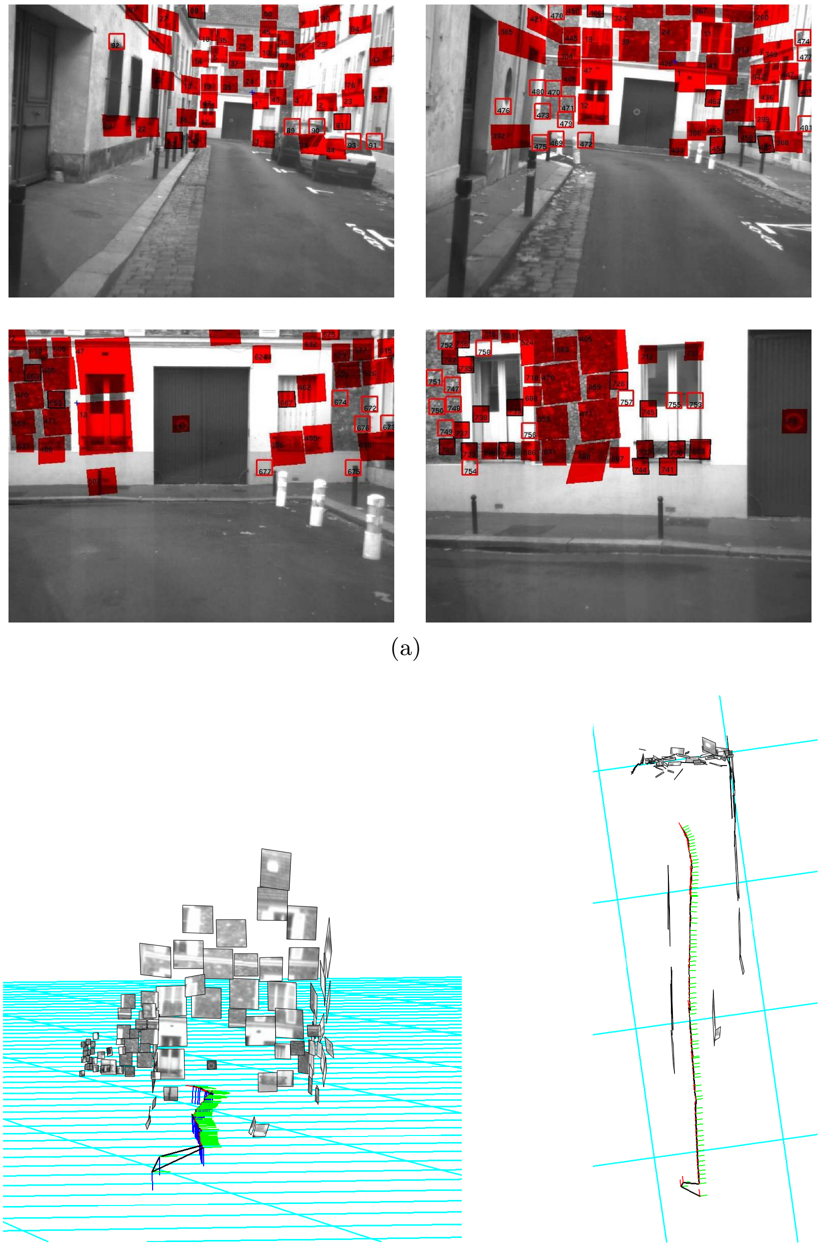
5.3.2 Real data

The 'Hangar' sequence. The application of the proposed technique to this outdoor sequence (see Fig. 5.1) also has a twofold objective. First, it aims at offering a didactic overview of the method, especially concerning the insertion of new information (the second region). Second, it shows its degree of robustness to different kinds of noise, e.g. shaking motion, image blur, etc. Very importantly, although we model the scene as a collection of planar regions, some occluding non-planar objects have appeared throughout the sequence, e.g. the tree in Fig. 5.1(a). These disturbances have not significantly perturbed the estimation process since they carry substantially less information compared to other parts of the patches.

For the requested accuracy, the approach performed along the sequence a median of 5 iterations, and returned a median photometric error of 13.37 levels of gray-scale. The recovered angle between the two walls is of 89.7°, using a median of 22.59% of each (320×240) image. This geometric measure is also an important benchmark for evaluating the technique (considering that these walls are truly perpendicular), since pose and structure are intimately tied together. The total displacement of the camera is of approximately 50 m, and the images were captured by a hand-held camcorder at 25 Hz.

The 'Canyon' sequence. We also run the proposed algorithm on a representative urban sequence, captured at approximately 12 Hz. It is also a challenging sequence in the sense that large inter-frame displacements are carried out, the objects are disposed at very different distances from the camera, and because there exists a significant change in scale. Furthermore, it corresponds to a typical urban scenario where cameras can be of particular importance for localization: narrow streets. In this case, positions from GPS may not be available or not sufficiently reliable. The obtained results are shown in Fig. 5.4, where the visual SLAM is successfully performed. The starting image was chosen such that the dominant plane is further away from the initial camera pose, compared to (Silveira et al., 2007). This choice aims to show the limitation of the optimization approach, which is local by nature. Notice that in the beginning of the task, despite the fact that the regions are effectively aligned in the images, the recovered motion and structure are not coherent with the true ones (see first camera poses in Fig. 5.4). This means that the algorithm got wedged in a local minimum. Thanks to the solution proposed in Subsection 3.3.2, this minimum is adequately treated and the correct parameters are subsequently obtained.

For the requested accuracy, the approach performed along the sequence a median of 12 iterations, returned a median photometric error of 10.77 levels of gray-scale, used a median of 34 image regions of size 31×31 pixels (at the time they are selected), and exploited a median of 17.01% of each (760×578) image. The total displacement of the camera is of approximately 60 m.



(a)

(b)

Figure 5.4. (a) Excerpts from the 81-frame ‘Canyon’ sequence superimposed with the regions registered (in red) by using the proposed approach. Observe the significant change in scale between first and last image. (b) Reconstructed structure and motion seen from different viewpoints. Recovered poses are represented by frames, and only the most stable regions are shown. See the parallelism and/or perpendicularity between most of them.

The 'Round-about' sequence. This sequence is also illustrative since other different types of noise are present, e.g. pedestrians and moving vehicles. It has been captured at approximately 12 Hz by a camera-mounted car, where the path length measured by Google Earth is of approximately 150 m. See Fig. 5.5 for a satellite image of this challenging scenario.

Nevertheless, the technique automatically copes with such outliers. Excerpts from this sequence and the obtained visual SLAM results can be seen in Fig. 5.6. We can observe that coherent motion and structure are recovered. For the requested accuracy, the approach performed along the sequence a median of 10 iterations, returned a median photometric error of 11.37 levels of gray-scale, used a median of 37 image regions of size 31 × 31 pixels (at the time they are selected), and exploited a median of 10.84% of each (760 × 578) image.

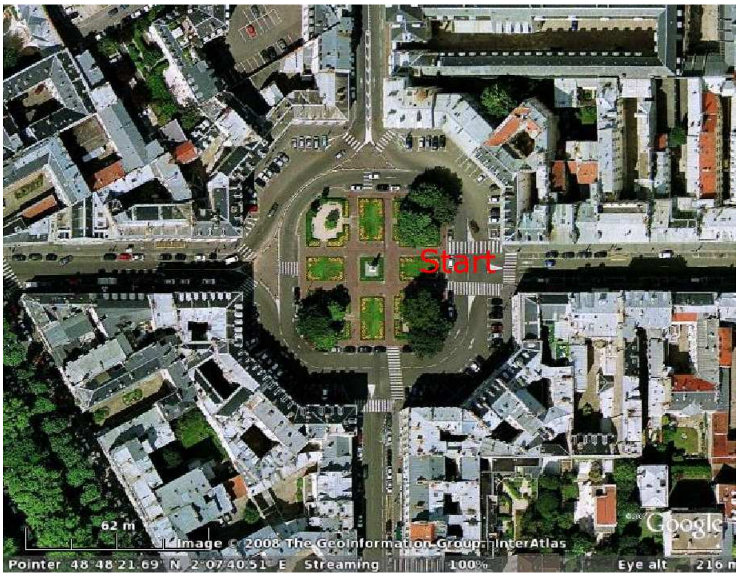
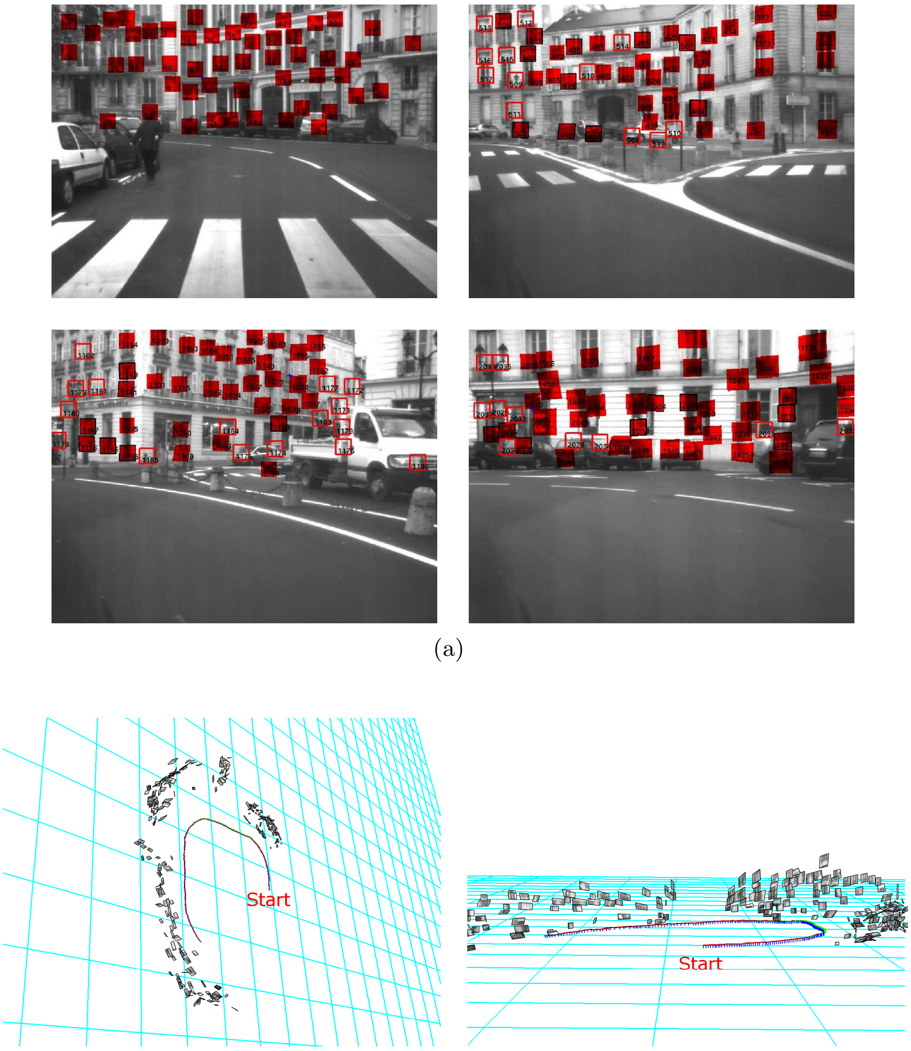

Start
62 m
Image © 2008 The GeoInformation Group, InterAtlas
©2007 Google
Pointer 48°48'21.69" N 2°07'40.51" E
Streaming 100%
Eye alt 216 m

Figure 5.5. Satellite image of the scenario where the 'Round-about' sequence was captured. See the corresponding results in Fig. 5.6.



(a)

(b)

Figure 5.6. (a) Excerpts from the 230-frame ‘Round-about’ sequence superimposed with the regions aligned (in red) by using the proposed approach. Observe the presence of a pedestrian in the first image, and of a moving car in the third image. (b) Reconstructed structure and motion. Recovered poses are represented by very small frames. The path length is of approximately 150 m.

5.4 Summary

This chapter proposes a different formulation of the vision-based SLAM problem. The technique is based on image registration using appropriate motion, structure and illumination parameters, without first having to find feature correspondences.

The strengths concern its high accuracy and absence of feature extraction process. Additionally, we have proved that standard methods need to add more frequently new features to track, especially under either significant lighting variations or lengthy camera displacements. Hence, the proposed method reduces the drift by maintaining for longer the estimation of the displacement with respect to the same reference frame. This is an important issue, especially in monocular frameworks. On the other hand, in order to be tractable in real-time, we use a local optimization procedure to obtain the related parameters. Alternatives to avoid getting trapped in local minima are discussed in Subsection 3.3.2.

Another important research topic regards loop closure, which was not an objective here. Nevertheless, we believe that the proposed direct technique is promising since existing ones (which have a smaller convergence domain) have already performed this task. Other future works may also focus on merging/growing regions with similar structure, which may lead to more stable and faster estimates.

Part III

Direct control from visual data

Chapter 6

Vision-based control given a reference image

This chapter addresses the teach-by-showing approach to visual servoing robots. In other terms, it is considered here the framework where the desired pose to be reached by the camera (i.e. the control objective) is specified by means of a reference image.

In addition to the standard definition of uncalibrated vision-based control, let us refer to it here as the case where none of the following information is either required a priori or estimated on-line so as to control the full 6 dofs: (*i*) precise camera's and/or robot's calibration parameters; and (*ii*) metric information about the observed target. Existing techniques which fall into this class (and are able to control all 6 dofs) require prior knowledge about the object's shape and/or the camera's motion.

Here, we propose a new uncalibrated visual servoing technique that does not require or estimate any of the above information. The technique exploits the projective parameters estimated from our generic uncalibrated direct visual tracking method presented in Chapter 4. Both theoretical and simulation results are provided to demonstrate that visual servoing can indeed be highly accurate and robust despite unknown objects and unknown imaging conditions. This naturally encompasses the case of color images.

6.1 Related work

Visual servoing consists in controlling the motion of a robot through the feedback of images (Chaumette and Hutchinson, 2006). Visually servoed systems can then be viewed as regulators of an appropriate control error, also referred to as task function (Samson et al., 1990). This chapter considers the task functions that can be constructed from the current and the reference images, i.e. the teach-by-showing approach. Further, we focus on techniques that both do not use metric information about the observed target and take control of all 6 dofs.

Existing methods which fall into this class require prior knowledge of the object's shape and/or of the camera's motion. Indeed, any method which solely relies on the Essential matrix, e.g. (Basri et al., 1999), though not using an explicit metric model of the object, requires a non-planar target as well as a sufficient amount of translation to be carried out in order to avoid the degeneracies (Faugeras et al., 2001; Hartley and Zisserman, 2000). With regard to the technique proposed in (Vargas and Malis, 2005; Benhimane and Malis, 2006a), although also not requiring metric information, they are designed for planar targets only. These are the only existing uncalibrated techniques (in the sense we have defined in the preamble of this chapter) available to date.

We remark that even for image-based visual servoing approaches, e.g. (Weiss and Anderson, 1987; Espiau et al., 1992; Silveira et al., 2002), minimal metric knowledge (the depth distribution) is necessary to provide a stable control law (Malis and Rives, 2003). The 2.5D visual servoing strategy (Malis et al., 1999) was then proposed to enlarge that domain of stability. However, it requires a coarse metric estimate of the normal vector of the planar target, in order to decide between the two possible solutions of the reconstruction (Faugeras and Lustman, 1988). Another alternative to augment the domain of convergence is to perform a path planning (Mezouar and Chaumette, 2002). However, this latter method also requires this coarse metric estimate of the planar target.

In this work, we propose a new visual servoing technique that does not either require or estimate any metric information about the observed target. The proposed control error as well as the control law are fully based on image measurements. We provide the theoretical proof that the control law ensures local asymptotic stability for the servoing, if the camera is perfectly calibrated. Nevertheless, the results presented in Subsection 6.3 demonstrate that the system is largely robust to errors in the intrinsics camera parameters. The control error proposed in this chapter generalizes the one presented in (Benhimane and Malis, 2006a), which is designed for planar surfaces only. In fact, the proposed method is independent of the object's shape and of the camera's motion as well. Moreover, another generalization concerns well-established techniques such as (Malis and Chaumette, 2002), which performs a partial Euclidean reconstruction. Our projective formulation also naturally encompasses this latter solution. The theoretical proof of all of these statements is provided in Appendix B. In addition to these attractive generalizations, other improvements are achieved. The proposed control error is locally isomorphic to the camera pose, and is also injective around the equilibrium for the entire domain of rotations. This local isomorphism represents an extremely important property of the system. Indeed, although within a limited region, it avoids the situation where the control error is null and the camera pose is different from the desired one. This is a well-known issue in standard 2D vision-based control (Chaumette, 1998). The fact of being injective around the equilibrium for the entire domain of rotations guarantees that the null control error corresponds to a unique camera pose in the large, e.g. even for a rotation of 180°. The theoretical proof of these properties is also given in Appendix B. Furthermore, another important strength of our control error is that it allows for simple, smooth, and physically valid path planning. This procedure can considerably enlarge the domain of convergence of the visual servoing.

As throughout this thesis, we directly exploit here the intensity value of the pixels. Therefore, higher accuracy for the servoing is achieved since noise is not introduced (there is no feature extraction process) and much more information is exploited. Again, only few works have been conducted in this spirit. To our knowledge, only the methods proposed in (Benhimane and Malis, 2006a; Kallem et al., 2007; Collewet et al., 2008) directly exploit pixel intensities. The first one (Benhimane and Malis, 2006a), which has been described above, require prior knowledge of either the object's shape or the camera's motion. The visual servoing technique proposed in (Kallem et al., 2007) is restricted to controlling only a subset of all six degrees-of-freedom, whereas the method proposed in (Collewet et al., 2008) requires approximate metric information about the observed target. Thus, these latter two strategies are not uncalibrated (in the sense we have defined in the preamble of this chapter).

The proposed method uses the same geometric parameters from our uncalibrated direct visual tracking method presented in Chapter 4. Indeed, we strongly believe that both vision and control aspects are intrinsically coupled processes, and are treated here as such. This represents a rupture of paradigm with respect to the vast majority of existing visual servoing techniques to date, where feature extraction process and control computation are formulated separately. Although conceptually appealing, this latter uncoupled, feature-based framework presents some relevant drawbacks. For example, global constraints are not easy to embed into feature correspondence algorithms (Jin et al., 2003), such as the fact that large portions of the scene move with a coherent rigid motion, or that the appearance changes due to motion of the scene relative to the lights. Attempts to impose these constraints are usually performed a posteriori within this framework. On the other hand, we have previously shown in this thesis (see Part II) that the rigidity of the scene, and the robustness to lighting changes, can both be effectively incorporated within direct methods.

Besides the theoretical analysis, the proposed approach is also validated using synthetic data. Various simulations are reported with objects of different shapes, large initial displacements, large errors in the camera's intrinsic parameters, as well as for challenging illumination conditions.

6.2 The direct visual servoing

Consider a rigid object of unknown shape being imaged under unknown conditions by an n-channel camera, $n \geq 1$. Of course, $n = 1$ refers to a gray-level image. The uncalibrated direct image registration method presented in Chapter 4 simultaneously recovers the optimal set of parameters $\mathbf{x}^{\mathrm{u}} = \{\mathbf{g}^{\mathrm{u}}, \mathbf{h}\}$. The photometric parameters $\mathbf{h}$ are estimated so as to achieve effective robustness to generic illumination changes, even in color images. On the other hand, the geometric parameters $\mathbf{g}^{\mathrm{u}} = \{\mathbf{G}, \mathbf{e}, \rho^*\}$ can be used for visual servoing purposes. In this subsection a new technique is proposed for positioning the camera-mounted robot to the reference (desired) pose, which is defined by means of a reference image. In the visual servoing community, this framework is commonly known as teach-by-showing.

6.2.1 Generic control error and some properties

The generic control error uses the projective information $\mathbf{g}^{\mathrm{u}}$ as follows. Firstly, normalized entities are obtained:

$$\mathbf{m}^{*\prime} = \mathbf{K}^{-1}\mathbf{p}^{*} \tag{6.1}$$

$$\mathbf{e}^{\prime} = \mathbf{K}^{-1}\mathbf{e} \tag{6.2}$$

$$\mathbf{H} = \mathbf{K}^{-1}\mathbf{G}\,\mathbf{K}, \tag{6.3}$$

where $\mathbf{p}^{*}$ corresponds to a chosen point (not necessarily an interest point), also called control point, and $\mathbf{K}$ gathers the camera's intrinsic parameters (see Subsection 1.2.1). We remark that, in the most general case of controlling all 6 dofs (our objective), at least a coarse estimate of $\mathbf{K}$ is *always* needed because the camera displaces in the real Euclidean space, whilst the observations are defined in the projective space.

Before stating the proposed control error and some of its properties, let us give first some definitions about the used terms.

Definition 6.1. A *"projective axis of rotation"* $\boldsymbol{\mu} \in \mathbb{R}^3$ is defined here as

$$[\boldsymbol{\mu}]_\times = \frac{1}{2}\big(\mathbf{H} - \mathbf{H}^\top\big). \tag{6.4}$$

This axis of rotation does not necessarily have unit norm.

Definition 6.2. A *"projective angle of rotation"* $\vartheta \in \,]-\pi, \pi]$ is defined here as

$$\vartheta = \begin{cases} \mathrm{real}\big(\arcsin(\|\boldsymbol{\mu}\|)\big), & \text{if } \mathrm{tr}(\mathbf{H}) \geq 1, \\ \pi - \mathrm{real}\big(\arcsin(\|\boldsymbol{\mu}\|)\big), & \text{otherwise}, \end{cases} \tag{6.5}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. The function $\mathrm{real}(\cdot)$ is needed since ϑ is a real-valued scalar and $\boldsymbol{\mu}$ does not necessarily have unit norm.

From those definitions above we are now able to present our proposed control error.

Definition 6.3. The *control error* $\boldsymbol{\varepsilon}^{\mathrm{u}}$ is defined here as

$$\boldsymbol{\varepsilon}^{\mathrm{u}} = \begin{bmatrix} \boldsymbol{\varepsilon}^{\mathrm{u}}_{\nu} \\ \boldsymbol{\varepsilon}^{\mathrm{u}}_{\omega} \end{bmatrix} = \begin{bmatrix} (\mathbf{H} - \mathbf{I})\,\mathbf{m}^{*\prime} + \rho^{*}\mathbf{e}^{\prime} \\ \vartheta \dfrac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|} \end{bmatrix}. \tag{6.6}$$

Before presenting a relevant property of this control error, let us state important relations involving it.

Lemma 6.1 (Control error and camera pose). *The proposed control error (6.6) can be expressed as a function of the camera pose. That is, as a function of the rotation* $\mathbf{R} = \exp([\theta\mathbf{u}]_\times) \in \mathbb{SO}(3)$*, and of the translation* $\mathbf{t} \in \mathbb{R}^3$ *between the current frame and the reference one. See Appendix A.2 for these relations.*

Proof. The proof of these relations is presented in Appendix B.2. ∎

In Lemma 6.1, we have preferred to refer the reader to Appendix A.2 (instead of fully presenting it here) since they are only used in theoretical demonstrations, i.e. they are not used for servoing the system.

Theorem 6.1 (Local isomorphism). *The proposed control error* (6.6) *is locally isomorphic to the camera pose. Moreover, it is injective around the equilibrium for the largest possible domain of rotations, since only* $\theta = 0$ *is mapped to by* $\boldsymbol{\varepsilon}^{\mathrm{u}}_{\omega} = \mathbf{0}$.

Proof. The proof of this isomorphism is presented in Appendix B.3, which uses the results from Lemma 6.1. ∎

Remark 6.1. A very important note about the control error defined in (6.6) is that it is constructed without measuring or requiring any metric information about the object.

Remark 6.2. Since the epipole is computed in the tracking process, we could use it solely to construct a decoupled translation error, e.g. by defining

$$\boldsymbol{\varepsilon}^{\mathrm{u}}_{\nu} = \mathbf{e}', \tag{6.7}$$

instead of that defined in (6.6). The translation error (6.7) is decoupled from the rotation motion since $\mathbf{e}' = \mathbf{K}^{-1}\mathbf{e} \propto \mathbf{K}^{-1}\mathbf{K}\,\mathbf{t} \propto \mathbf{t}$, using Eq. (1.28). However, if the object is planar then one is not sure if the recovered epipole corresponds to the true solution because, in this case, more than one admissible solution do exist (Faugeras and Lustman, 1988). Nevertheless, the coupling present in (6.6) is not a major concern to the stability of the system because a straightforward path planning can be performed (see Subsection 6.2.3).

Remark 6.3. In addition to the possible modification (6.7), we could also have defined the rotational error differently, such as (Benhimane and Malis, 2006a):

$$\begin{aligned} [\boldsymbol{\varepsilon}^{\mathrm{u}}_{\omega}]_{\times} &= 2[\boldsymbol{\mu}]_{\times} &(6.8)\\ &= \mathbf{H} - \mathbf{H}^{\top}, &(6.9)\end{aligned}$$

in our general, unified framework. However, remarkable improvements are achieved through $\boldsymbol{\varepsilon}^{\mathrm{u}}_{\omega}$ as defined in (6.6), which are stated in Corollary 6.1.

Corollary 6.1 (Generality and improvements). *The proposed control error* (6.6) *is a generalization of the one presented in (Benhimane and Malis, 2006a) for coping with objects of arbitrary shape and camera motion. Moreover, the proposed control error allows for a straightforward path planning (it will be demonstrated in Subsection 6.2.3). Furthermore, our projective formulation naturally encompasses the hybrid control error proposed in (Malis and Chaumette, 2002), which requires a coarse metric estimate of the normal vector of the planar target.*

Proof. The proof of these statements is presented in Appendix B.4. ∎

6.2.2 Control law and stability analysis

Consider a camera-mounted holonomic robot or an omnidirectional mobile robot.

Definition 6.4. Let $\mathbf{v} = \left[\boldsymbol{\nu}^\top, \boldsymbol{\omega}^\top\right]^\top \in \mathbb{R}^6$ represent the translational and rotational velocities of the camera. The *control law*

$$\mathbf{v} = \boldsymbol{\Lambda}\,\boldsymbol{\varepsilon}^{\mathrm{u}}, \tag{6.10}$$

with the control gain

$$\begin{aligned} \boldsymbol{\Lambda} &= \operatorname{diag}(\lambda_\nu \mathbf{I}_3, \lambda_\omega \mathbf{I}_3) && (6.11) \\ &= \begin{bmatrix} \lambda_\nu\,\mathbf{I}_3 & \mathbf{0} \\ \mathbf{0} & \lambda_\omega\,\mathbf{I}_3 \end{bmatrix}, \quad \lambda_\nu, \lambda_\omega > 0, && (6.12) \end{aligned}$$

uses the control error $\boldsymbol{\varepsilon}^{\mathrm{u}} = \left[\boldsymbol{\varepsilon}_\nu^{\mathrm{u}\top}, \boldsymbol{\varepsilon}_\omega^{\mathrm{u}\top}\right]^\top$ defined in (6.6) as feedback to compute the input signals (i.e. the velocities).

Theorem 6.2 (Local stability). *The proposed control law* (6.10) *ensures local asymptotic stability provided that the control point* $\mathbf{p}^*$ (6.1) *is chosen such that its parallax relative to the dominant plane*[1] *of the object is sufficiently small.*

Proof. The proof is presented in Appendix B.5, which uses the results from Lemma 6.1. ■

Corollary 6.2 (Parallax condition). *There always exists a point* $\mathbf{p}^*$ *which has zero parallax (and thus can be chosen as the control point) since, in the formulation, the dominant plane always crosses the object. Therefore, the closed-loop system is always locally asymptotically stable.*

Although one could always choose the control point such that its parallax is zero, for robustness reasons it is convenient to choose this image point close to the center of the object. This reduces the possibility of this point getting out of the field-of-view due to measurement noise or camera calibration errors.

Remark 6.4. It can be noted that the control law (6.10) has a positive sign. This is due to how the control error (6.6) is defined, which exploits the geometric set of parameters $\mathbf{g}^{\mathrm{u}} = \{\mathbf{G}, \mathbf{e}, \rho^*\}$. This set is computed using the uncalibrated registration method presented in Chapter 4, which in turn defines these entities as a mapping from the reference frame to the current frame, e.g. $\mathbf{G} = {}^c\mathbf{G}_r$.

[1] if the object is not planar, this plane is virtual.

6.2.3 Path planning

Although the technique is robust to large camera calibration errors, it is desirable that the trajectory of the control point in the image be as closely as possible to a straight line. With this, a large domain of convergence for the visual servoing is achieved since we enforce that at least such a point always remains in the image. To this end, instead of regulating $\boldsymbol{\varepsilon}^{\mathrm{u}}(t) \to \mathbf{0}$, an appropriate path tracking $\boldsymbol{\varepsilon}^{\mathrm{u}}(t) \to \boldsymbol{\varepsilon}^{\mathrm{u}*}(t)$ can be performed, where the latter represents a desired time-varying signal. This requires then the definition of a time-varying control error.

Definition 6.5. Path tracking is accomplished by regulating a *time-varying control error*

$$\boldsymbol{\varepsilon}^{\mathrm{u}\prime}(t) = \boldsymbol{\varepsilon}^{\mathrm{u}}(t) - \boldsymbol{\varepsilon}^{\mathrm{u}*}(t), \quad \forall t \in [0, T], \tag{6.13}$$

given a desired time-varying signal

$$\boldsymbol{\varepsilon}^{\mathrm{u}*}(t) = \left[\boldsymbol{\varepsilon}_{\nu}^{\mathrm{u}*\top}(t),\, \boldsymbol{\varepsilon}_{\omega}^{\mathrm{u}*\top}(t)\right]^{\top}. \tag{6.14}$$

The strategy presented in this subsection is different from (Mezouar and Chaumette, 2002), where this latter is composed of three phases and requires a coarse metric estimate of the normal vector of the planar target. Indeed, a simple strategy is shown to be sufficient to attain our purposes, i.e. without requiring any metric information and being independent of the object's shape and of the camera's motion. This is possible owing to the properties of the proposed control error (6.6):

- we need to plan the trajectory of only one point, which means that physically valid camera situations are always specified;
- the projective axis-angle parametrization already provides for a smooth trajectory;
- given the local isomorphism, there is no singularity or local minima in a region.

Particular case (Linear path). An example of special interest consists in specifying (6.14) as a linear desired path such that $\boldsymbol{\varepsilon}^{\mathrm{u}*}(0) = \boldsymbol{\varepsilon}^{\mathrm{u}}(0)$ and $\boldsymbol{\varepsilon}^{\mathrm{u}*}(T) = \mathbf{0}$, i.e.

$$\boldsymbol{\varepsilon}^{\mathrm{u}*}(t) = \boldsymbol{\varepsilon}^{\mathrm{u}*}(0) + \left(\boldsymbol{\varepsilon}^{\mathrm{u}*}(T) - \boldsymbol{\varepsilon}^{\mathrm{u}*}(0)\right)\frac{t}{T} \tag{6.15}$$

$$= \boldsymbol{\varepsilon}^{\mathrm{u}*}(0)\left(1 - \frac{t}{T}\right). \tag{6.16}$$

In all cases, motivated by the fact (see the results from Lemma 6.1) that

$$\boldsymbol{\varepsilon}_{\omega}^{\mathrm{u}} = \vartheta\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|} \to \theta\mathbf{u} \qquad \text{as} \qquad \mathbf{t} \to \mathbf{0}, \tag{6.17}$$

which means that if $\mathbf{t} = \mathbf{0}$ then geodesic rotations will be induced, the rotational part of (6.14) can be slightly changed into

$$\boldsymbol{\varepsilon}^{\mathrm{u}*}_{\omega}(t) = \boldsymbol{\varepsilon}^{\mathrm{u}}_{\omega}(t-1)\left(1 - \frac{t}{T}\right), \tag{6.18}$$

where the notation $\boldsymbol{\varepsilon}^{\mathrm{u}}_{\omega}(t-1)$ refers to the last value of $\boldsymbol{\varepsilon}^{\mathrm{u}}_{\omega}$.

Definition 6.6. considering a motionless target and willing to regulate the time-varying control error (6.13), the *control law* (6.10) is transformed into

$$\mathbf{v} = \mathbf{\Lambda}(t)\,\boldsymbol{\varepsilon}^{\mathrm{u}\prime}(t) + \frac{\partial \boldsymbol{\varepsilon}^{\mathrm{u}*}(t)}{\partial t}, \tag{6.19}$$

where the feed-forward term $\partial \boldsymbol{\varepsilon}^{\mathrm{u}*}(t)/\partial t$ allows compensation of the tracking error and

$$\mathbf{\Lambda}(t) = \mathrm{diag}\big(\lambda_\nu\,\mathbf{I}_3, \lambda_\omega(t)\,\mathbf{I}_3\big) \tag{6.20}$$

$$= \begin{bmatrix} \lambda\,\mathbf{I}_3 & \mathbf{0} \\ \mathbf{0} & \lambda\exp\big(-\gamma\|\boldsymbol{\varepsilon}^{\mathrm{u}}_{\nu}(t)\|\big)\,\mathbf{I}_3 \end{bmatrix}, \quad \lambda, \gamma > 0, \tag{6.21}$$

is an adaptive gain matrix also motivated by (6.17): $\lambda_\omega(t)$ is small for large $\|\boldsymbol{\varepsilon}^{\mathrm{u}}_{\nu}(t)\|$, and $\lambda_\omega(t) \to \lambda$ as $\|\boldsymbol{\varepsilon}^{\mathrm{u}}_{\nu}(t)\| \to 0$.

6.3 Results

This section reports various simulations using the proposed direct visual servoing technique. We use the word “direct” to express that there is no feature extraction process, and that the control error and control law are computed using only image measurements. For their computation, all pixels within an area of interest (also called reference template) are exploited. For all results presented here, this area is delimited by a red grid.

The considered visual servoing task consists in positioning the camera with respect to a rigid object independently of its shape. To this end, a reference image is stored at the reference (desired) pose. After displacing the camera to another pose, the objective is then to drive the camera back to this desired pose. The reader is referred to Subsection 3.3.2 for a discussion on how to initialize the direct image registration procedure between the reference and initial images. A possible technique is to perform a global optimization procedure only for this first image. Nonetheless, if those two images present a sufficient amount of overlapping, then no special initialization routine is necessary. It should be noted that the visual servoing technique is also independent of the displacement between the initial and desired poses, i.e. it may comprise pure translations, pure rotations or a combination of both.

To have a real ground truth, we constructed synthetic objects of different shapes and, in order to simulate realistic situations as closely as possible, textured images are mapped onto them. For all images shown here, blue marks are used to depict the motion of the control point in the image plane, whilst its planned path is projected in green. This latter is typically composed of 1000 points with an adaptive gain (6.21) with $\lambda = \gamma = 10$.

Despite the fact that only a local stability proof with a calibrated camera has been established in this chapter, the results shown here demonstrate that the technique can cope with large initial displacements, as well as is robust to large errors on the camera parameters.

A planar object. It is shown in Fig. 6.1 that the proposed method can cope with planar objects (a-priori unknown by the technique) including thus, the existing technique that is designed for this particular surface (see Corollary 6.1). The control law is stable: both translational and rotational velocities converge to zero. At the convergence, the visual information coincides with the reference image, and the camera is positioned at the reference pose very accurately. Errors less than 1 mm for the translation and less than 0.1° for the rotation are simultaneously achieved. Figure 6.1 also shows the evolution of the Cartesian displacement (in meters and in degrees) and of the input signals along the entire visual servoing task. The blue marks in the reference image depict the straight line performed by the control point, as desired.

A hyperbolic paraboloid. In this second set of results, we set up a challenging scenario: the object is an hyperbolic paraboloid (the horse's saddle); the used focal lengths are almost the double of the true ones, i.e. instead of $\alpha_u = \alpha_v = 500$ pixels, we used $\widehat{\alpha}_u = 900$ and $\widehat{\alpha}_v = 800$; and a large initial displacement is carried out. The visual servoing technique successfully performs the positioning task, despite all of these large perturbations. See Fig. 6.2 for the corresponding results. This demonstrates that the proposed technique also copes with non-planar objects, that the strategy is robust to large errors in the camera's internal parameters, and that the servoing has a very large domain of convergence.

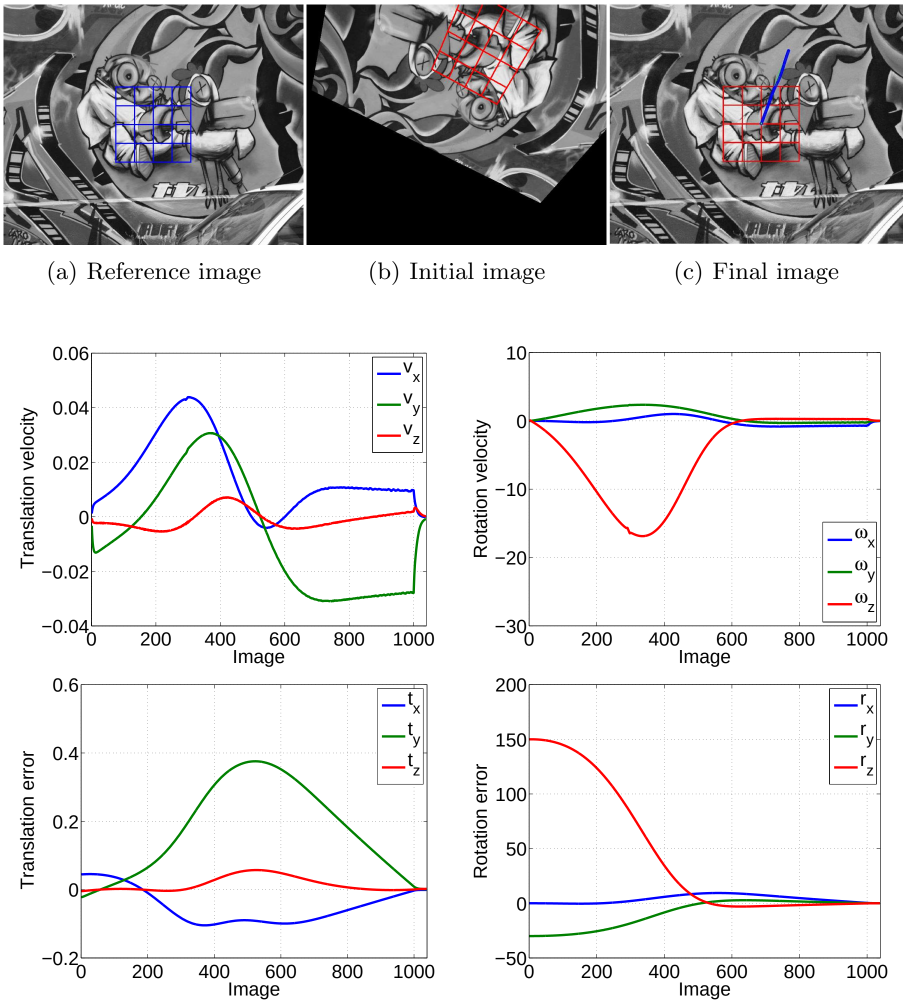
The particular task of rotation of 180°**.** Yet another improvement concerns a positioning task for a rotation of $\theta = 180°$, which cannot be performed by existing uncalibrated vision-based control methods. The results achieved by the proposed technique for this case are presented in Fig. 6.3, and without any path planning, i.e. using Eq. (6.10) with $\lambda_\nu = \lambda_\omega = 1$. Since a pure rotation is given, the result will be the same regardless the shape of the object and so we use a sphere.

A sphere imaged by a color camera under generic lighting variations. In this last set of results, a sphere is also used, although this knowledge is not a-priori provided to the algorithm. Here, we created illuminants, which simulate a white noise, and rendered the images considering an infinite-bandwidth color camera. Specular reflections are due to a light source rigidly attached to the virtual camera. It points towards the object with a slightly different direction with respect to the camera's optical axis. This latter simulates a misalignment between the camera and a carrying light. We have then applied a fully coupling model of illumination changes within the image registration method presented in Chapter 4 so as to ensure robustness to those generic lighting variations. The corresponding surfaces are modeled through a discretization into blocks of size 50×50 pixels for computational efficiency.

The visual servoing results for this challenging scenario are shown in Fig. 6.4. The control law is stable: both translational and rotational velocities converge to zero. At the convergence, the camera is positioned at the desired pose very accurately. The norm of the final Cartesian error is around 1 mm for the translation, and $0.1°$ for the rotation. The remark here is that accuracy is obtained despite large specular reflections even at the final image (compare final image with the desired one). See Fig. 6.5 both for the synthetic reflection present in the image at the convergence, and for a particular surface related to the illumination changes, reconstructed by the image registration method.

6.4 Summary

This chapter proposes a new approach to visual servoing all 6 dofs of a robot, given a reference image. The approach does not require or estimate any metric information about the observed rigid object. Further, our general technique is independent of the object's shape and of the camera's motion. Thus, it does not rely on prior knowledge (leading to system flexibility), and ensures robustness to errors in the calibration parameters. Moreover, owing to the application of the direct image registration method proposed in Chapter 4 to recover the used parameters, high levels of accuracy for the positioning can be attained, whilst ensuring robustness to arbitrary illumination changes (even in color images). Finally, a very large domain of convergence for the servoing is obtained due to a (straightforward) path planning scheme. Hence, visual servoing tasks can be performed despite large initial displacements.



(a) Reference image (b) Initial image (c) Final image

Figure 6.1. Direct visual servoing with respect to a planar object (a-priori unknown) using an uncalibrated pinhole camera.

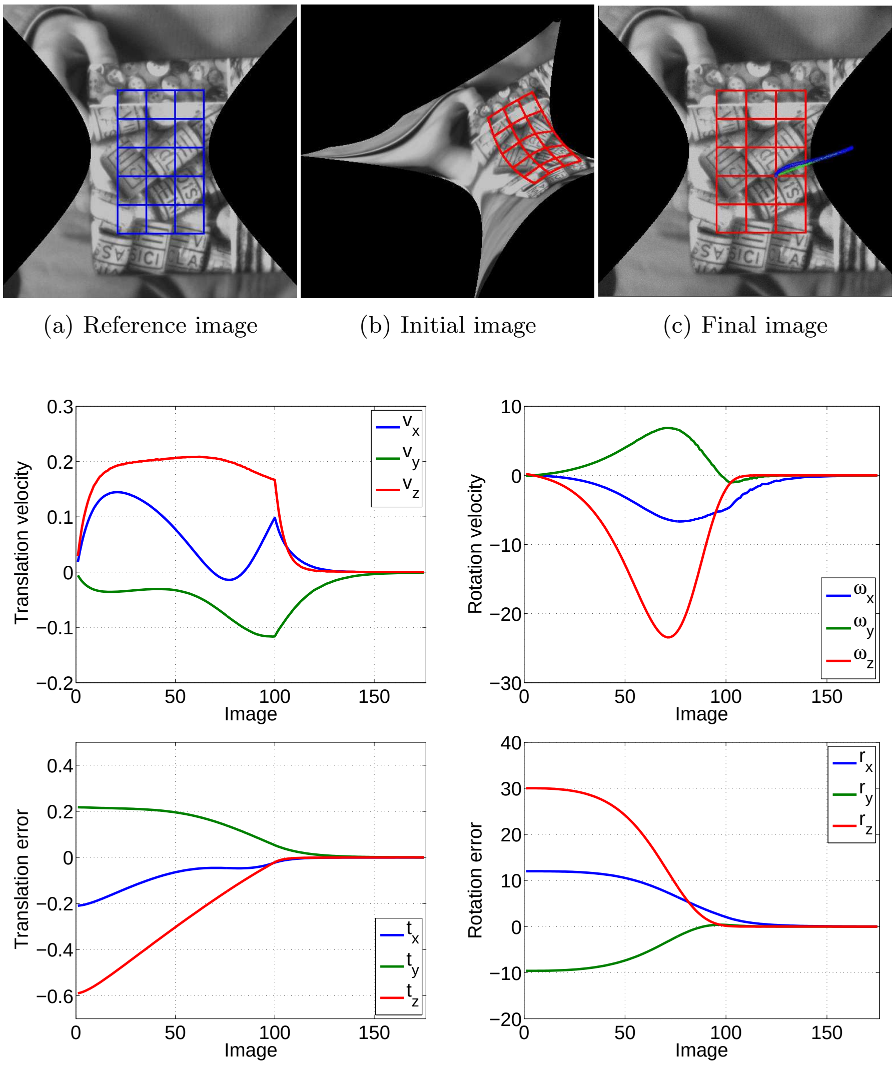
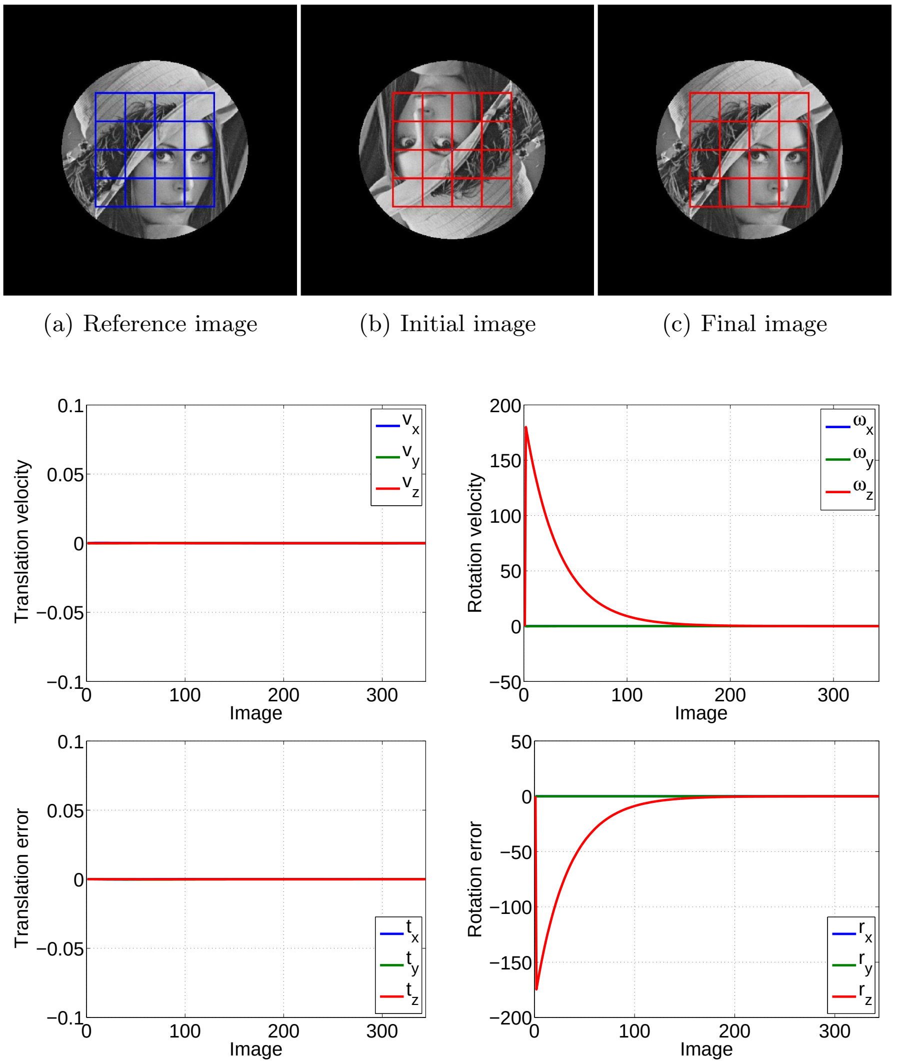

(a) Reference image (b) Initial image (c) Final image

Figure 6.2. Direct visual servoing with respect to a hyperbolic paraboloid (a-priori unknown) using an uncalibrated pinhole camera.

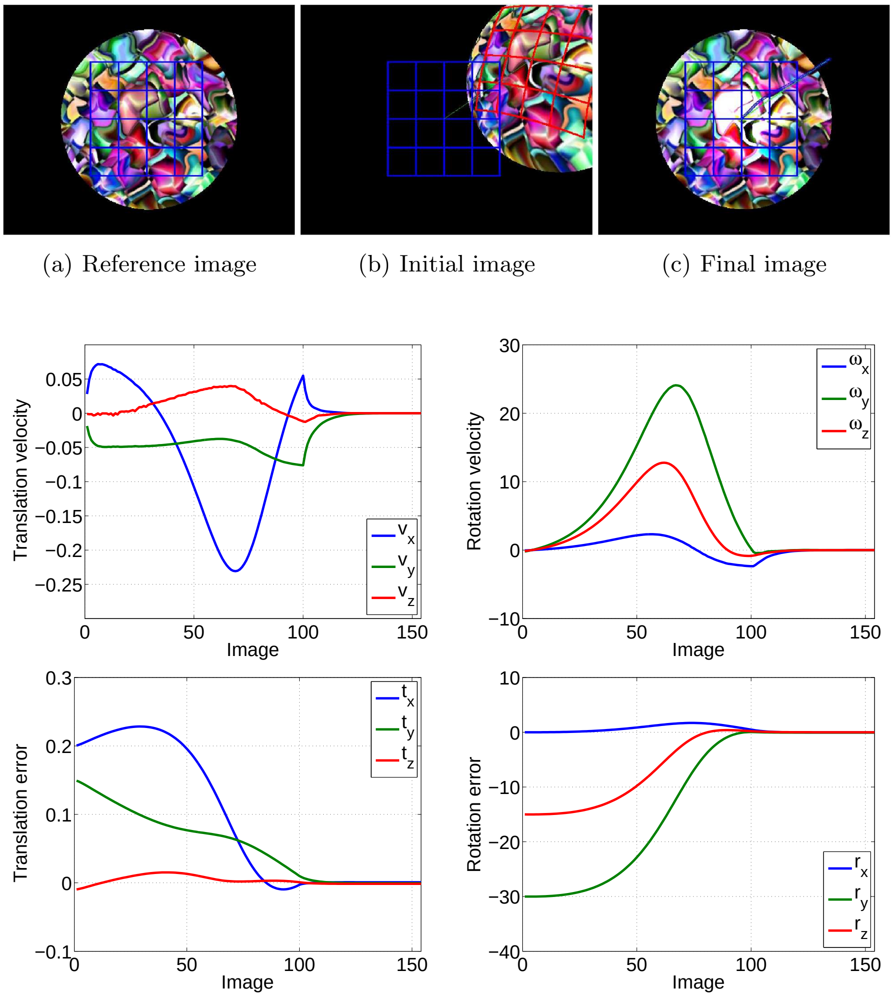


(a) Reference image

(b) Initial image

(c) Final image

Figure 6.3. Direct visual servoing for the particular task of rotation of 180° using an uncalibrated pinhole camera.



(a) Reference image (b) Initial image (c) Final image

Figure 6.4. Direct visual servoing with respect to a sphere (a-priori unknown) using an uncalibrated pinhole camera.

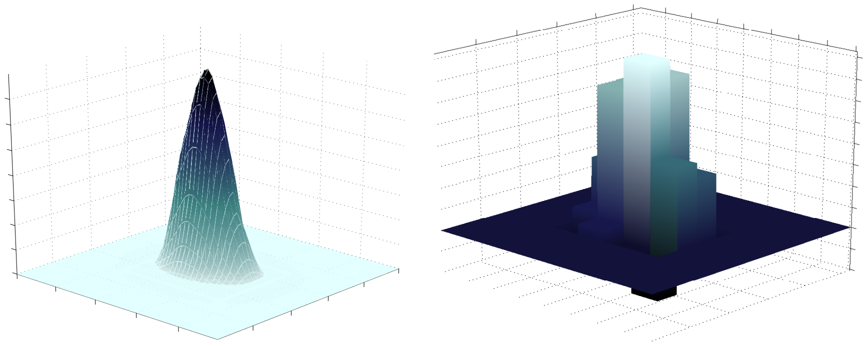


Figure 6.5. (Left) The synthetic specular reflection at the convergence for the visual servoing task showed in Fig. 6.4. (Right) A particular reconstructed surface ($\mathcal{S}_{22}$) to counterbalance the illumination changes.

Chapter 7

Vision-based control given a reference pose

This chapter proposes a vision-based control scheme where the reference pose is specified directly in the Euclidean space. Considering the case where the robot has never reached this reference pose, the corresponding reference image is hence not available. Thus, the visual servoing technique described in Chapter 6 cannot be applied (in fact, none that is based on the teach-by-showing approach). Furthermore, let the scene be a-priori unknown. In this case, standard pose reconstruction algorithms cannot be applied either. In effect, the proposed framework is well-suited to autonomously navigating mobile robots over extensive, unexplored scenes.

Given that the Euclidean space in this case corresponds to the space of the control error by definition, the camera's intrinsic parameters are needed. The proposed method is in fact based on the generic calibrated direct image registration presented in Chapter 5 to accurately recover the camera pose and then, to pursue our objective.

7.1 Related work

This chapter focuses on automatically driving a camera-mounted robot to a given desired Cartesian pose relatively to a given reference frame (i.e. coordinate system). See Fig. 7.1 for a graphical illustration. Since everything is relative, the reference frame is also defined by the user. That is, the desired pose can be specified relatively to a particular camera frame (e.g. the first frame), or even to a particular known object by attaching a frame to this latter. For example, the robot may be controlled to visually move in a particular direction with respect to its current pose. Therefore, standard 3D visual servoing strategies, for instance (Wilson et al., 1996; Thuilot et al., 2002), fall into this class of methods. However, these strategies require the prior knowledge of the object's (i.e. scene's) metric model.

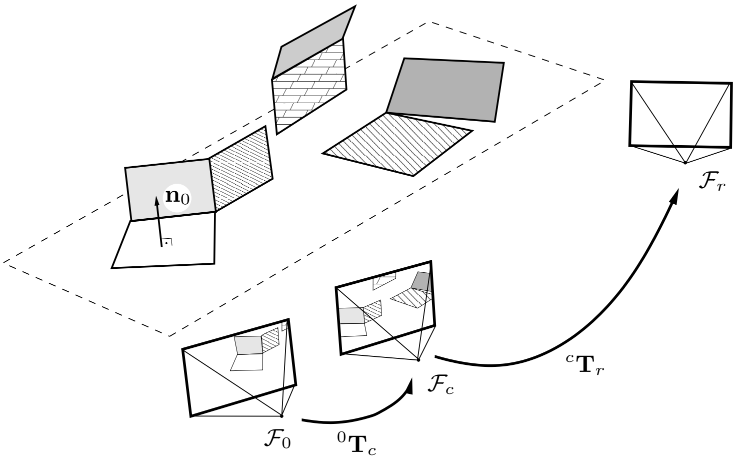


Figure 7.1. Main objective of the efficient E-3D visual servoing (explained in Subsection 7.2.2): to perform a vision-based navigation task where neither the desired image (corresponding to the given desired pose) nor the metric model of the scene are available a priori.

Very importantly, this chapter considers the case where the scene is a-priori unknown. Thus, standard model-based techniques for pose recovery cannot be applied. Furthermore, we consider navigation (or positioning) tasks where the given desired pose has never been reached by the robot beforehand. Therefore, the corresponding desired image is neither available nor can be rendered. This fact makes impossible to use 'metric model'-free visual servoing methods that are based on the teach-by-showing approach, such as the uncalibrated technique described in Chapter 6.

On one hand, the desired orientation can be fully specified and accurately tracked. On the other hand, if no other sensory device than a single camera is used, the translational part of the task is defined up to a scale factor (Rives, 2000). That is, only the specified direction of translation is ensured to be tracked with high accuracy. Therefore, we suppose that the scale factor is provided by, for example, an exteroceptive sensor or using an initialization pattern.

Given that no other sensory device than a single camera is used, the control problem at hand is closely related to an active monocular SLAM problem (see Chapter 5). Although the mapping does not necessarily have to be performed to recover the pose (by using an appropriate tensor, e.g. the Essential matrix), precision may be rapidly lost within monocular frameworks if it is not simultaneously carried out. This happens because important structural constraints, e.g. scene rigidity, are not effectively exploited in a long run. As a remark, the use of multiple cameras for pose recovery represents a different type of problem, as far as the baselines are sufficiently large with respect to the scene depths. It constitutes a different type of problem due to this important prior knowledge concerning the baselines. See for example the binocular system described in (Comport et al., 2007), or the trinocular one in (Saeedi et al., 2006). In these cases and under that baseline condition, visual odometry can indeed be sufficiently accurate despite not explicitly recovering the scene structure.

Nevertheless, the proposed approach is also different from existing monocular SLAM techniques. Firstly, the vast majority of existing methods do not control the robot. Whilst SLAM aims at a wide range of applications, e.g. a dense mapping of the environment, the specific objective here is to efficiently displace the camera from the starting pose to the given desired one. To avoid post-processing steps, and given that we can control the system at arbitrarily small motions, one should exploit the best suited information from the partially observed scene.

Additionally, the majority of visual SLAM techniques are feature-based. Although they may afford relatively larger motions of the object in the image, they inevitably introduce errors which are never corrected. Since we consider real-time vision-based control, we can suppose that the frame rate is sufficiently high such that only relatively small inter-frame displacements of the object are observed. Furthermore, the robustness to illumination changes is somewhat limited within feature extraction and matching procedures. On the other hand, the robustness to arbitrary lighting variations can be effectively incorporated within direct methods (see Chapter 2), amongst other structural constraints. Therefore, using all possible image information and avoiding the inherent difficulties of feature-based methods, the accuracy of direct pose reconstruction procedures is significantly improved.

7.2 The E-3D visual servoing

The E-3D visual servoing technique[1] is based on an appropriate visual SLAM approach so as to accurately and robustly estimate the camera pose. Thus, the generic calibrated system presented in Chapter 5 can be adapted to this particular task. The above-mentioned appropriateness not only refers to suitable transformation models and optimization methods, but also to which image information should be exploited by the system. Here, new information is inserted into (and thus, exploited by) the system only after assuring that it complies with the adopted models. This minimizes the probability of inducing discontinuities in the estimates, which is important when using them in feedback control loops. To this end, another key component of the proposed scheme consists in robustly identifying new rigidly attached, fitted objects as the robot displaces, since known ones may get out of the field-of-view. This will be demonstrated in Subsection 7.2.2 for a particular class of objects. Finally, once the optimal current camera pose is recovered, our control objective can be pursued.

In this way, the proposed scheme can be applied on large-scale scenes, i.e. for extensive navigation tasks. In fact, the (unavailable) corresponding desired image may not have anything in common with the initial one, but the desired Cartesian path can still be tracked precisely. Once again, let us denote the reference frame by either $\mathcal{F}_r$ or $\mathcal{F}^*$, and the current frame by either $\mathcal{F}_c$ or $\mathcal{F}$.

[1] E-3D is an acronym for Extended-3D.

7.2.1 Localization through direct image registration

Vision-based localization can be formulated as a calibrated direct image registration task (Cobzas and Sturm, 2005; Benhimane and Malis, 2006b). In this case, a metric model of the scene is required. To obtain a generic and robust localization technique, we can adapt the proposed transformation model (5.9) such that the optimization variables are composed of the illumination parameters

$$\mathbf{h} = \{\boldsymbol{\mathcal{S}}, \boldsymbol{\beta}\}, \tag{7.1}$$

and only a subset of the geometric parameters $\mathbf{g}^{\mathrm{c}} = \{\mathbf{R}, \mathbf{t}, (z^*)^{-1}\}$, i.e. only

$$\mathbf{g}^{\mathrm{c}\prime} = \{\mathbf{R}, \mathbf{t}\} = \mathbf{T} \in \mathbb{SE}(3). \tag{7.2}$$

The superscript ‘c’ written in (7.2) in this standard roman font denotes the calibrated case. The structure parameters

$$\mathbf{s}^* = \big\{(z^*)^{-1}\big\} \tag{7.3}$$

are supposed to be already identified (it will be discussed in the next subsection), and are required only to perform the warping, i.e. they are not optimization variables.

In respect to the parametrization, whilst $\mathbf{h}$ is again represented by

$$\mathbf{z}_h = \{\boldsymbol{\Gamma}, \boldsymbol{\beta}\} \tag{7.4}$$

with $\boldsymbol{\Gamma} = \{\gamma_{kj}\}$, the geometric parametrization is only the coordinates of the related Lie algebra $\mathfrak{se}(3)$, i.e.

$$\mathbf{z}_g^{\mathrm{c}\prime} = \mathbf{v}. \tag{7.5}$$

As discussed in Subsection 1.1.2, the mechanism for passing information from the Lie algebra to the Lie group is the exponential mapping.

In this case, the transformation model (5.9) becomes

$$\boldsymbol{\mathcal{I}}'\big(\mathbf{T}(\widetilde{\mathbf{v}})\,\widehat{\mathbf{T}}, \mathbf{s}^*, \mathbf{h}(\widetilde{\mathbf{z}}_h) \circ \widehat{\mathbf{h}}, \mathbf{p}^*\big) = \boldsymbol{\mathcal{S}}\big(\widetilde{\boldsymbol{\Gamma}} \circ \widehat{\boldsymbol{\Gamma}}, \mathbf{p}^*\big) \bullet \boldsymbol{\mathcal{I}}\big(\mathbf{w}(\mathbf{T}(\widetilde{\mathbf{v}})\,\widehat{\mathbf{T}}, \mathbf{s}^*, \mathbf{p}^*)\big) + \widetilde{\boldsymbol{\beta}} \circ \widehat{\boldsymbol{\beta}}, \tag{7.6}$$

where the symbol ‘$\circ$’ refers to the related composition rule (see Subsection 3.3) 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. Therefore, a robust and generic vision-based localization technique can be formulated as

$$\min_{\widetilde{\mathbf{z}}^{\mathrm{c}\prime} = \{\widetilde{\mathbf{v}}, \widetilde{\mathbf{z}}_h\}} \; \frac{1}{2} \sum_i \Big[\boldsymbol{\mathcal{I}}'\big(\mathbf{T}(\widetilde{\mathbf{v}})\,\widehat{\mathbf{T}}, \mathbf{s}^*, \mathbf{h}(\widetilde{\mathbf{z}}_h) \circ \widehat{\mathbf{h}}, \mathbf{p}_i^*\big) - \boldsymbol{\mathcal{I}}^*(\mathbf{p}_i^*)\Big]^2, \tag{7.7}$$

where the optimal $\widehat{\mathbf{T}} = {}^{c}\widehat{\mathbf{T}}_0$ (the superscript ‘c’ is written in italic) encodes the current camera pose relatively to the origin (the reference frame), since the input estimate represents the camera displacement from the origin until the preceding image (or the identity $\mathbf{I}_4$, at the start). Finally, the same optimization procedure presented in Subsections 3.3 and 3.3.2 can then be applied to solve this registration problem efficiently and with nice convergence properties.

As a remark, given that all objects share the same incremental camera motion, the rigidity of the scene is directly enforced in that formulation. This enforcement, along with the fact that all possible information is exploited, significantly increase the accuracy of the pose estimates. Moreover, robustness to generic illumination changes are ensured.

For generality, the formulation (7.7) has to be extended so that multiple objects $\{\mathbf{s}_j^*\}$ can be taken into consideration. Each object has its own reference template $\{\boldsymbol{\mathcal{I}}_j^*\}$, and may have been identified at different poses $\{\mathbf{T}_j\}$. The identification of a particular class of objects and their insertion into the full system are discussed below.

7.2.2 An efficient localization method

Besides the use of an efficient optimization method, computational efficiency of the proposed generic and robust localization technique (7.7) can also be improved by modeling all surfaces using only first-order approximations (Szeliski and Torr, 1998; Simon and Berger, 2002), i.e. as planar surfaces. To this end, a planar region detector is needed in order to both segment the regions in the image $\{\boldsymbol{\mathcal{I}}_j^*\}$ and characterize these regions in the Euclidean space $\{\mathbf{s}_j^* \equiv \mathbf{n}_{d\,j}^*\}$, where the scaled normal vector is given by

$$\mathbf{n}_d^* = (d^*)^{-1}\mathbf{n}^* = \|\mathbf{n}_d^*\|\,\mathbf{n}^*. \tag{7.8}$$

Identification of planar regions

The interest in finding planar regions in images is not new, and a number of different approaches is available in the literature. Many of existing methods rely on scene assumptions, e.g. presence of lines (Baillard and Zisserman, 1999; Simon and Berger, 2008) or perpendicularity assumptions (Dick et al., 2000). It is also possible to assume a particular configuration of the camera with respect to the scene. For example, having a car-mounted camera always pointing toward the road plane is a constraint that can help to reduce the complexity of the problem. However, these methods cannot be applied here, since we deal with unknown scenes. Another class of existing methods endeavors to perform a preliminary step of Euclidean scene reconstruction, e.g. (Okada et al., 2001). Nevertheless, these methods usually require several images to converge, and are in general too computationally intensive to be applied to real-time systems, such as visual-servoed systems.

To circumvent these shortcomings, a generic Planar Region Detector (PRD) is developed here by exploiting the two-view geometry. The proposed technique is based on an efficient voting procedure from the solution of a linear system. Let us first describe how this linear system is derived. Then, we will discuss the applied voting procedure. For the sake of generality, let $\mathbf{T} = \{\mathbf{R}, \mathbf{t}\}$ represent in this subsection the rigid displacement between any two views. Nevertheless, for the purposes of plane identification using a pair of images, the considered displacement usually involves the current frame and the frame (indexed by 'τ') from where the plane was first viewed, i.e. ${}^c\mathbf{T}_\tau$.

The linear system is constructed by exploiting the two-view epipolar geometry. Indeed, by injecting (1.30) into (1.27) allows for rewriting the equation that links the projection of the same 3D point onto $\boldsymbol{\mathcal{I}}$ and $\boldsymbol{\mathcal{I}}^*$ (i.e. $\boldsymbol{\mathcal{I}}^*_\tau$) as:

$$\mathbf{p}_i \propto \mathbf{K}\,\mathbf{R}\,\mathbf{K}^{-1}\,\mathbf{p}_i^* + \mathbf{K}\,\mathbf{t}\,\mathbf{x}^\top \mathbf{p}_i^*, \tag{7.9}$$

where

$$\mathbf{x} = \mathbf{K}^{-\top}\mathbf{n}_d^*, \tag{7.10}$$

and $\mathbf{K}$ gathers the camera's intrinsic parameters.

Definition 7.1. Pre-multiplying both members of (7.9) by $[\mathbf{p}_i]_\times$ and using the fact that $\mathbf{x}^\top\mathbf{p}_i^* = \mathbf{p}_i^{*\top}\mathbf{x}$, the *linear system* is finally obtained:

$$\mathbf{A}_i\,\mathbf{x} = \mathbf{b}_i, \tag{7.11}$$

with

$$\begin{cases} \mathbf{A}_i = [\mathbf{p}_i]_\times\,\mathbf{K}\,\mathbf{t}\,\mathbf{p}_i^{*\top} \\ \mathbf{b}_i = -[\mathbf{p}_i]_\times\,\mathbf{K}\,\mathbf{R}\,\mathbf{K}^{-1}\,\mathbf{p}_i^*. \end{cases} \tag{7.12}$$

However, matrix $\mathbf{A}_i \in \mathbb{R}^{3\times3}$ has maximum rank 1 since it can be seen as a product of two 3-vectors, i.e. as $\mathbf{A}_i = \mathbf{c}_i\,\mathbf{p}_i^{*\top}$, where $\mathbf{c}_i = [\mathbf{p}_i]_\times\,\mathbf{K}\,\mathbf{t}$. This is an obvious statement from a geometric point of view since at least 3 points are needed to constraint the 3 dofs of a plane (2 dofs for $\mathbf{n}^*$ and 1 dof for d^*). Hence, the parameters related to a plane is recovered by stacking three equations (7.11), one for each pair of corresponding points $\mathbf{p}_i \leftrightarrow \mathbf{p}_i^*$:

$$\bar{\mathbf{A}}\,\mathbf{x} = \bar{\mathbf{b}}, \tag{7.13}$$

with the augmented matrix $\bar{\mathbf{A}} = \left[\{\mathbf{A}_i\}_{i=1}^3\right] \in \mathbb{R}^{9\times3}$ and the augmented vector $\bar{\mathbf{b}} = \left[\{\mathbf{b}_i\}_{i=1}^3\right] \in \mathbb{R}^9$. The solution of such a rectangular linear system is obtained in the least-squares sense by solving its normal equations

$$\bar{\mathbf{A}}^\top\bar{\mathbf{A}}\,\mathbf{x} = \bar{\mathbf{A}}^\top\,\bar{\mathbf{b}}, \tag{7.14}$$

which is performed extremely fast given its low dimensionality. Furthermore, if noise is not too large then those 9 equations can be reduced to 6 by using only the first 2 equations of each $\mathbf{A}_i$. The linearly independent equation is either the first or the second one. Now, it is important to study in which conditions the solution (the vote) of such a system is unique.

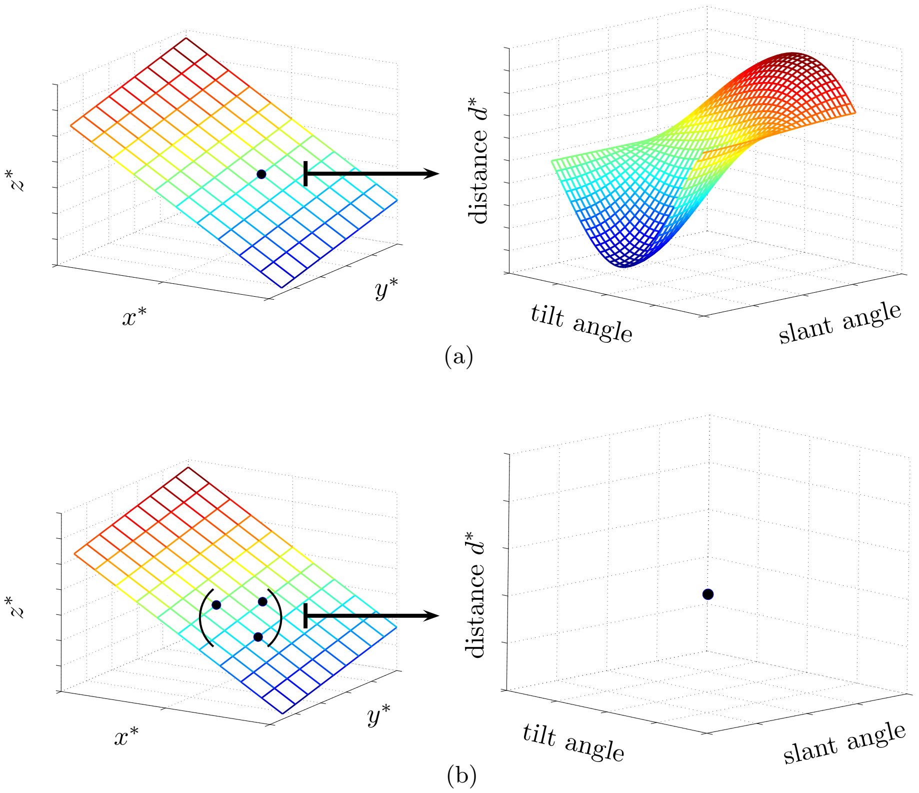
Lemma 7.1 (Existence and uniqueness). *The assembled linear system* (7.14) *from 3 pairs of corresponding points* $\mathbf{p}_i \leftrightarrow \mathbf{p}_i^*$ *is consistent and has a unique solution if:*

1. $\mathbf{t} \neq \mathbf{0}$*;*
2. *the 3 points are non-collinear.*

Proof. The proof is presented in Appendix B.6. ■

Voting procedures (e.g. the Hough Transform) are amongst the most important robust techniques in computer vision (Stewart, 1999). As it will be experimentally shown in Section 7.3, even if the set of camera parameters is miscalibrated (i.e. only an estimate is provided) and/or even if there exist mismatched corresponding points, it is still possible to cluster planar regions in the image. This robustness property is an attractive characteristic of the approach since it is able to tolerate large errors in its inputs. Additionally, multiple highly reliable planar regions can be partitioned in a global optimal sense (Meer, 2004), instead of producing only an inlier/outlier dichotomy as in RANSAC-based procedures (Fischler and Bolles, 1981).

A major difference between the used voting technique and the standard Hough Transform is related to the performed mapping. As in (Xu and Oja, 1993) for detecting lines, the entire (a-priori fixed) parameter space is also not voted here. The solution of the constructed linear system represents a single vote (see Fig. 7.2). Various advantages of this convergence mapping for robustly detecting planes are discussed in (Silveira et al., 2006a), e.g. reduction of both memory and computational complexities.



(a)

(b)

Figure 7.2. (a) Illustration of the divergence mapping performed by a standard Hough transform to detect planes. In this case, a point is mapped to an hypersurface. (b) In the case of a convergence mapping, a chosen triplet of points maps to a single point.

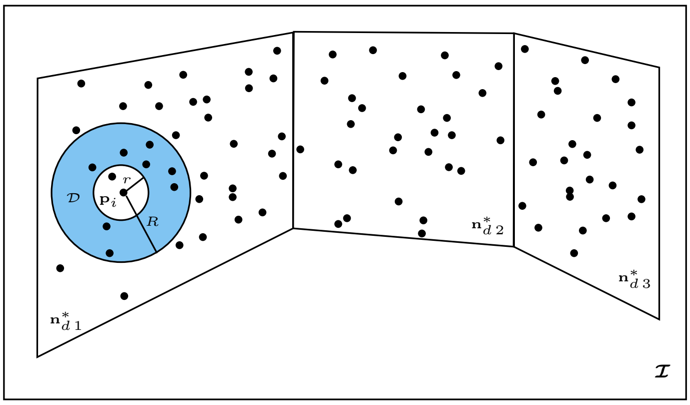


Figure 7.3. Illustration of a geometric-based local grouping in the image by using a disk $\mathcal{D}$ of radius $r, R > 0$ centered at $\mathbf{p}_i$. In the case of a photometric measure, the intensity of the pixel plays the role of devising the region.

Moreover, all possible combinations of three points are not necessarily voted. As in (Galambos et al., 1999) for detecting lines, a progressive procedure is performed here over triplets of points likely to be coplanar. Points likely to be coplanar are chosen here in function of a photo-geometric distance amongst them (see Fig. 7.3). This dynamic subdivision of the image also contributes to avoid clustering dominant (virtual) planes.

Thus, complexities are further reduced since a plane is clustered, and all of its points are removed from input data, as soon as the contents of a given accumulator permits such a decision. This decision involves checking if the number of votes is sufficiently large, together with a plane verification step (this latter will be described next). The number of votes of an accumulator is incremented every time a chosen triplet of points produces a vote, i.e. a solution of (7.14), that already exists according to an user-defined resolution. Otherwise, a new accumulator is created for the unmatched solution, without any boundaries on the parameter space. In this way, the method also features an infinite range. As a remark, the final precision of the algorithm is therefore an user input, since that resolution represents the criterion used to decide if two normals correspond to the same plane.

Finally, a plane is formed in the image, i.e. a template from $\boldsymbol{\mathcal{I}}^*$, by means of the convex hull:

$$\mathcal{H}^* \equiv \left\{ \sum_i \mu_i \, \mathbf{p}_i^* \; : \; \mu_i \geq 0, \; \forall i, \; \text{and} \; \sum_i \mu_i = 1 \right\}. \tag{7.15}$$

The convex hull, also referred to as the convex envelope, is the smallest set of points containing all clustered points of a sufficiently voted accumulator.

Refinement through direct image registration

After defining a plane, a refinement can be conducted since its current estimate $\mathbf{n}_d^*$ from (7.10) is obtained from the sparse set of points contained in the accumulator (and thus, is not very accurate). To this end, one can apply a direct image registration using only the structure and illumination parameters as optimization variables. The motion parameters are used to obtain the transformed image $\boldsymbol{\mathcal{I}}'$ but are kept constant, i.e. they do not represent optimization variables.

More formally, this particular task can be accomplished for a particular j-th plane by adapting the generic framework (7.7) as

$$\min_{\widetilde{\mathbf{z}}^{c\prime\prime}=\{\widetilde{\mathbf{y}},\widetilde{\mathbf{z}}_h\}} \frac{1}{2}\sum_i \left[\boldsymbol{\mathcal{I}}'\big(\widehat{\mathbf{T}}\mathbf{T}_j^{-1}, \mathbf{n}_{d\,j}^*\big(\mathbf{z}^*(\widetilde{\mathbf{y}})\circ\widehat{\mathbf{z}}^*\big), \mathbf{h}(\widetilde{\mathbf{z}}_h)\circ\widehat{\mathbf{h}}, \mathbf{p}_{ij}^*\big) - \boldsymbol{\mathcal{I}}_j^*(\mathbf{p}_{ij}^*)\right]^2, \quad (7.16)$$

where $\mathbf{T}_j$ represents the relative transformation ${}^{\tau}\mathbf{T}_0$ from the origin to the frame where the j-th plane has been identified (indexed by 'τ'), $\widehat{\mathbf{T}} = {}^{c}\widehat{\mathbf{T}}_0$ encodes the current camera pose relatively to the origin, and using the relation between normal vector and inverse of the depths of 3 image points expressed in (5.13),

$$\mathbf{n}_d^* = \mathbf{M}\left[\,(z_1^*)^{-1}, (z_2^*)^{-1}, (z_3^*)^{-1}\,\right]^\top,$$

with $\mathbf{M} \in \mathbb{R}^{3\times 3}$ as defined in (5.14), so that the cheirality constraint can also be enforced within the minimization procedure.

This task is also useful as a plane verification step. Indeed, the resulting cost value can be used to discard regions that contain outliers, such as non-planar objects within the template. Notice that, whereas image features are needed to vote and form the templates, all pixels within the regions are to be exploited by the direct localization method.

The full efficient system

After having all plane normals $\{\mathbf{n}_{d\,j}^*\}$ refined, an efficient localization system from (7.7) can be formulated as

$$\min_{\widetilde{\mathbf{z}}^{c\prime}=\{\widetilde{\mathbf{v}},\widetilde{\mathbf{z}}_h\}} \frac{1}{2}\sum_j\sum_i \left[\boldsymbol{\mathcal{I}}'\big(\mathbf{T}(\widetilde{\mathbf{v}})\,\widehat{\mathbf{T}}\mathbf{T}_j^{-1}, \mathbf{n}_{d\,j}^*, \mathbf{h}(\widetilde{\mathbf{z}}_h)\circ\widehat{\mathbf{h}}, \mathbf{p}_{ij}^*\big) - \boldsymbol{\mathcal{I}}_j^*(\mathbf{p}_{ij}^*)\right]^2. \quad (7.17)$$

Once again, the optimal $\widehat{\mathbf{T}} = {}^{c}\widehat{\mathbf{T}}_0$ encodes the current camera pose relatively to the origin, and the same optimization procedure presented in Subsections 3.3 and 3.3.2 can be applied. In this plane-based system, computational efficiency in estimating the camera pose is improved. This corresponds to an efficient version of the E-3D visual servoing (see Fig. 7.1).

Nevertheless, even though suitable regions are explicitly identified, outliers may still appear in the image during the robot navigation (e.g. an independently moving object). Hence, one needs to detect and discard them. To this end, the same indexes described in Subsection 5.2.3 can also be used here.

7.2.3 Control aspects

Consider a camera-mounted holonomic robot or an omnidirectional mobile robot. Under the assumption that known objects can leave the field-of-view without destabilizing the system (i.e. assumption that, if known objects leave, new ones can be identified), the control error can be entirely constructed in the Cartesian space. In this case, this error is designed using the recovered current pose $\widehat{\mathbf{T}} = {}^{c}\widehat{\mathbf{T}}_0$ (see Section 7.2.1) and the user-defined reference ${}^{0}\mathbf{T}_r$.

Following the conventions adopted throughout this thesis, let us express the control error with respect to $\mathcal{F}_c$, instead of the reference frame $\mathcal{F}_r$. To this end, define first

$$\begin{aligned} {}^{c}\widehat{\mathbf{T}}_r &= {}^{c}\widehat{\mathbf{T}}_0 \; {}^{0}\mathbf{T}_r && (7.18) \\ &= \begin{bmatrix} {}^{c}\widehat{\mathbf{R}}_r & {}^{c}\widehat{\mathbf{t}}_r \\ \mathbf{0} & 1 \end{bmatrix} \quad \in \mathbb{SE}(3). && (7.19) \end{aligned}$$

Definition 7.2. Given that

$$ {}^{c}\mathbf{R}_r = \exp([\theta\, {}^{c}\mathbf{u}_r]_\times), \quad \in \mathbb{SO}(3), \qquad (7.20)$$

the *control error* vector in the calibrated setting can be defined as

$$\begin{aligned} \boldsymbol{\varepsilon}^{\mathrm{c}} = \left[\boldsymbol{\varepsilon}_{\nu}^{\mathrm{c}\top}, \boldsymbol{\varepsilon}_{\omega}^{\mathrm{c}\top} \right]^\top &= \left[{}^{c}\widehat{\mathbf{t}}_r^{\top}, \widehat{\theta}\, {}^{c}\widehat{\mathbf{u}}_r^{\top} \right]^\top && (7.21) \\ &= \left[\mathbf{t}^\top, \theta\mathbf{u}^\top \right]^\top \quad \in \mathbb{R}^6, && (7.22) \end{aligned}$$

which, by dropping the indices from (7.21), respectively denotes the error in translation and in rotation of the reference frame with respect to the current one.

Remark 7.1. We emphasize that this particular control error corresponds to a positioning task whose desired pose is specified relative to the initial robot pose ${}^{0}\mathbf{T}_r$. Another possible task could be, for instance, to drive the camera to a given desired pose relative to a particular known object.

Definition 7.3. Let $\mathbf{v} = \left[\boldsymbol{\nu}^\top, \boldsymbol{\omega}^\top \right]^\top \in \mathbb{R}^6$ respectively represent the translational and rotational velocities of the camera. The *control law* can be defined simply as

$$\mathbf{v} = \boldsymbol{\Lambda}\, \boldsymbol{\varepsilon}^{\mathrm{c}}, \qquad (7.23)$$

with the control gain

$$\begin{aligned} \boldsymbol{\Lambda} &= \operatorname{diag}(\lambda_\nu \mathbf{I}_3, \lambda_\omega \mathbf{I}_3) && (7.24) \\ &= \begin{bmatrix} \lambda_\nu\, \mathbf{I}_3 & \mathbf{0} \\ \mathbf{0} & \lambda_\omega\, \mathbf{I}_3 \end{bmatrix}, \quad \lambda_\nu, \lambda_\omega > 0, && (7.25) \end{aligned}$$

and the control error given in Definition 7.2.

Theorem 7.1 (Global stability). *Consider the efficient E-3D visual servoing technique. The control law* (7.23) *ensures a global*[2] *asymptotic stability of the system under the assumption that, if known objects leave the field-of-view, then new planes can be accurately identified as the camera displaces toward the desired pose.*

Proof. The proof is presented in Appendix B.7. ∎

Corollary 7.1 (Straight-line path). *The control law* (7.23) *induces a full decoupling of translational and rotational motions of the camera. Thus, if the estimated poses are perfect, then the camera performs a straight-line path in the Cartesian space.*

Remark 7.2. It can be noted that the control law (7.23) has a positive sign. This is due to how the control error (6.6) is defined, which is relative to the current frame $\mathcal{F}_c$. That is, $\boldsymbol{\varepsilon}^{\mathrm{c}} = {}^{c}\boldsymbol{\varepsilon}_r^{\mathrm{c}}$.

7.3 Results

This section reports some representative results concerning the proposed Planar Region Detector and the efficient E-3D visual servoing.

Planar Region Detector. To assess the performance of the Planar Region Detector (PRD), we have tested it against a large data set of both simulated and real images. In all cases, the resolution for the normal vector is set to 5° and to 0.05 m for the distance to the plane. With respect to their boundaries, as already stated, they do not need to be defined a priori. Also, the disk parameters were set to $r = 5$ and $R = 50$ pixels. As for the corresponding points, they are provided here by a standard technique. For this, the sub-pixel Harris detector together with a correlation-based matching algorithm are used. Additionally, in accordance with probabilistic Hough-like transforms, where as low as 2% of the number of points is used (~ 300 here), the threshold on the minimum number of votes was then set to 15 from the binomial $\binom{0.02*300}{3}$. Those parameters remained constant for all experiments.

To have a ground truth for a large range of variations for each input variable, the same synthetic 3D scene described in Chapter 5 is used here. This scene is composed of four planes disposed in pyramidal form, but cut by another plane on its top. Onto each one of the five planes, a different texture is applied (see Fig. 7.4). The reference camera frame $\mathcal{F}^*$ is positioned at the center of the pyramid pointing downward, and whose perpendicular distance to the farthest plane (the top plane) is of $d^* = 1$m. This distance does not represent a restricting fact since it is the amount of scaled translation $\|\mathbf{t}\|/d^*$ between the two frames (along with the focal length) that plays an important geometric role for scene reconstruction from a pair of images. This represents

[2] in the domain $\mathbb{R}^3 \times \{\mathbf{R} = \exp([\theta\mathbf{u}]_\times) \in \mathbb{SO}(3) \colon \theta \in \,]-\pi, \pi]\}$.

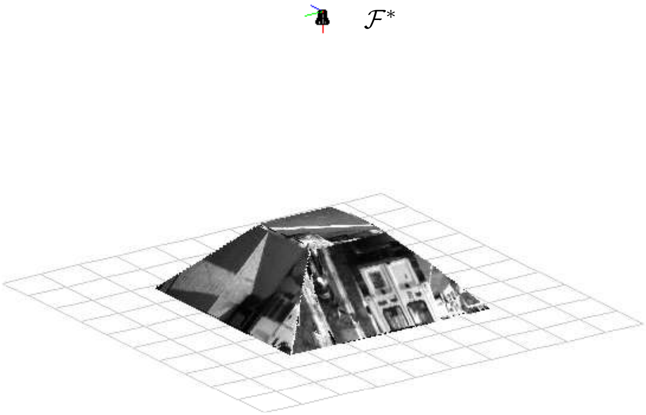


Figure 7.4. The textured synthetic scene designed for the systematic tests.

the baseline with respect to depth in case of stereoscopic images. We have then conducted more than $10,000$ simulations to investigate the performance of the PRD algorithm. For every simulation, a normally distributed, independent noise η_i with mean 0 and standard deviation 1/6 is added to every input camera parameter: $\widehat{a}_i = a_i(1 + \frac{1}{6}\eta_i)$. This means that such an input has an error of up to 50% in 99.7% of the cases. From $\mathcal{F}^*$, random directions of translation as well as random rotations were used to displace the camera by a varying amount of $\|\mathbf{t}\|/d^* \in [0.01, 0.5]$. We remark that the image may contain fewer planes for large displacements (large baselines), since we do not enforce that all planes must remain in the image. The median number of corresponding points, along with the interquartile range, and of the percentage of outliers in the data are shown in Fig. 7.5. A corresponding point is said to be an outlier here if the known warping (ground truth) of the extracted point in the first image and the extracted point in the second view gives an error over 5σ pixels. It was considered that the point detector has a standard deviation of 1 pixel.

Then, from such a large, noisy input data set, two measures have been computed for assessing the performance of the PRD: the median number of detected planar regions as well as of the rate of false positives. The results are shown in Fig. 7.6. Firstly, as theoretically demonstrated in Appendix B.6, if $\|\mathbf{t}\|/d^*$ is too small then *any* scene may be viewed as a single plane (the plane at infinity). That explains the high rate of false positives for $\|\mathbf{t}\|/d^* = 0.01$. However, for all the other cases, a median of zero false positive planes is obtained. Moreover, it can be noted that this happens even if a large number of mismatched points (outliers) is present in the process (compare Figs. 7.5 and 7.6 for large displacements), though reducing the number of detected planes. Such a result confirms the robustness of the PRD algorithm to large errors in the camera parameters and to the presence of outliers.

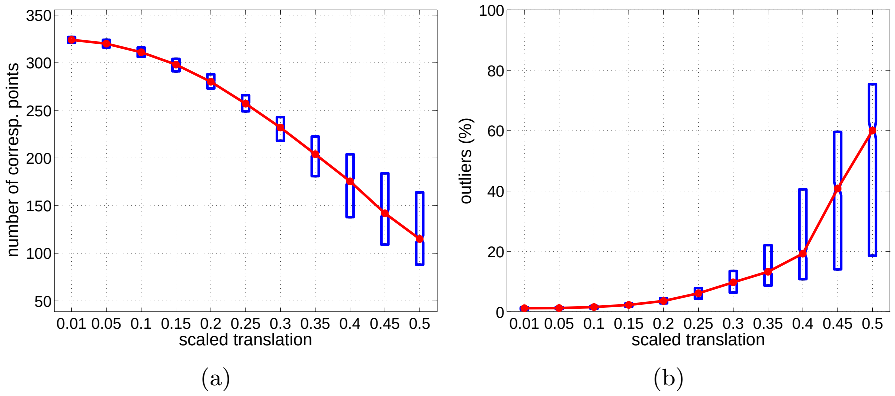


Figure 7.5. (a) Median number of corresponding points along with the interquartile range, as well as (b) the median percentage of outliers present in the simulated data, as the amount of the scaled translation is varied.

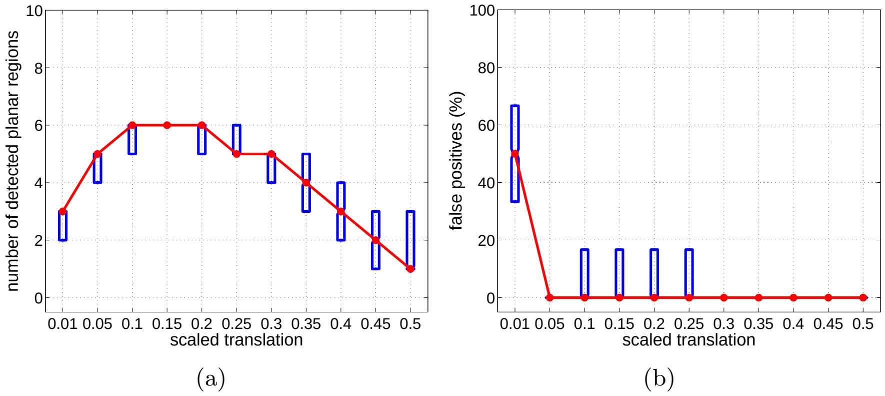
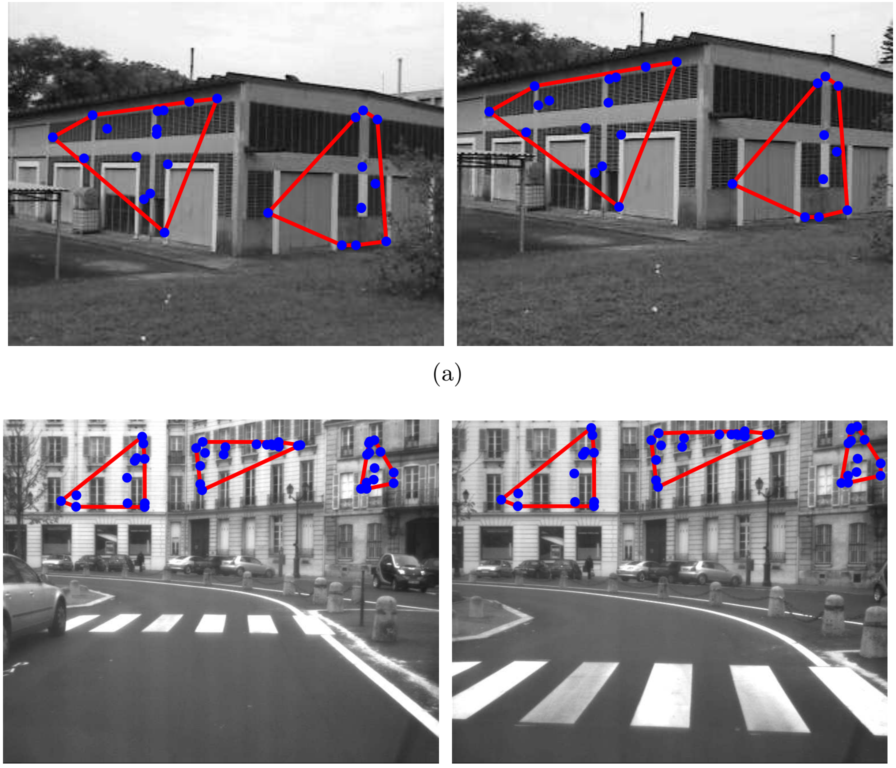


Figure 7.6. (a) Median number of the detected planar regions, and (b) median number of the rate of false positives, obtained from such a large, noisy input data set (see Fig. 7.5). In the simulations, the planes are not enforced to always remain in the image.

With respect to experimental results, some examples are shown in Fig. 7.7. Again to illustrate the robustness characteristics of this detector, aside from the unavoidable mismatched features, we have used erroneous both intrinsic and extrinsic camera parameters. For all pairs of images tested, we set $\widehat{\alpha}_u = \widehat{\alpha}_v = 500$ pixels, principal point as the middle of the image, zero skew, as well as $\widehat{\mathbf{R}} = \mathbf{I}_3$ and $\widehat{\mathbf{t}} = [-0.1, 0, -1]^\top$ m for the rotation and translation motions, respectively. Despite all these sources of noise, actual Euclidean planes are detected, confirming the robustness properties found in the simulations. Since the approach aims to cluster planar regions *in the image*, large errors on the camera parameters are tolerated. The effect of erroneous camera parameters appears on the Cartesian values. Obviously, the used pairs of images verify the geometric conditions (see Lemma 7.1) for segmenting real planes. In order to satisfy real-time requirements, only a part of each plane is clustered. Nevertheless, a region growing process could be used to partition a larger extent of them. For example, by iteratively verifying if other input features (not shown in the figure for the sake of clarity) projectively fit a given plane model.


(a)

(b)

Figure 7.7. Results obtained by the Planar Region Detector (a) on a pair of outdoor images and (b) on a pair of urban images. The other input features are not shown for the sake of clarity. Due to real-time requirements, only a part of each plane is segmented. A larger extent can be obtained by region growing.

Efficient E-3D visual servoing. With respect to the navigation task, a desired Cartesian trajectory with loop closing is specified and afterward subdivided into 10 elementary positioning tasks. The trajectory has a total displacement of approximately 3.3 m. An elementary task is said to be completed here when the translational error drops below a certain precision (it was set when $\|\boldsymbol{\varepsilon}_\nu\| < 1$ mm). It is evident that the total amount of time (and hence the total number of images) needed to perform the task also depends on the chosen control gain, which is set here to $\lambda_\nu = \lambda_\omega = 0.5$. The images obtained at the convergence for some of these tasks are shown in Fig. 7.8, where the detected and exploited planes are superimposed as well. Note that even though a known plane (shown in the third image of Fig. 7.8) leaves the field-of-view, the entire navigation task is successfully performed since new planes are identified. In addition, when such a known plane reenters the image it is automatically re-detected.

The true errors obtained by the pose recovery process along the entire task are shown in Fig. 7.9, since the ground truth is available. One can observe that when the image loses resolution (e.g. the camera moves away from the object), the precision of the reconstruction also decreases. Nevertheless, one important benchmark is obtained from performing a closed-loop trajectory: errors smaller than 0.1 mm and than 0.01° are obtained after the camera comes back to the same pose at the beginning (compare the first and last images of Fig. 7.8). This demonstrates the degree of accuracy achieved by the framework in simulation conditions.

Another important result from the approach concerns the reconstruction of the scene in the 3D space (up to a scale factor), which is shown in Fig. 7.10 for different views of the scene. This demonstrates that the proposed efficient E-3D visual servoing approach can be also used as a "Plane-based Structure from Controlled Motion" technique, improving the stability, the accuracy and the rate of convergence of Structure From Motion methods.

7.4 Summary

This chapter proposes a visual servoing scheme where the desired pose is directly provided in the Euclidean space. Further, it is considered here that the desired image (corresponding to the given desired pose) and the metric model of the scene are both not available a priori. To accomplish the task, an accurate and robust localization method is formulated.

A special attention has been given to the particular case of modeling the scene as a collection of planar regions. The main interests concern its versatility and computational efficiency. We have thus specialized the localization method and proposed a new planar region detector. Hence, new planes can be identified (and thus exploited), since known ones may eventually leave of the field-of-view during an extended navigation task. The proposed robust detector clusters multiple planar regions in a global optimal sense, featuring fast speed, small storage, infinite range, and high resolution. Navigation tasks are performed and only very small Cartesian errors are obtained using this framework.

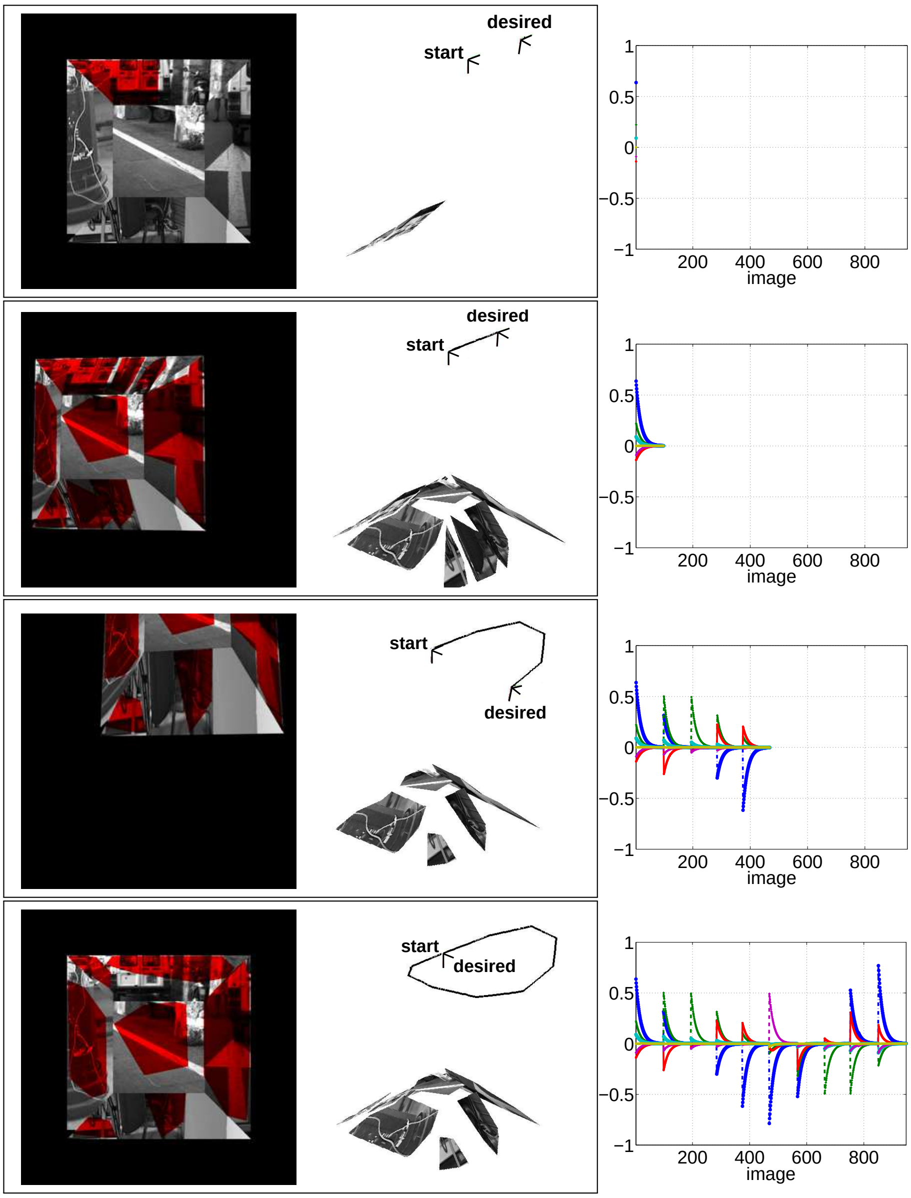


Figure 7.8. A closed-loop navigation task comprised of 10 elementary positioning ones. A plane is initialized in the first image. For each elementary task shown, it is drawn respectively from left to right: the obtained image at the convergence superimposed by the exploited planes, the corresponding reconstructed pose and scene, and the control input (in m/s and radians/s). Observe that a plane gets out of the field-of-view (image in the third row), but when it reenters it is again identified (image in the last row).

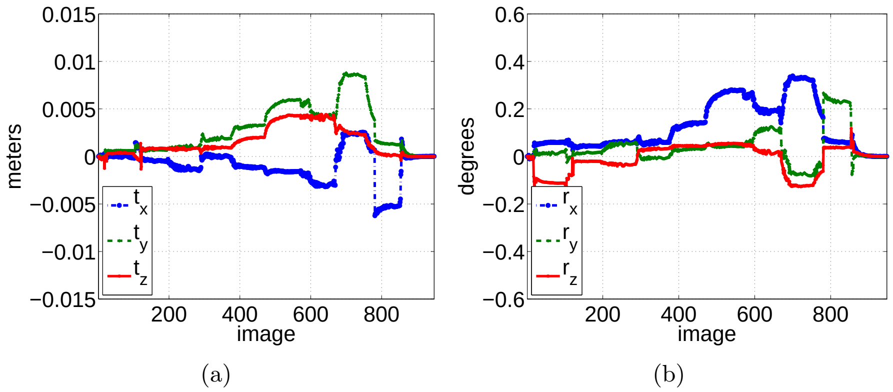


Figure 7.9. (a) Errors in the position recovery and (b) errors in the attitude recovery, with respect to ground truth, along the entire navigation task (≈3.3 m). The Euclidean norm of these errors at the end of this closed-loop trajectory (camera comes back to the same pose at the beginning) is smaller than 0.1 mm and than 0.01°, respectively for the position and orientation.

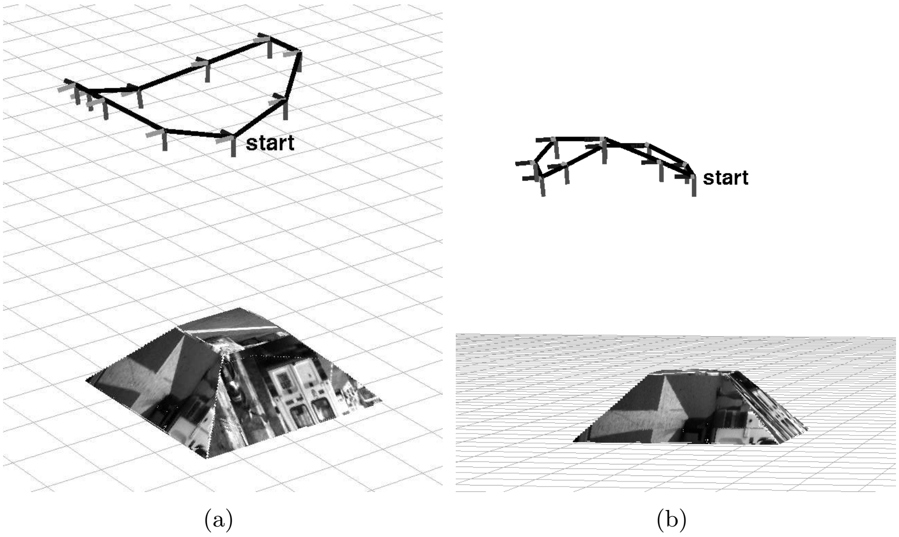


Figure 7.10. Different viewpoints of the reconstructed 3D scene (after performing a region growing of the exploited planes), of the trajectory performed by the camera (line linking the frames), and the desired poses to be reached (represented by frames).

Conclusions and future research

Image registration is one of the key tools developed in this thesis to perform estimation and control from visual information. It could also be called image regulation — a term borrowed from control theory — because it comprises a feedback loop both whilst estimating parameters and whilst controlling a system from visual data. An important difference between the two applications concerns how the corresponding systems are transformed from their initial state to the reference one. Whilst estimating, the computed signals (i.e. the parameters) are iteratively fed into a generative model so as to synthetically transform images. Whilst controlling, the input signals (e.g. robot velocities) are fed into a dynamic system in order to physically transform it.

To date, the overwhelming majority of vision-based techniques for both estimation and control consider a feature-based scheme. A large portion of this thesis has focused on how to adequately exploit pixel intensities directly, without having to first extract and match some image features. In other terms, we have focused on how to appropriately perform direct registration for both estimation and control from visual data. The results shown here suggest that direct methods are well-suited to these applications. This is especially true in the robotics field since, in this case, the main limitation of these methods (i.e. the domain of convergence) is not so restrictive. Indeed, within robotic applications we can suppose that the image acquisition rate is sufficiently high, such that only relatively small inter-frame displacements of the objects are perceived. Furthermore, their strengths are relevant to that field and have been thoroughly discussed and demonstrated here. Principally, the degree of accuracy that direct methods can attain in practice. The advantages are mainly owing both to the possibility of exploiting all possible image information, even from where no distinctive features exist, and to the simultaneous enforcement of various structural constraints.

Considering pinhole cameras, we have also shown in this thesis that, depending on the task at hand, the registration can be formulated either in the calibrated setting or in the uncalibrated one directly. That is, one does not necessarily have to rely on the Euclidean stratum (and hence, on accurate camera parameters) to perform vision-based estimation and/or control tasks. Moreover, both cases share a common framework. The proposed framework is composed of efficient and generic photo-geometric transformation models and optimization procedures. The differences between the two settings mainly regard to the needed parametrization so as to deal with their particular specificities. For example, in order to enforce the cheirality constraint in the calibrated case.

Specifically to estimation, we have proposed an efficient and robust solution to visual tracking both in the uncalibrated case and in the calibrated one. As a matter of fact, the calibrated visual tracking technique directly provides the camera pose and the scene structure. Therefore, it represents a new solution to the visual SLAM problem. As for control, we have proposed a flexible and reliable teach-by-showing strategy to drive all six degrees-of-freedom to the reference state. The technique does not require either precise parameters of the vision system or metric knowledge of the observed scene. In the calibrated case, we integrate a vision-based control into the visual SLAM approach. This controlled visual SLAM scheme in fact allows for autonomous navigation of mobile robots over previously unexplored scenes.

Every effort has been made to devise a framework that suits the requirements of robotic applications. Indeed, we have searched for accurate, simple yet generic solutions to those tasks. Nevertheless, the design of such systems usually involves making compromises. As for simplicity, a hierarchical approach in the number of parameters is applied. Indeed, this strategy is also defined in terms of sufficiency so that real-time performance can be achieved. As for achieving high levels of accuracy, direct methods are developed. Generality is also important in order to construct systems as flexible as possible. That is, systems that can work independently of its configuration, using minimal (or no) prior knowledge. Indeed, we have proposed visual tracking and visual servoing techniques that can be highly accurate despite unknown objects and unknown imaging conditions. In all case, since absolutely generic systems are still out of reach, we have focused on widely adopted classes of them, such as the pinhole camera model.

Hence, we plan to work on extending the models and methods proposed in this thesis to a broader class of vision systems. For example, on extending them to (multiple) central catadioptric cameras, both in terms of estimation and control. A more ambitious research direction consists in studying how to relax the assumption of central cameras. Some of non-central cameras present the benefits of their counterparts (e.g. large field-of-view), whilst having constant resolution (one of their major difficulties).

Specifically to estimation, the limitation of direct methods concerning the domain of convergence should be addressed as well. This is important in order to avoid system failures in the case of rapidly moving objects and/or the camera itself. Here, convergence difficulties are partially overcome owing to an (eventual) usage of feature-based and/or other predictors, such as the Kalman filtering. Nevertheless, they also have their limitations in providing sufficiently good predictions. As for feature-based methods, see the discussion presented in Subsection 2.1.1. In particular, feature detection and matching are not fully invariant to all possible changes in all photometric and geometric parameters. As for the filtering, the assumption on the type of noise (e.g. Gaussian) and/or on the model of motion (e.g. constant velocity) may not be realistic in many scenarios.

Another application that can be benefited from direct methods concerns loop closing within visual SLAM approaches. That is, the detection of previously visited locations in a map. This constitutes an important component to be developed in order to produce an accurate and reliable long-term navigation system. Again, the majority of working systems rely on image features to perform this task. We believe that these features, by themselves, are not particularly discriminative to allow loop closing under disparate imaging conditions.

Still for estimation purposes, the direct registration framework can also be of particular usefulness for fusioning data from multiple sensory modalities, including range sensors. This is likely to improve accuracy and robustness of vision-only estimation systems, notably in the monocular case.

Specifically to control, several problems and analysis issues still remain open. This is especially the case of the proposed approach where the reference pose is defined by means of a reference image. Firstly, we plan to carry out real-world experiments to validate the proposed technique. In addition, an important analysis to be conducted for this strategy concerns its robustness to errors in the camera intrinsics parameters. Although we have observed a large degree of robustness in extensive simulations, no theoretical analysis has been performed on this aspect.

Last but not least, this thesis has focused on controlling only non-critical non-linear systems. That is, the case of classical manipulator robots or omnidirectional mobile robots. One of research directions we plan to pursue within vision-based control concerns the generalization of the proposed schemes to deal with critical non-linear systems, such as ground and aerial robots.

Part IV

Appendices

Appendix A

Useful relations

A.1 Normal vector and homography

A closed-form solution is given here to determine the Euclidean parameters $\boldsymbol{\pi}^* = \left[\mathbf{n}^{*\top}, -d^*\right]^\top \in \mathbb{R}^4$ of a plane, with respect to the reference frame, provided the camera motion between two views $\{\mathbf{R}, \mathbf{t}\}$, and the associated projective homography $\mathbf{G} \in \mathbb{R}^{3\times3}$. It can be noted that $\mathbf{G}$ is not necessarily parametrized here as an element of $\mathbb{SL}(3)$ so that a general relation is deduced. Of course, the camera's internal parameters $\mathbf{K}$ are always necessary to perform the upgrade from projective to Euclidean stratum.

To this end, multiplying Eq. (1.31), i.e.

$$\mathbf{G} \propto \mathbf{K}\,\mathbf{H}\,\mathbf{K}^{-1}$$

where

$$\mathbf{H} = \mathbf{R} + (d^*)^{-1}\,\mathbf{t}\,\mathbf{n}^{*\top}, \tag{A.1}$$

on the left by $\mathbf{K}^{-1}$ and on the right by $\mathbf{K}$, one obtains

$$\mathbf{H} = \alpha\,\mathbf{K}^{-1}\,\mathbf{G}\,\mathbf{K}, \tag{A.2}$$

where $\alpha \in \mathbb{R}$ is a normalizing factor. Then, using (A.1) yields

$$\mathbf{t}\,\mathbf{n}_d^{*\top} = \alpha\,\mathbf{K}^{-1}\,\mathbf{G}\,\mathbf{K} - \mathbf{R}, \tag{A.3}$$

with

$$\mathbf{n}_d^* = (d^*)^{-1}\,\mathbf{n}^* = \|\mathbf{n}_d^*\|\,\mathbf{n}^*. \tag{A.4}$$

Definition A.1. Pre-multiplying both members of Eq. (A.3) by $\mathbf{t}^\top$ and using the Euclidean norm

$$\|\mathbf{t}\| = \sqrt{\mathbf{t}^\top\mathbf{t}}, \tag{A.5}$$

a *closed-form solution* can be defined for determining the normal vector relative to the reference frame:

$$\mathbf{n}_d^* = \frac{\left(\alpha\,\mathbf{K}^{-1}\,\mathbf{G}\,\mathbf{K} - \mathbf{R}\right)^\top\mathbf{t}}{\|\mathbf{t}\|^2}. \tag{A.6}$$

In order to determine $\boldsymbol{\pi}^*$ using (A.6) and then (A.4), the normalizing $\alpha \in \mathbb{R}$ can be obtained as follows. Let $\operatorname{svd}(\mathbf{H}) = [\,\sigma_1,\, \sigma_2,\, \sigma_3\,]^\top$ be the singular values of $\mathbf{H}$ in decreasing order, $\sigma_1 \geq \sigma_2 \geq \sigma_3 > 0$. Such a homography can be normalized by the median singular value (Faugeras and Lustman, 1988). In this case, it is possible to use the facts that $x = \operatorname{sign}(x)\,|x|$, $\forall x \in \mathbb{R}$, that $\det(\mathbf{H}) = \prod_{k=1}^{3} \lambda_k(\mathbf{H})$, and also that the strictly positive σ_k are the square-roots of the eigenvalues $\lambda(\mathbf{H}^\top\mathbf{H})$, in order to define

$$\alpha = \frac{\operatorname{sign}\big(\det(\mathbf{H})\big)}{\sigma_2(\mathbf{H})}, \tag{A.7}$$

where $\operatorname{sign}(\cdot)$ denotes the signum function.

Lemma A.1 (Normal Vector Characterization). *The necessary and sufficient geometric conditions for the normal vector determination expressed in Eq. (A.6) are such that:*

1. $\|\mathbf{t}\| > 0$;
2. $|\det(\mathbf{G})\,| > 0$.

Proof. The proof is presented in Appendix B.1. ∎

A.2 Control error and camera pose

Let us state an important result which is largely used throughout the Part III (related to vision-based control) of this thesis.

Lemma A.2 (Control error and camera pose). *The proposed control error* $\boldsymbol{\varepsilon}^{\mathrm{c}} = \big[\boldsymbol{\varepsilon}_\nu^{\mathrm{c}\top}, \boldsymbol{\varepsilon}_\omega^{\mathrm{c}\top}\big]^\top$ *in* (6.6) *is expressed as a function of the camera pose* $\{\mathbf{R}, \mathbf{t}\}$ *(or equivalently,* $\{\theta\mathbf{u}, \mathbf{t}\}$*), through*

$$\boldsymbol{\varepsilon}_\nu^{\mathrm{c}} = \frac{\beta_\nu}{z^*}\big((\mathbf{R} - \mathbf{I}_3)\underline{\mathbf{m}}^* + \mathbf{t}\big), \tag{A.8}$$

for some normalization factor $\beta_\nu > 0$*, and*

$$\boldsymbol{\varepsilon}_\omega^{\mathrm{c}} = \vartheta \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|} \tag{A.9}$$

with

$$\boldsymbol{\mu} = \beta_\omega \left(\sin(\theta)\mathbf{u} + \frac{1}{2}[\mathbf{q}^{*\prime}]_\times \mathbf{t}\right), \tag{A.10}$$

for some normalization factor $\beta_\omega > 0$*. The "projective angle of rotation"* $\vartheta \in \,]-\pi, \pi]$ *is defined in Eq.* (6.5)*. The factors* β_ν *and* β_ω *depend on the reconstruction algorithm, for example, on how homographies are parametrized.*

Proof. The proof is presented in Appendix B.2. ∎

Appendix B

Theoretical demonstrations

B.1 Proof of the Lemma A.1

Proof (Normal Vector Characterization). The proof comes directly from analyzing (A.6), together with the knowledge that $\mathbf{K} > 0$ and $\mathbf{R} \in \mathbb{SO}(3)$.

The first condition $\|\mathbf{t}\| > 0$ is necessary so that (A.6) is well-defined. In fact, it states that a sufficient amount of translation relative to the distance of the plane has to be carried out, i.e. $\|\mathbf{t}\|/d^* > 0$. Otherwise, its Euclidean structure cannot be recovered reliably. Indeed, manipulating Eqs. (A.6) and (A.4) gives

$$\mathbf{n}^* = \frac{\left(\alpha\, \mathbf{K}^{-1}\, \mathbf{G}\, \mathbf{K} - \mathbf{R}\right)^\top \mathbf{t}}{\|\mathbf{t}\|^2/d^*}. \tag{B.1}$$

The last condition comes from the fact that $\alpha \neq 0$ also to avoid the trivial solution. From Eq. (A.7), given that $\sigma_k > 0$, $\forall k$, one must then have $|\det(\mathbf{H})| > 0$. Hence, using (A.2)

$$|\det(\mathbf{H})| \;>\; 0 \tag{B.2}$$

$$\left|\alpha^3\right| \left|\det(\mathbf{K}^{-1})\, \det(\mathbf{G})\, \det(\mathbf{K})\right| \;>\; 0 \tag{B.3}$$

$$|\det(\mathbf{G})| \;>\; 0. \tag{B.4}$$

The $\det(\mathbf{G})$ can in this case be used as a measure of degeneracy of the plane (in order to discard it, for instance). The plane is in a degenerate configuration when is projected in the image as a line. In this case, $\det(\mathbf{G}) = 0$. ■

B.2 Proof of the Lemma 6.1

Proof (Control error and camera pose). We provide a constructive proof.

Let us start with $\varepsilon_\nu^{\mathrm{u}}$. Consider the generic relation between corresponding points in uncalibrated images of a rigid object expressed in (1.23), i.e.

$$\mathbf{p} \,\propto\, \mathbf{G}\, \mathbf{p}^* + \rho^*\, \mathbf{e}. \tag{B.5}$$

Multiplying the above equation (B.5) on the left by $\mathbf{K}^{-1}$, and injecting Eqs. (1.16), (6.1), (6.2), and (6.3), respectively

$$\mathbf{p} = \mathbf{K}\,\mathbf{m}' \tag{B.6}$$

$$\mathbf{m}^{*\prime} = \mathbf{K}^{-1}\mathbf{p}^{*} \tag{B.7}$$

$$\mathbf{e}' = \mathbf{K}^{-1}\mathbf{e} \tag{B.8}$$

$$\mathbf{H} = \mathbf{K}^{-1}\mathbf{G}\,\mathbf{K}, \tag{B.9}$$

one obtains

$$\mathbf{m}' \;\propto\; \mathbf{H}\,\mathbf{m}^{*\prime} + \rho^{*}\mathbf{e}'. \tag{B.10}$$

This result can be rewritten as

$$\alpha\,\mathbf{m}' \;=\; \mathbf{H}\,\mathbf{m}^{*\prime} + \rho^{*}\mathbf{e}', \tag{B.11}$$

with the scale factor given by

$$\alpha = \beta_{\nu}\frac{z}{z^{*}} > 0, \tag{B.12}$$

where $\beta_{\nu} > 0$ is only a normalization factor which depends on the reconstruction algorithm. Expanding the proposed control error $\boldsymbol{\varepsilon}_{\nu}^{\mathrm{u}}$ in (6.6) and using (B.11), we have

$$\boldsymbol{\varepsilon}_{\nu}^{\mathrm{u}} = (\mathbf{H} - \mathbf{I})\,\mathbf{m}^{*\prime} + \rho^{*}\mathbf{e}' \tag{B.13}$$

$$= \mathbf{H}\,\mathbf{m}^{*\prime} + \rho^{*}\mathbf{e}' - \mathbf{m}^{*\prime} \tag{B.14}$$

$$= \alpha\,\mathbf{m}' - \mathbf{m}^{*\prime}. \tag{B.15}$$

Then, using the scale factor (B.12) and Thales' theorem, we can rewrite (B.15) as

$$\boldsymbol{\varepsilon}_{\nu}^{\mathrm{u}} = \frac{\beta_{\nu}}{z^{*}}(\underline{\mathbf{m}} - \underline{\mathbf{m}}^{*}). \tag{B.16}$$

Finally, by injecting the equation of rigid-body motion (1.1) in the equation above, we obtain the desired relation between $\boldsymbol{\varepsilon}_{\nu}^{\mathrm{u}}$ and the camera pose:

$$\boldsymbol{\varepsilon}_{\nu}^{\mathrm{u}} = \frac{\beta_{\nu}}{z^{*}}\big((\mathbf{R} - \mathbf{I}_{3})\underline{\mathbf{m}}^{*} + \mathbf{t}\big). \tag{B.17}$$

With respect to $\boldsymbol{\varepsilon}_{\omega}^{\mathrm{u}}$, consider the possible characterization of $\mathbf{G}$ expressed by (1.24), i.e.

$$\mathbf{G} \;\propto\; \mathbf{G}_{\infty} + \mathbf{e}\,\mathbf{q}^{*\top}, \tag{B.18}$$

within the generic relation between corresponding image points in uncalibrated images in (B.5). Multiplying (B.18) on the left by $\mathbf{K}^{-1}$ and on the right by $\mathbf{K}$, and then using (B.9), we obtain

$$\mathbf{H} \;\propto\; \mathbf{K}^{-1}\mathbf{G}_{\infty}\mathbf{K} + \mathbf{K}^{-1}\mathbf{e}\,\mathbf{q}^{*\top}\mathbf{K}. \tag{B.19}$$

Next, by injecting the relations (1.28) and (1.32), respectively

$$\mathbf{e} \propto \mathbf{K}\,\mathbf{t} \tag{B.20}$$

$$\mathbf{G}_{\infty} \propto \mathbf{K}\,\mathbf{R}\,\mathbf{K}^{-1}, \tag{B.21}$$

in (B.19) gives

$$\mathbf{H} = \beta_\omega \big(\mathbf{R} + \mathbf{t}\, \mathbf{q}^{*\prime\top} \big) \tag{B.22}$$

with

$$\mathbf{q}^{*\prime} = \mathbf{K}^\top \mathbf{q}^* \tag{B.23}$$

and the normalization factor $\beta_\omega > 0$. The "projective axis of rotation" (6.4) can then be developed:

$$[\boldsymbol{\mu}]_\times = \frac{1}{2}\big(\mathbf{H} - \mathbf{H}^\top \big) \tag{B.24}$$

$$= \frac{\beta_\omega}{2}\big(\mathbf{R} + \mathbf{t}\, \mathbf{q}^{*\prime\top} - \mathbf{R}^\top - \mathbf{q}^{*\prime}\mathbf{t}^\top \big). \tag{B.25}$$

Using Rodrigues' formula

$$\mathbf{R} = \mathbf{I}_3 + \sin(\theta)[\mathbf{u}]_\times + \big(1 - \cos(\theta)\big)[\mathbf{u}]_\times^2, \tag{B.26}$$

whose transpose is given by

$$\mathbf{R}^\top = \mathbf{I}_3 - \sin(\theta)[\mathbf{u}]_\times + \big(1 - \cos(\theta)\big)[\mathbf{u}]_\times^2, \tag{B.27}$$

we have

$$\mathbf{R} - \mathbf{R}^\top = 2\sin(\theta)[\mathbf{u}]_\times. \tag{B.28}$$

By using (B.28) together with the property

$$[\mathbf{a}]_\times\, [\mathbf{b}]_\times - [\mathbf{b}]_\times\, [\mathbf{a}]_\times = \mathbf{b}\, \mathbf{a}^\top - \, \mathbf{a}\, \mathbf{b}^\top \, = \big[\, [\mathbf{a}]_\times \mathbf{b} \,\big]_\times \tag{B.29}$$

from the definition

$$[\mathbf{a}]_\times\, [\mathbf{b}]_\times = \mathbf{b}\, \mathbf{a}^\top - \, (\mathbf{a}^\top \mathbf{b})\, \mathbf{I}_3, \tag{B.30}$$

an important relation from Eq. (B.25) is obtained:

$$\boldsymbol{\mu} = \beta_\omega \left(\sin(\theta)\mathbf{u} + \frac{1}{2}[\mathbf{q}^{*\prime}]_\times \mathbf{t} \right). \tag{B.31}$$

The "projective angle of rotation" follows directly by injecting the norm of (B.31) in (6.5). Hence, the desired relation between $\boldsymbol{\varepsilon}_\omega^{\mathrm{u}}$ and the camera pose is also achieved. ■

B.3 Proof of the Theorem 6.1

This demonstration uses the results from Lemma 6.1, whose relations are fully presented in Appendix A.2.

Proof (Local isomorphism). The proof consists in demonstrating that $\boldsymbol{\varepsilon}^{\mathrm{u}} = \big[\boldsymbol{\varepsilon}_\nu^{\mathrm{u}\top}, \boldsymbol{\varepsilon}_\omega^{\mathrm{u}\top}\big]^\top = \mathbf{0}$ if and only if $\theta = 0$ and $\mathbf{t} = \mathbf{0}$. We remark that, even if the domain of the angle of rotation also includes $\theta = \pi$, it will be demonstrated here only the isomorphism around the equilibrium $\boldsymbol{\varepsilon}^{\mathrm{u}} = \mathbf{0}$.

First of all, it is evident that if $\theta = 0$ and $\mathbf{t} = \mathbf{0}$ then $\boldsymbol{\varepsilon}^{\mathrm{u}} = \mathbf{0}$ ($\Longleftarrow$). However, we need to prove the implication in the other direction ($\Longrightarrow$): if $\boldsymbol{\varepsilon}^{\mathrm{u}} = \mathbf{0}$ then $\theta = 0$ and $\mathbf{t} = \mathbf{0}$. That is, we have to show that the homogeneous non-linear system of equations $\boldsymbol{\varepsilon}^{\mathrm{u}} = \mathbf{0}$ has a unique solution which is $\theta = 0$ and $\mathbf{t} = \mathbf{0}$, $\forall \mathbf{q}^*$ and $\forall \underline{\mathbf{m}}^*$ such that $z^* > 0$.

We start by constructing such a system of equations, which is given as

$$\begin{cases} (\mathbf{R} - \mathbf{I}_3)\underline{\mathbf{m}}^* + \mathbf{t} = \mathbf{0} \\ \sin(\theta)\mathbf{u} + \dfrac{1}{2}[\mathbf{q}^{*\prime}]_\times \mathbf{t} = \mathbf{0} \\ \dfrac{1}{2}\big(\mathrm{tr}(\mathbf{H}) - 1\big) \geq 0. \end{cases} \tag{B.32}$$

The first equation of (B.32) comes directly from $\boldsymbol{\varepsilon}^{\mathrm{u}}_{\nu}$ (A.8) since $\beta_\nu > 0$ and $z^* > 0$. Thus, one obtains directly

$$\mathbf{t} = (\mathbf{I}_3 - \mathbf{R})\underline{\mathbf{m}}^* \tag{B.33}$$

$$= \big(\mathbf{I}_3 - \exp([\mathbf{u}\theta]_\times)\big)\underline{\mathbf{m}}^*. \tag{B.34}$$

Both the second equation and the inequality in (B.32) were constructed by injecting (A.10) in (6.5) and (6.6), together with the following facts. The statement $\boldsymbol{\varepsilon}^{\mathrm{u}}_{\omega} = \vartheta\boldsymbol{\mu}/\|\boldsymbol{\mu}\| = \mathbf{0}$ implies $\vartheta = 0$ since $\boldsymbol{\mu}/\|\boldsymbol{\mu}\|$ is an unit (projective) axis of rotation, if $\|\boldsymbol{\mu}\| \neq 0$. Having $\vartheta = 0$ implies the inequality in (B.32) given that $\mathrm{real}\big(\arcsin(\|\boldsymbol{\mu}\|)\big) \in [0, \pi/2]$. In turn, this implication yields the result that $\vartheta = 0$ if and only if $\|\boldsymbol{\mu}\| = 0$. Further, we have $\beta_\omega > 0$.

Pre-multiplying the second equation of (B.32) by $\mathbf{t}^\top$ and injecting (B.34), one obtains

$$\big[\big(\mathbf{I}_3 - \exp([\mathbf{u}\theta]_\times)\big)\underline{\mathbf{m}}^*\big]^\top \sin(\theta)\mathbf{u} = 0, \tag{B.35}$$

using the property

$$\mathbf{b}^\top [\mathbf{a}]_\times \mathbf{b} = 0. \tag{B.36}$$

Developing (B.35), we have

$$\underline{\mathbf{m}}^{*\top}\mathbf{u} = \exp([\mathbf{u}\theta]_\times)\, \underline{\mathbf{m}}^{*\top}\mathbf{u}. \tag{B.37}$$

Since $\mathbf{u}$ is an unit axis of rotation and $z^* > 0$ by definition, the only possible solutions to (B.37) are:

(*i*) $\theta = 0$;

(*ii*) $\theta = \pi$;

(*iii*) $\underline{\mathbf{m}}^{*\top}\mathbf{u} = 0, \ \forall\theta$.

This latter case signifies that $\underline{\mathbf{m}}^* = \exp([\mathbf{u}\theta]_\times)\underline{\mathbf{m}}^*$. However, using (B.34), all those cases imply $\mathbf{t} = \mathbf{0}$. Further, using this result in (B.32) implies that the only possible cases are in fact (*i*) $\theta = 0$ or (*ii*) $\theta = \pi$.

Then, we only need to show now that one must have $\theta = 0$. Using the implication that $\mathbf{t} = \mathbf{0}$, together with (B.22) and the identity

$$\cos(\theta) = \frac{1}{2}\big(\mathrm{tr}(\mathbf{R}) - 1\big), \tag{B.38}$$

permit to make conclusions from the inequality in (B.32):

$$\begin{aligned} \frac{1}{2}\big(\mathrm{tr}(\mathbf{H}) - 1\big) \geq 0 \implies & \frac{1}{2}\big(\beta_\omega \mathrm{tr}(\mathbf{R}) - 1\big) \geq 0 && \text{(B.39)} \\ \implies & \cos\big(\mathrm{real}(\theta)\big) \geq -\frac{1}{2} && \text{(B.40)} \\ \implies & |\theta| \leq \frac{2\pi}{3} && \text{(B.41)} \end{aligned}$$

since $\beta_\omega > 0$ and θ must be a real-valued scalar. Therefore, the only solution to (B.32) is $\mathbf{t} = \mathbf{0}$ and $\theta = 0$, $\forall \mathbf{q}^*$ and $\forall \underline{\mathbf{m}}^*$ such that $z^* > 0$. ∎

B.4 Proof of the Corollary 6.1

Proof (Generality and improvements). The generality of the proposed direct visual servoing technique regards to coping with rigid objects of unknown shape, and without requiring or estimating any of its metric attributes.

The homography-based technique proposed in (Benhimane and Malis, 2006a) is designed to cope with planar objects. Indeed, for a planar object defined by $\Pi = [\mathbf{n}^{*\top}, -d^*]^\top$ (though neither requiring nor estimating it), they propose to regulate the control error $\boldsymbol{\varepsilon}^{\mathrm{u}}_{\Pi} = \big[\boldsymbol{\varepsilon}^{\mathrm{u}\top}_{\nu\Pi}, \boldsymbol{\varepsilon}^{\mathrm{u}\top}_{\omega\Pi}\big]^\top \in \mathbb{R}^6$ with:

$$\begin{cases} \boldsymbol{\varepsilon}^{\mathrm{u}}_{\nu\Pi} = (\mathbf{H}_\Pi - \mathbf{I})\,\mathbf{m}^{*\prime} \\ [\boldsymbol{\varepsilon}^{\mathrm{u}}_{\omega\Pi}]_\times = \mathbf{H}_\Pi - \mathbf{H}_\Pi^\top. \end{cases} \tag{B.42}$$

Then, it is easy to show that, aside from coping with non-planar objects, the proposed control error (6.6) comprises (B.42) as well. Given that the object is planar, we have the parallax $\rho_i^* = 0$, $\forall \mathbf{m}^{*\prime}$. Moreover, the dominant plane for this particular target is in fact the plane Π on which the object lies, i.e.

$$\mathbf{q}^* = \mathbf{K}^{-\top}\mathbf{n}^*. \tag{B.43}$$

Applying this knowledge to (B.23) gives

$$\mathbf{q}^{*\prime} = \mathbf{K}^\top \mathbf{q}^* = \mathbf{K}^\top\mathbf{K}^{-\top}\mathbf{n}^* = \mathbf{n}^*. \tag{B.44}$$

For this particular case, the proposed task function (6.6) is rewritten as

$$\begin{cases} \boldsymbol{\varepsilon}^{\mathrm{u}}_{\nu} = (\mathbf{H} - \mathbf{I})\,\mathbf{m}^{*\prime} + \rho^*\mathbf{e}' = (\mathbf{H}_\Pi - \mathbf{I})\,\mathbf{m}^{*\prime} \\ \boldsymbol{\varepsilon}^{\mathrm{u}}_{\omega} = \vartheta \dfrac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|} = \dfrac{\vartheta_\Pi}{\|\boldsymbol{\mu}_\Pi\|}\boldsymbol{\mu}_\Pi, \end{cases} \tag{B.45}$$

where

$$
\begin{aligned}
\left[\boldsymbol{\mu}_\Pi\right]_\times &\propto \left[\sin(\theta)\mathbf{u} + \frac{1}{2}[\mathbf{n}^*]_\times \mathbf{t}\right]_\times && \text{(B.46)}\\
&\propto \mathbf{H}_\Pi - \mathbf{H}_\Pi^\top, && \text{(B.47)}
\end{aligned}
$$

by injecting (B.44) in (B.31), and then using (B.24).

Besides this generality, improvements in the behavior of the servoing is attained by using the proposed rotational control error, i.e. with

$$
\frac{\vartheta_\Pi}{\|\boldsymbol{\mu}_\Pi\|} \neq 1. \qquad \text{(B.48)}
$$

In fact, it explicitly determines in which quadrant the "projective angle of rotation" operates, instead of using simply (6.8) (or Eq. (B.42) for a planar object). This is particularly important for the initial conditions $\theta_0 > \pi/2$ and $\mathbf{t}_0 \approx \mathbf{0}$. In this situation, using for example (B.42) and then (B.46), we have $\boldsymbol{\varepsilon}^{\mathrm{u}}_{\omega\Pi} \approx 2\sin(\theta)\mathbf{u}$. Hence, the norm of this error is initially increased during the servoing since $\sin(\theta_0) < 1$ and $\sin(\theta) \to 1$ as $\theta \to \pi/2$, because the second quadrant is never specified. This may lead to system failure. In addition, this non-injection does not allow for a straightforward path planning. The control error (B.42) and the corresponding (6.8) are also non-injective around the equilibrium point if $\theta = \pi$ belongs to their codomain.

Furthermore, we remark that the knowledge of $\mathbf{q}^{*\prime}$ is not required in our projective framework, regardless of the object's shape. In an Euclidean framework, besides that $\mathbf{q}^{*\prime} = \mathbf{0}$ must be ensured, it requires perfect camera parameters. This setting is in fact only a stratum of the general relation. This demonstrates that our task function (6.6) is also a generalization of the hybrid control error $\boldsymbol{\varepsilon}^{\mathrm{u}\prime}_{\Pi} = \left[\boldsymbol{\varepsilon}^{\mathrm{u}\prime\top}_{\nu\Pi}, \boldsymbol{\varepsilon}^{\mathrm{u}\prime\top}_{\omega\Pi}\right]^\top \in \mathbb{R}^6$ with

$$
\begin{cases}
\boldsymbol{\varepsilon}^{\mathrm{u}\prime}_{\nu\Pi} = \left[\dfrac{z}{z^*}x' - x^{*\prime}, \dfrac{z}{z^*}y' - y^{*\prime}, \dfrac{z}{z^*} - 1\right]^\top = \alpha\underline{\mathbf{m}}' - \underline{\mathbf{m}}^{*\prime}\\
\boldsymbol{\varepsilon}^{\mathrm{u}\prime}_{\omega\Pi} = \theta\mathbf{u}
\end{cases}
\qquad \text{(B.49)}
$$

proposed in (Malis and Chaumette, 2002), with the advantage of not requiring the coarse metric estimate of the normal vector to perform its required partial Euclidean reconstruction (to recover $\boldsymbol{\varepsilon}^{\mathrm{u}\prime}_{\omega\Pi} = \mathbf{u}\theta$). That is, if $\mathbf{q}^{*\prime} = \mathbf{0}$ (or $\mathbf{t} = \mathbf{0}$) then (B.31) yields the equivalence

$$
\vartheta\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|} = \theta\mathbf{u}. \qquad \text{(B.50)}
$$

Also, the translational error $\boldsymbol{\varepsilon}^{\mathrm{u}\prime}_{\nu\Pi}$ in (B.49) is shown to be equivalent to ours by using the result (B.15), with $\beta_\nu = 1$ in (B.12) from their feature-based projective reconstruction. ■

B.5 Proof of the Theorem 6.2

To this end, we need to derivate the proposed control error $\boldsymbol{\varepsilon}^{\mathrm{u}} = \left[\boldsymbol{\varepsilon}_{\nu}^{\mathrm{u}\top}, \boldsymbol{\varepsilon}_{\omega}^{\mathrm{u}\top}\right]^{\top}$ in (6.6) with respect to time

$$\dot{\boldsymbol{\varepsilon}}^{\mathrm{u}} = \begin{bmatrix} \dot{\boldsymbol{\varepsilon}}_{\nu}^{\mathrm{u}} \\ \dot{\boldsymbol{\varepsilon}}_{\omega}^{\mathrm{u}} \end{bmatrix} = \mathbf{L}^{\mathrm{u}} \begin{bmatrix} \boldsymbol{\nu} \\ \boldsymbol{\omega} \end{bmatrix} = \mathbf{L}^{\mathrm{u}}\, \mathbf{v}, \tag{B.51}$$

in order to obtain the closed-loop equation for the proposed control law (6.10):

$$\dot{\boldsymbol{\varepsilon}}^{\mathrm{u}} = \mathbf{L}^{\mathrm{u}}\, \boldsymbol{\Lambda}\, \boldsymbol{\varepsilon}^{\mathrm{u}}, \tag{B.52}$$

where $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_{\nu}\mathbf{I}_3, \lambda_{\omega}\mathbf{I}_3), \lambda_{\nu}, \lambda_{\omega} > 0$ and $\mathbf{L}^{\mathrm{u}} \in \mathbb{R}^{6\times 6}$ denotes the interaction matrix, which is never used to perform the visual servoing if the proposed technique is applied. In this case, the matrix $\mathbf{L}^{\mathrm{u}}$ is necessary only for analysis purposes.

Proof (Local stability). The proof consists in analyzing the behavior of the closed-loop system (B.52) around the equilibrium $\boldsymbol{\varepsilon}^{\mathrm{u}} = \mathbf{0}$. Hence, only local stability will be demonstrated here. This is performed bellow by using the results from Lemma 6.1.

Let us start with $\boldsymbol{\varepsilon}_{\nu}^{\mathrm{u}}$:

$$\dot{\boldsymbol{\varepsilon}}_{\nu}^{\mathrm{u}} = (\dot{\mathbf{R}}\underline{\mathbf{m}}^{*} + \dot{\mathbf{t}})\frac{\beta_{\nu}}{z^{*}}. \tag{B.53}$$

By injecting the relation (1.5), i.e.

$$\dot{\mathbf{t}} = -\boldsymbol{\nu} - [\boldsymbol{\omega}]_{\times}\mathbf{t} \tag{B.54}$$

$$\dot{\mathbf{R}} = -[\boldsymbol{\omega}]_{\times}\mathbf{R} \tag{B.55}$$

in (B.53), and using (A.8) together with the property

$$[\mathbf{a}]_{\times}\mathbf{b} = -[\mathbf{b}]_{\times}\mathbf{a}, \tag{B.56}$$

we obtain

$$\dot{\boldsymbol{\varepsilon}}_{\nu}^{\mathrm{u}} = -\frac{\beta_{\nu}}{z^{*}}\boldsymbol{\nu} + \left[\boldsymbol{\varepsilon}_{\nu}^{\mathrm{u}} + \beta_{\nu}\underline{\mathbf{m}}^{*}\right]_{\times}\boldsymbol{\omega}. \tag{B.57}$$

With respect to $\boldsymbol{\varepsilon}_{\omega}^{\mathrm{u}}$, we have:

$$\dot{\boldsymbol{\varepsilon}}_{\omega}^{\mathrm{u}} = \beta_{\omega}\frac{d\sin(\theta)\mathbf{u}}{dt} + \frac{\beta_{\omega}}{2}[\mathbf{q}^{*\prime}]_{\times}\dot{\mathbf{t}}. \tag{B.58}$$

By using the relation

$$\frac{d\sin(\theta)\mathbf{u}}{dt} = -\mathbf{L}_{\omega}^{\mathrm{u}}\boldsymbol{\omega}, \tag{B.59}$$

with

$$\mathbf{L}_{\omega}^{\mathrm{u}} = \mathbf{I}_3 - \frac{\sin(\theta)}{2}[\mathbf{u}]_{\times} - \sin^2\left(\frac{\theta}{2}\right)(2\mathbf{I}_3 + [\mathbf{u}]_{\times}^2), \tag{B.60}$$

and Eqs. (B.54), (B.56), we obtain:

$$\dot{\boldsymbol{\varepsilon}}^{\mathrm{u}}_{\omega} = -\frac{\beta_\omega}{2}[\mathbf{q}^{*\prime}]_\times \boldsymbol{\nu} - \beta_\omega \Big(\mathbf{L}^{\mathrm{u}}_{\omega} - \frac{1}{2}[\mathbf{q}^{*\prime}]_\times [\mathbf{t}]_\times\Big)\boldsymbol{\omega}. \tag{B.61}$$

By using Eqs. (B.57) and (B.61), the interaction matrix is finally given as

$$\mathbf{L}^{\mathrm{u}} = \begin{bmatrix} -\dfrac{\beta_\nu}{z^*}\mathbf{I}_3 & [\boldsymbol{\varepsilon}_\nu + \beta_\nu \underline{\mathbf{m}}^*]_\times \\ -\dfrac{\beta_\omega}{2}[\mathbf{q}^{*\prime}]_\times & -\beta_\omega \mathbf{L}^{\mathrm{u}}_{\omega} + \dfrac{\beta_\omega}{2}[\mathbf{q}^{*\prime}]_\times [\mathbf{t}]_\times \end{bmatrix}. \tag{B.62}$$

Then, we may proceed to the evaluation of (B.52) around the equilibrium $\boldsymbol{\varepsilon}^{\mathrm{u}} = \big[\boldsymbol{\varepsilon}^{\mathrm{u}\top}_{\nu}, \boldsymbol{\varepsilon}^{\mathrm{u}\top}_{\omega}\big]^\top = \mathbf{0}$:

$$\dot{\boldsymbol{\varepsilon}}^{\mathrm{u}} = \mathbf{L}^{\mathrm{u}}\big|_{\boldsymbol{\varepsilon}^{\mathrm{u}}=\mathbf{0}} \boldsymbol{\Lambda}\, \boldsymbol{\varepsilon}^{\mathrm{u}} \tag{B.63}$$

$$= -\begin{bmatrix} \lambda_\nu \dfrac{\beta_\nu}{z^*}\mathbf{I}_3 & -\lambda_\omega \beta_\nu [\underline{\mathbf{m}}^*]_\times \\ \lambda_\nu \dfrac{\beta_\omega}{2}[\mathbf{q}^{*\prime}]_\times & \lambda_\omega \beta_\omega \mathbf{I}_3 \end{bmatrix} \boldsymbol{\varepsilon}^{\mathrm{u}}, \tag{B.64}$$

whose eigenvalues of $\mathbf{L}^{\mathrm{u}}\big|_{\boldsymbol{\varepsilon}^{\mathrm{u}}=\mathbf{0}} \boldsymbol{\Lambda}$ are given by

$$\begin{bmatrix} -\lambda_\omega \beta_\omega \\ -\lambda_\nu \dfrac{\beta_\nu}{z^*} \\ -\dfrac{\lambda_\omega \beta_\omega z^* + \lambda_\nu \beta_\nu - \sqrt{\Delta}}{2z^*} \\ -\dfrac{\lambda_\omega \beta_\omega z^* + \lambda_\nu \beta_\nu + \sqrt{\Delta}}{2z^*} \\ -\dfrac{\lambda_\omega \beta_\omega z^* + \lambda_\nu \beta_\nu - \sqrt{\Delta}}{2z^*} \\ -\dfrac{\lambda_\omega \beta_\omega z^* + \lambda_\nu \beta_\nu + \sqrt{\Delta}}{2z^*} \end{bmatrix} \tag{B.65}$$

with $\beta_\nu, \beta_\omega, z^* > 0$ and

$$\Delta = \lambda_\omega^2 \beta_\omega^2 z^{*2} + \lambda_\nu^2 \beta_\nu^2 - 2\lambda_\nu \lambda_\omega \beta_\nu \beta_\omega z^* \big(1 - z^* \mathbf{q}^{*\prime\top} \underline{\mathbf{m}}^*\big). \tag{B.66}$$

Therefore, if $\lambda_\nu > 0$, $\lambda_\omega > 0$, and

$$\Delta \;<\; (\lambda_\omega \beta_\omega z^* + \lambda_\nu \beta_\nu)^2, \tag{B.67}$$

whose substitution of (B.66) into (B.67) and using Thales' theorem gives

$$\mathbf{q}^{*\prime\top} \underline{\mathbf{m}}^* < 2, \tag{B.68}$$

then all eigenvalues of $\mathbf{L}^{\mathrm{u}}\big|_{\boldsymbol{\varepsilon}^{\mathrm{u}}=\mathbf{0}}\,\boldsymbol{\Lambda}$ shown in (B.65) have strictly negative real part. The condition (B.68) expresses the perpendicular distance between the chosen control point and the reference plane. Given that this reference plane represents the dominant plane of the object in our projective formulation, this condition can be easily satisfied if the control point is chosen such that its parallax ρ^* is sufficiently small. In fact, we could use simply a point which has $\rho^* = 0$ (since in the formulation the dominant plane always crosses the object). Consequently, the closed-loop system (B.52) is always locally asymptotically stable. ■

B.6 Proof of the Lemma 7.1

Proof (Existence and uniqueness). First of all, associated systems of normal equations

$$\bar{\mathbf{A}}^\top \bar{\mathbf{A}}\,\mathbf{x} = \bar{\mathbf{A}}^\top\,\bar{\mathbf{b}}, \tag{B.69}$$

are always consistent since $\bar{\mathbf{A}}^\top\,\bar{\mathbf{b}} \in \mathcal{R}(\bar{\mathbf{A}}^\top) = \mathcal{R}(\bar{\mathbf{A}}^\top\bar{\mathbf{A}})$, where $\mathcal{R}(\bar{\mathbf{A}}^\top)$ denotes the range of the matrix $\bar{\mathbf{A}}^\top$. Thus, we only need to proof the uniqueness of solution $\mathbf{x}$ for such a system under the stated conditions. The proof consists in demonstrating that the null space $\mathcal{N}(\bar{\mathbf{A}}^\top\bar{\mathbf{A}}) = \mathcal{N}(\bar{\mathbf{A}}) = \mathbf{0}$ or, equivalently, that $\bar{\mathbf{A}}$ is a full rank matrix if those conditions are verified.

We start by observing that $\mathbf{t} \neq \mathbf{0}$ is a necessary and sufficient condition to avoid a null coefficient matrix $\bar{\mathbf{A}}$. This can be seen directly from its submatrices in (7.11). In fact, as discussed in Subsection 1.3.2, if $\mathbf{t} = \mathbf{0}$ then the entire image corresponds to the plane at infinity $\boldsymbol{\pi}_\infty$, since there exists a solution such that

$$\lim_{\|\mathbf{x}\|\to 0^+} d^* = \frac{1}{\|\mathbf{K}^\top\mathbf{x}\|} = \infty, \tag{B.70}$$

using (7.8) and (7.10). Hence, actual Euclidean planes cannot be detected in this case.

However, this is a necessary condition but is not sufficient to guarantee that $\bar{\mathbf{A}}$ is a full rank, i.e. $\mathrm{rank}(\bar{\mathbf{A}}) = 3$ in that case. In fact, $\exists \mathbf{y} \neq \mathbf{0} : \bar{\mathbf{A}}\,\mathbf{y} = \mathbf{0}$ when the third image point is a linear combination of the first two, i.e. $\mathbf{p}_3^* = \alpha\mathbf{p}_1^* + \beta\mathbf{p}_2^*,\ \alpha,\beta \neq 0$. In this case, $\mathbf{y} = \gamma[\mathbf{p}_1^*]_\times\mathbf{p}_2^*,\ \forall\gamma \neq 0$, is such a vector. Therefore, if an image point is collinear with the others, then $\bar{\mathbf{A}}$ is also rank-deficient independently of the amount of translation. ■

B.7 Proof of the Theorem 7.1

Proof (Global stability). In standard 3D visual servoing techniques, the proof that the control law ensures asymptotic stability is straightforward. Let $V\colon \mathbb{R}^6 \to \mathbb{R}$ be a scalar function. In fact, it is immediate by using the Lyapunov candidate function

$$V(\boldsymbol{\varepsilon}^{\mathrm{c}}) = \frac{1}{2}\boldsymbol{\varepsilon}^{\mathrm{c}\top}\boldsymbol{\varepsilon}^{\mathrm{c}}, \tag{B.71}$$

which is radially unbounded: $\|\boldsymbol{\varepsilon}^{\mathrm{c}}\| \to \infty \implies V(\boldsymbol{\varepsilon}^{\mathrm{c}}) \to \infty$. However, only local stability is theoretically proved because the visibility constraints are not taken into consideration. It is well-known that servoing failure can happen under these techniques since no control is performed in the image.

On the other hand, under the assumption that new planes can be accurately identified (see Lemma 7.1 for the geometric conditions) if needed, then the efficient E-3D visual servoing ensures global asymptotic stability of the system since a dedicated identification algorithm (the Planar Region Detector) is employed. In this case, the time derivative of (B.71)

$$\dot{V}(\boldsymbol{\varepsilon}^{\mathrm{c}}) = \frac{\partial V(\boldsymbol{\varepsilon}^{\mathrm{c}})}{\partial \boldsymbol{\varepsilon}^{\mathrm{c}}}\, \dot{\boldsymbol{\varepsilon}}^{\mathrm{c}} \tag{B.72}$$

$$= \boldsymbol{\varepsilon}^{\mathrm{c}\top} \dot{\boldsymbol{\varepsilon}}^{\mathrm{c}} \tag{B.73}$$

is strictly negative in the large, as demonstrated bellow.

To this end, we need first to obtain the closed-loop system, which is obtained by deriving the control error $\boldsymbol{\varepsilon}^{\mathrm{c}} = \left[\mathbf{t}^{\top}, \theta\mathbf{u}^{\top}\right]^{\top} \in \mathbb{R}^6$ in (7.22) with respect to time,

$$\dot{\boldsymbol{\varepsilon}}^{\mathrm{c}} = \mathbf{L}^{\mathrm{c}}\,\mathbf{v} \tag{B.74}$$

where $\mathbf{L}^{\mathrm{c}} \in \mathbb{R}^{6\times 6}$ is the interaction matrix in the calibrated case, and then by applying the control law (7.23) in (B.74),

$$\dot{\boldsymbol{\varepsilon}}^{\mathrm{c}} = \mathbf{L}^{\mathrm{c}}\,\boldsymbol{\Lambda}\,\boldsymbol{\varepsilon}^{\mathrm{c}}, \tag{B.75}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_{\nu}\mathbf{I}_3, \lambda_{\omega}\mathbf{I}_3), \lambda_{\nu}, \lambda_{\omega} > 0$. The interaction matrix is obtained by using the relation (1.5), i.e.

$$\dot{\mathbf{t}} = -\boldsymbol{\nu} - [\boldsymbol{\omega}]_{\times}\mathbf{t} \tag{B.76}$$

$$\dot{\mathbf{R}} = -[\boldsymbol{\omega}]_{\times}\mathbf{R}, \tag{B.77}$$

together with the property

$$[\mathbf{a}]_{\times}\mathbf{b} = -[\mathbf{b}]_{\times}\mathbf{a}, \tag{B.78}$$

which give

$$\mathbf{L}^{\mathrm{c}} = \begin{bmatrix} -\mathbf{I}_3 & [\mathbf{t}]_{\times} \\ \mathbf{0} & -\mathbf{L}^{\mathrm{c}}_{\omega} \end{bmatrix}. \tag{B.79}$$

The interaction matrix $\mathbf{L}^{\mathrm{c}}_{\omega}$, which is related to the parametrization of the rotation

$$\frac{d(\theta\mathbf{u})}{dt} = -\mathbf{L}^{\mathrm{c}}_{\omega}\,\boldsymbol{\omega}, \tag{B.80}$$

is obtained using the Rodrigues' formula from (B.77):

$$\mathbf{L}^{\mathrm{c}}_{\omega} = \mathbf{I}_3 + \frac{\theta}{2}[\mathbf{u}]_{\times} + \left(1 - \frac{\mathrm{sinc}(\theta)}{\mathrm{sinc}^2(\frac{\theta}{2})}\right)[\mathbf{u}]^2_{\times}. \tag{B.81}$$

Finally, Equation (B.73) can be written as

$$\dot{V}(\boldsymbol{\varepsilon}^{\mathrm{c}}) = \left[\, \mathbf{t}^{\top}, \theta\mathbf{u}^{\top} \,\right] \left[\begin{array}{cc} -\lambda_{\nu}\mathbf{I}_3 & \lambda_{\omega}[\mathbf{t}]_{\times} \\ \mathbf{0} & -\lambda_{\omega}\mathbf{L}_{\omega}^{\mathrm{c}} \end{array} \right] \left[\begin{array}{c} \mathbf{t} \\ \theta\mathbf{u} \end{array} \right] \tag{B.82}$$

and therefore, given that $\lambda_{\nu}, \lambda_{\omega} > 0$,

$$\dot{V}(\boldsymbol{\varepsilon}^{\mathrm{c}}) = -\lambda_{\nu}\|\mathbf{t}\|^2 - \lambda_{\omega}\theta^2 \quad < 0, \quad \forall \boldsymbol{\varepsilon}^{\mathrm{c}} \in \mathbb{R}^6 \setminus \{\mathbf{0}\}, \tag{B.83}$$

using the properties $\mathbf{a}^{\top}[\mathbf{a}]_{\times} = [\mathbf{a}]_{\times}\mathbf{a} = \mathbf{0}$ and $\|\mathbf{u}\| = 1$. ■

Bibliography

Baillard, C. and Zisserman, A. (1999). Automatic reconstruction of piecewise planar models from multiple views, *Proc. of the IEEE Computer Vision and Pattern Recognition*, pp. 559–565.

Baker, S., Gross, R. and Matthews, I. (2003). Lucas-kanade 20 years on: A unifying framework: Part 3, *Technical Report CMU-RI-TR-03-35*, Carnegie Mellon University, USA.

Bartoli, A. (2006). Groupwise geometric and photometric direct image registration, *Proc. of the British Machine Vision Conference*.

Basri, R., Rivlin, E. and Shimshoni, I. (1999). Visual homing: surfing on the epipoles, *International Journal of Computer Vision* **33**(2): 22–39.

Benhimane, S. and Malis, E. (2004). Real-time image-based tracking of planes using Efficient Second-order Minimization, *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Benhimane, S. and Malis, E. (2006a). Homography-based 2D visual servoing, *Proc. of the IEEE International Conf. on Robotics and Automation*, USA.

Benhimane, S. and Malis, E. (2006b). Integration of Euclidean constraints in template based visual tracking of piecewise-planar scenes, *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, China, pp. 1218–1223.

Benhimane, S., Malis, E., Rives, P. and Azinheira, J. R. (2005). Vision-based control for car platooning using homography decomposition, *Proc. of the IEEE International Conference on Robotics and Automation*, Spain.

Berger, M.-O. and Simon, G. (1998). Robust image composition algorithms for augmented reality, *Proc. of the Asian Conference on Computer Vision*, China, pp. 360–367.

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, *SIGGRAPH*, pp. 192–198.

Broida, T. J., Chandrashekhar, S. and Chepalla, R. (1990). Recursive 3-D motion estimation from a monocular image sequence, *IEEE Transactions on Aerospace and Electronic Systems* **26**(4): 639–656.

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

Bruss, A. R. and Horn, B. K. P. (1983). Passive navigation, *Computer Vision, Graphics, and Image Processing* **21**: 3–20.

Bueno, S. S., Azinheira, J. R., Ramos, J. J. G., de Paiva, E. C., Carvalho, J. R. H., Rives, P., Elfes, A. and Silveira, G. F. (2002). Project AURORA: Towards an autonomous robotic airship, *Workshop on Aerial Robotics, IEEE/RSJ International Conference on Intelligent Robots and Systems*, Switzerland, pp. 43–53.

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

Chaumette, F. (1998). Potential problems of stability and convergence in image-based and position-based visual servoing, *in* D. J. Kriegman, G. D. Hager and A. S. Morse (eds), *The Confluence of Vision and Control*, Vol. 237 of *LNCIS*, Springer-Verlag, pp. 66–78.

Chaumette, F. and Hutchinson, S. (2006). Visual servo control part I: Basic approaches, *IEEE Robotics & Automation Magazine* pp. 82–90.

Cobzas, D. and Sturm, P. (2005). 3D SSD tracking with estimated 3D planes, *Proc. Canadian Conf. on Comp. and Rob. Vision*, pp. 129–134.

Collewet, C., Marchand, E. and Chaumette, F. (2008). Visual servoing set free from image processing, *Proc. of the IEEE International Conference on Robotics and Automation*, USA.

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

Comport, A., Malis, E. and Rives, P. (2007). Accurate quadrifocal tracking for robust 3D visual odometry, *Proc. of the IEEE International Conference on Robotics and Automation*, Italy.

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

Davison, A. (2003). Real-time simultaneous localization and mapping with a single camera, *Proc. of the IEEE International Conference on Computer Vision*.

Dennis, J. E. and Schnabel, R. B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Classics in Applied Mathematics 16, SIAM.

Dick, A., Torr, P. and Cipolla, R. (2000). Automatic 3D modelling of architecture, *Proc. of the British Machine Vision Conference*, pp. 372–381.

Eade, E. and Drummond, T. (2006). Scalable monocular SLAM, *Proc. of the IEEE Computer Vision and Pattern Recognition.*

Espiau, B., Chaumette, F. and Rives, P. (1992). A new approach to visual servoing in robotics, *IEEE Transactions on Robotics and Automation* **8**(3): 313–326.

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

Faugeras, O. and Lustman, F. (1988). Motion and structure from motion in a piecewise planar environment, *International Journal of Pattern Recognition and Artificial Intelligence* **2**(3): 485–508.

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

Fischler, M. and Bolles, R. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, *Communications of the ACM* **24**: 381–385.

Galambos, C., Matas, J. and Kittler, J. (1999). Progressive probabilistic Hough transform for line detection, *Proc. of the IEEE Computer Vision and Pattern Recognition*, pp. 554–560.

Gouiffès, M., Collewet, C., Fernandez-Maloigne, C. and Trémeau, A. (2006). Feature points tracking using photometric model and colorimetric invariants, *Proc. of the European Conference on Colour in Graphics, Imaging, and Vision*, pp. 18–23.

Hager, G. and Belhumeur, P. (1998). Efficient region tracking with parametric models of geometry and illumination, *IEEE PAMI* **20**(10): 1025–1039.

Hall, B. (2003). *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*, Vol. 222 of *Graduate Texts in Mathematics*, Springer.

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 Transactions on Pattern Analysis and Machine Intelligence* **23**(6).

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

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

Hummel, R. and Sundareswaran, V. (1993). Motion parameter estimation from global flow field data, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15**(5): 459–476.

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

Isaacson, E. and Keller, H. (1966). *Analysis of numerical methods*, John Wiley & Sons.

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, *Proc. of the British Machine Vision Conference*, UK, pp. 123–131.

Kallem, V., Dewan, M., Swensen, J., Hager, G. and Cowan, N. (2007). Kernel-based visual servoing, *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, USA.

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 Transactions on Pattern Analysis and Machine Intelligence* **22**: 322–336.

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

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision* pp. 91–110.

Lucas, B. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision, *Proc. of the International Joint Conference on Artificial Intelligence*, pp. 674–679.

Luenberger, D. G. (1984). *Linear and Nonlinear Programming*, Addison-Wesley.

Ma, Y., Soatto, S., Kosecka, J. and Sastry, S. S. (2003). *An Invitation to 3-D Vision: From images to Geometric Models*, Springer-Verlag.

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, *Proc. of the IEEE International Conference on Robotics and Automation*, USA.

Malis, E. (2007). An efficient unified approach to direct visual tracking of rigid and deformable surfaces, *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, USA.

Malis, E. and Chaumette, F. (2002). Theoretical improvements in the stability analysis of a new class of model-free visual servoing methods, *IEEE Transactions on Robotics and Automation* **18**(2): 176–186.

Malis, E., Chaumette, F. and Boudet, S. (1999). 2D 1/2 visual servoing, *IEEE Transactions on Robotics and Automation* **15**(2): 238–250.

Malis, E. and Rives, P. (2003). Robustness of image-based visual servoing with respect to depth distribution errors, *Proc. of the IEEE International Conference on Robotics and Automation*.

Maya-Mendez, M., Morin, P. and Samson, C. (2006). Control of a nonholonomic mobile robot via sensor-based target tracking and pose estimation, *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5612–5618.

Meer, P. (2004). *Emerging Topics in Computer Vision*, Prentice Hall, chapter Robust techniques for computer vision.

Mei, C., Benhimane, S., Malis, E. and Rives, P. (2006). Constrained multiple planar template tracking for central catadioptric cameras, *Proc. of the British Machine Vision Conference*.

Mezouar, Y. and Chaumette, F. (2002). Path planning for robust image-based control, *IEEE Transactions on Robotics and Automation* **18**: 534–549.

Molton, N. D., Davison, A. J. and Reid, I. D. (2004). Locally planar patch features for real-time structure from motion, *Proc. of the British Machine Vision Conference*.

Montemerlo, M., Thrun, S., Koller, D. and Wegbreit, B. (2003). FastSLAM 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges, *Proc. of International Joint Conference on Artificial Intelligence*, pp. 1151–1156.

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.

Morin, P. (2004). Stabilisation de systèmes non linéaires critiques et application à la commande de véhicules, *Habilitation à diriger des recherches*, Université de Nice-Sophia Antipolis.

Nastar, C., Moghaddam, B. and Pentland, A. (1996). Generalized image matching: Statistical learning of physically-based deformations, *Proc. of the European Conference on Computer 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.

Nistér, D. (2003). An efficient solution to the five-point relative pose problem, *Proc. of the IEEE Computer Vision and Pattern Recognition*, Vol. 2, pp. 195–202.

Okada, K. et al. (2001). Plane segment finder: Algorithm, implementation and applications, *Proc. of the IEEE International Conference on Robotics and Automation*, pp. 2120–2125.

Rives, P. (2000). Visual servoing based on epipolar geometry, *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems Robots and Systems*, pp. 602–607.

Saeedi, P., Lawrence, P. D. and Lowe, D. G. (2006). Vision-based 3-D trajectory tracking for unknown environments, *IEEE Transactions on Robotics* **22**(1): 119–136.

Samson, C., Espiau, B. and le Borges, M. (1990). *Robot Control: the Task Function Approach*, Oxford University Press.

Siciliano, B. and Khatib, O. (eds) (2008). *Springer Handbook of Robotics*, Springer.

Silveira, G. F., Carvalho, J. R. H., Rives, P., Azinheira, J. R., Bueno, S. S. and Madrid, M. K. (2002). Optimal visual servoed guidance of outdoor autonomous robotic airships, *Proc. of the American Control Conference*, USA, pp. 779–784.

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

Silveira, G. and Malis, E. (2007b). Direct visual servoing with respect to rigid objects, *Research Report 6265*, INRIA.
URL: *https://hal.inria.fr/inria-00166417/en*

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

Silveira, G. and Malis, E. (2007d). Suivi visuel efficace et robuste aux changements d'éclairage quelconques, *Proc. of the Journées ORASIS*, France.

Silveira, G. and Malis, E. (2008). L'asservissement visuel direct, *Proc. of the Reconnaissance des Formes et Intelligence Artificielle*, France.

Silveira, G., Malis, E. and Rives, P. (2006a). Real-time robust detection of planar regions in a pair of images, *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, China, pp. 49–54.

Silveira, G., Malis, E. and Rives, P. (2006b). Visual servoing over unknown, unstructured, large-scale scenes, *Proc. of the IEEE International Conference on Robotics and Automation*, USA, pp. 4142–4147.

Silveira, G., Malis, E. and Rives, P. (2007). An efficient direct method for improving visual SLAM, *Proc. of the IEEE International Conference on Robotics and Automation*, Italy.

Silveira, G., Malis, E. and Rives, P. (2008a). An efficient direct approach to visual SLAM, *IEEE Transactions on Robotics* **24**(5): 969–979. Special issue on Visual SLAM.

Silveira, G., Malis, E. and Rives, P. (2008b). The efficient E-3D visual servoing, *International Journal of Optomechatronics* **2**(3): 166–184. Special Issue on Visual Servoing.

Silveira, G., Malis, E. and Rives, P. (2008c). Une approche de SLAM visuel direct, *Proc. of the Reconnaissance des Formes et Intelligence Artificielle*, France.

Simon, G. and Berger, M.-O. (2002). Pose estimation for planar structures, *IEEE Comput. Graph. Appl.* **22**(6): 46–53.

Simon, G. and Berger, M.-O. (2008). Reconstruction et augmentation simultanées de scènes planes par morceaux, *Proc. of the Reconnaissance des Formes et Intelligence Artificielle*, France.

Smith, R. C. and Cheeseman, P. (1986). On the representation and estimation of spatial undertainty, *Int. J. of Rob. Res.* **5**(4): 56–68.

Stein, G. and Shashua, A. (2000). Model-based brightness constraints: On direct estimation of structure and motion, *Transactions on Pattern Analysis and Machine Intelligence* **22**: 992–1015.

Stewart, C. V. (1999). Robust parameter estimation in computer vision, *SIAM Rev.* **41**: 513–537.

Szeliski, R. (2005). Image alignment and stitching, *in* N. Paragios, Y. Chen and O. Faugeras (eds), *Handbook of Mathematical Models in Computer Vision*, Springer, pp. 273–292.

Szeliski, R. and Torr, P. H. S. (1998). Geometrically constrained structure from motion: Points on planes, *Proc. of the Workshop on 3D Structure from Multiple Images of Large-Scale Environments*, pp. 171 – 186.

Thuilot, B., Martinet, P., Cordesses, L. and Gallice, J. (2002). Position based visual servoing: Keeping the object in the field of vision, *Proc. of the IEEE International Conference on Robotics and Automation*, pp. 1624–1629.

Tomasi, C. and Kanade, T. (1992). Shape and motion from image streams under orthography: A factorization method, *International Journal on Computer Vision* **9**(2): 137–154.

Torr, P. H. S. and Zisserman, A. (1999). Feature based methods for structure and motion estimation, *Workshop on Vision Algorithms: Theory and Practice*, pp. 278–294.

Tsai, R. (1987). Versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses, *IEEE Journal of Robotics and Automation* **3**(4): 323–344.

Tuytelaars, T. and Van Gool, L. J. (2004). Matching widely separated views based on affine invariant regions, *International Journal of Computer Vision* **59**(1): 61–85.

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

Vargas, M. and Malis, E. (2005). Visual servoing based on an analytical homography decomposition, *Proc. of the Joint IEEE Conference on Decision and Control and European Control Conference*, Spain.

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

Weiss, L. E. and Anderson, A. C. (1987). Dynamic sensor-based control of robots with visual feedback, *IEEE Journal of Robotics and Automation* **3**(5): 404–417.

Wilson, W. J., Hulls, C. C. W. and Bell, G. S. (1996). Relative end-effector control using Cartesian position based visual servoing, *IEEE Transactions on Robotics and Automation* **12**(5): 684–696.

Xu, L. and Oja, E. (1993). Randomized Hough Transform (RHT): Basic mechanisms, algorithms, and computational complexities, *CVGIP: Image Understanding* **57**(2): 131–154.

Zhang, Z. (2000). A flexible new technique for camera calibration, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**: 1330–1334.

Abstract

The overwhelming majority of vision-based techniques for both estimation and control consider a feature-based scheme. This thesis investigates how to appropriately exploit pixel intensities directly, i.e. without having to resort to image features. The fact of using all image information, even from where no features exist, can considerably increase their accuracy and flexibility.

To this end, we propose generic photo-geometric transformation models and optimization methods for directly and efficiently registering images (including color ones) of rigid and deformable objects, all in a unified manner. In particular, the new photometric model ensures robustness to arbitrary illumination changes, are independent of the object's attributes and of the camera's characteristics, and naturally encompasses gray-level images. We then show that the framework can effectively be formulated using uncalibrated or calibrated pinhole cameras. The differences mainly regard to the needed parametrization.

A robust visual tracking technique is constructed by directly registering a reference image with successive frames. Then, using the optimal parameters that relate the reference image to the current one, a vision-based control strategy is proposed to drive all six degrees-of-freedom of a robot to the (desired) pose where the reference image was taken. This new technique does not require either precise parameters of the vision system or any metric structure of the observed rigid scene, leading to a flexible and reliable system.

If a calibrated camera is used, then the proposed robust visual tracking technique directly provides the optimal camera pose and scene structure. Since they are simultaneously and causally recovered, the technique represents a new solution to the visual Simultaneous Localization and Mapping (SLAM) problem. Finally, we propose a new visual servoing method that uses the estimates from this visual SLAM approach. Hence, this controlled visual SLAM scheme allows for autonomous navigation of mobile robots over previously unexplored scenes.

Comparisons results with existing techniques demonstrate significant improvements in the system performance. Various real-world experiments and simulations are reported to show that the proposed methods can indeed be highly accurate and robust despite unknown objects and unknown imaging conditions. The trade-offs to attain real-time efficiency are discussed in the text.

Keywords: Vision-based estimation, vision-based control, robotics, image registration, computer vision, visual tracking, visual SLAM, visual servoing.

Résumé

Dans leur grande majorité, les techniques d'estimation et de commande basées sur la vision s'appuient sur l'extraction d'informations géométriques dans les images. L'objectif de cette thèse, rédigée en anglais, est de développer une nouvelle approche exploitant directement l'intensité des pixels dans l'image en s'affranchissant de l'étape d'extraction de ces informations. Nous espérons montrer que le fait d'utiliser toute l'information contenue dans l'image permet en outre d'augmenter la précision et le domaine d'application.

Dans ce but, nous proposons un modèle générique de transformation prenant à la fois en compte les aspects géométriques et photométriques. Ce modèle est associé à une méthode efficace d'optimisation pour le recalage d'images, valide pour des modes d'acquisition variés (incluant les images couleurs) et pour des classes d'objets rigides ou déformables. En particulier, le nouveau modèle photométrique assure une robustesse aux variations d'éclairage quelconques, et il est indépendant des attributs des objets et des caractéristiques de la caméra. Ce cadre méthodologique est formulé, dans le cas d'un modèle sténopé, à la fois dans le cas calibré et non calibré; les différences portant principalement sur la nature de la paramétrisation choisie.

Une méthode robuste de suivi visuel est proposée permettant le recalage d'une image de référence tout au long de la séquence. A partir des paramètres estimés liant l'image de référence à l'image courante, nous proposons une nouvelle stratégie d'asservissement visuel permettant de contrôler les six degrés de liberté du mouvement de la caméra pour l'amener dans la pose où a été acquise l'image de référence. Cette nouvelle approche ne nécessite pas de connaissance précise sur les paramètres de la caméra ni sur la géométrie de l'objet observé, permettant ainsi d'obtenir une méthode générique et fiable.

Dans le cas de l'utilisation d'une caméra calibrée, la méthode de suivi robuste permet d'accéder directement à la pose de la caméra et à la structure géométrique de la scène. Elle peut donc être appliquée pour proposer une nouvelle solution au problème de SLAM (pour *Simultaneous Localization and Mapping*) visuel. Enfin, nous présentons une méthode d'asservissement visuel intégrant directement les estimées fournies par la méthode de suivi et permettant ainsi la navigation autonome de robots dans un environnement inconnu a priori.

Les méthodes développées tout au long de cette thèse ont été confrontées aux approches classiques de la littérature, et ont montré des avantages certains. Elles ont également été testées en condition réelle sur des séquences caractéristiques de différentes applications et dans des conditions variées. Les conditions et compromis à faire pour obtenir performances temps réel et précision, sont également discutés dans le document.

Mots-clefs : Estimation basée sur la vision, commande basée sur la vision, recalage d'image, vision par ordinateur, automatique, robotique.