3D Pose estimation of continuously deformable instruments in robotic endoscopic surgery

Paolo Cabras

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE ED 269 - Mathématiques, Sciences de l'Information et de l'Ingénieur

ICube, Equipe AVR (Automatique Vision et Robotique) - UMR 7357

THÈSE présentée par :

Paolo CABRAS

soutenue le : **24/02/2016**

pour obtenir le grade de: **Docteur de l'Université de Strasbourg**

Discipline/ Spécialité: **Robotique, Vision et Automatique**

3D Pose Estimation of Continuously Deformable Instruments in Robotic Endoscopic Surgery.

Rapporteurs

M. Guillaume Morel	Professeur, Université Pierre et Marie Curie, Paris
M. Nicolas Andreff	Professeur, Université de Franche-Comté, Besançon

Examinateurs

Mme. Alicia Casals	Professeur, Universitat Politècnica de Catalunya, Barcelone
M. Michel de Mathelin	Professeur, Université de Strasbourg
M. Florent Nageotte	Maître de Conférence, Université de Strasbourg, co-encadrant

Directeur de thèse

M. Christophe Doignon	Professeur, Université de Strasbourg

A Rosaura e Mateo,
inesauribili fonti di gioia.
Ai miei genitori e mia sorella,
inestimabili consiglieri di una vita.

Acknowledgments

Dopo quasi 8 anni lontano dalla terra natìa, mi voglio riappropriare della mia lingua in questi ringraziamenti di tesi e riassaporare la leggerezza e il piacere di scrivere senza pensare a quale accento va messo o che parola è la più adatta o come si scrive... come si dice...

Le persone (e organismi) che hanno reso questa tesi possibile sono molteplici e mi piacerebbe ringraziare ognuna di loro nella loro lingua o almeno nella lingua in cui abbiamo condiviso questo tempo e la nostra amicizia.

Avant tout, je voulais remercier ce qui a rendu possible mon travail à temps plein au sein de la recherche : un grand merci au Labex CAMI pour avoir soutenu financièrement cette thèse de trois ans avec des fonds gérés par l'ANR dans le programme "Investissement d'Avenir" sous la référence ANR-11-LABX-0004. Je remercie aussi l'IRCAD pour l'accès aux images endoscopiques in vivo et Karl Storz pour nous avoir fourni l'endoscope Anubis. Grâce à ce matériel, j'ai eu la possibilité d'expérimenter mes méthodes dans des conditions réelles, donnant ainsi, à ma thèse, des résultats plus forts.

Cette thèse n'aurait jamais été possible sans l'aide permanente de mes encadrants. Merci à Christophe Doignon, mon directeur de thèse, pour toujours m'avoir motivé à faire mieux et à me demander "pourquoi ?" à chaque pas pour comprendre en profondeur les problématiques et faire de mon travail, un travail de recherche de qualité. Merci à Florent Nageotte pour l'aide et le support de chaque jour, merci pour tous tes enseignements et pour ta forte implication comme si ce travail était le tien. Enfin, merci pour le soutien et l'amitié que tu m'as offerts chaque jour depuis le début.

Je ne peux pas ne pas remercier Philippe Zanne pour sa perpétuelle disponibilité, patience et gentillesse, merci de m'avoir expliqué tous les secrets du robot STRAS et de m'avoir conseillé les meilleures solutions à mes problèmes techniques. Grâce à toi, la validation avec le système stéréo a été possible.
Merci à David Goyard pour sa contribution précieuse à mon travail et le temps qu'il m'a dédié pour partager avec moi ses idées et ses connaissances sur le learning.
Merci, aussi, à Lucile Zorn et à Florent Le Bastard de m'avoir aidé avec toutes mes questions mécaniques, les plans du robot et l'impression 3D de pièces fondamentales pour mon travail.

Un énorme merci à toute l'équipe AVR du laboratoire ICUBE pour avoir fait de cette expérience de travail aussi une expérience forte de vie et d'amitié. Merci à chacune et à chacun de vous et, en particulier, à ceux qui ont commencé avec moi ce chemin du doctorat : Nadège, Arnaud et Laure-Anaïs qui m'ont fait me sentir français depuis le début (même si je ne parlais aucun mot de français à l'époque) et ont toujours été disponibles pour me donner un coup de main.
Merci à Laure Esteveny pour avoir été la maman présumée du groupe des doctorants quand je suis arrivé, et, merci à tous les doctorants qui ont rendu la création d'une

vraie petite famille possible.
A particular thanks goes to Rahim Kadkhodamohammadi for the great patience he had in explaining me learning methods and graph-based segmentation theory. Thanks to you and your beautiful family for our pleasant meetings... we should have done that more often.

Fuori dall'ambiente lavorativo sono molte le persone che mi hanno offerto quelle parole e quegli abbracci che mi hanno aiutato a rinnovare le mie forze e l'entusiasmo.

Grazie al movimento dei Focolari e in particolare a Guido, Mario, Antonella, Ita e Rosa che mi hanno sempre offerto una casa (in senso lato) dove ristorarmi.

Grazie ai miei amici italiani che nonostante la distanza mi sono sempre stati vicini in maniera concreta e costante. Grazie agli amici del liceo e al Re che ha saputo essere chioccia perfetta di un gruppo sempre unito che mi ha insegnato il vero significato dell'amicizia... un amicizia sincera e gratuita dal valore inestimabile. Grazie Valentina per essere un'amica sempre presente e interessata, grazie per le lunghe appassionanti e ricche chiacchierate e le mail di pettegolezzi che mi facevano restare con un piede in Italia.

Gracias a Pilar, Mark y Micaela... nos hemos dado fuerza, entusiasmo y confianza recíprocamente como solo las familias saben hacer, gracias por haber creado momentos de hogar juntos.

Un énorme merci à mes deux collègues de bureau : Nicole et Markus. Merci pour toutes les aides, tous les conseils, tous les mots et toute l'amitié que vous nous avez offerts, à ma famille et à moi. Merci pour cette perle rare : une amitié gratuite, patiente et affectueuse. Merci de tout coeur.

Un grande grazie alla mia famiglia d'origine: i miei genitori e mia sorella... per avermi educato e sostenuto in tutta la mia vita, reso felice e entusiasta della vita stessa e dei suoi valori più profondi. Grazie per avermi spinto sempre oltre e avermi offerto la possibilità di studiare e uscire di casa per scoprire la mia strada.

Grazie alla mia famiglia attuale. Gracias a mi maravillosa esposa Rosaura por ser amiga y compañera de vida, gracias por el apoyo constante, por escuchar y colmar mis miedos e inseguridades, por hacerme recordar mis verdaderas motivaciones profesionales, por hacer especial cada cosa que vivimos juntos y por enseñarme a disfrutar de la vida desde sus más pequeñas cosas. Grazie Mateo, figlio mio, per meravigliarmi ogni giorno con i tuoi sorrisi, con le tue nuove scoperte, i tuoi abbracci e il tuo entusiasmo. Hai fatto del nostro matrimonio una famiglia unica. Compartir la vida con vosotros es una maravilla y un regalo que llena cotidianamente.

Infine, grazie a Dio che ha accompagnato con misericordia e affetto la mia vita e ha dimostrato nei piccoli dettagli il suo amore personale.

Contents

Chapter 1

Introduction

Contents

With the advent of Minimally Invasive Surgery (MIS), the gaze of medicine on the patient health has become wider and start focusing not only on the success of the operation, but also on the post-operative and aesthetic sides. The dimension and number of scars after operation is, now, a sort of evaluation paradigm of the operation: the less invasive the operation is, the scanter the blood loss, the less the surgery and post-operative pain and the shorter the convalescence period are.

The first example of MIS dates back to 1985, when Eric Muhe performed the first LC (Laparoscopic Cholecystectomy). He encountered great skepticism by the German Medical Community, where the general surgical thinking was oriented to the concept that big problems required big incisions [Litynski 1998]. Eric Muhe work on LC was never published and for many years the merit for the first LC was attributed to Philippe Mouret who performed it in 1987 [Cuschieri 1991]. However, even in the case of Mouret the work had weak responses and the diffusion of these techniques was more due to private communities or private practice surgeons. Nowadays, on the contrary, Laparoscopy is a diffuse technique and, probably, the most performed among the MIS techniques. In conventional Laparoscopy three or four incisions are made on the skin of the patient to access the body cavity: one is usually for the camera (laparoscope) and the other for rigid instruments. With the aim of limiting the invasiveness, LESS (Laparo-Endoscopic Single Site) surgery reduces the incisions to a single one: in this procedure all the instruments and the camera are inserted in the body cavity through a single port [Curcillo 2010].

Along the same line, in the last decade, new surgical techniques have been developed with the aim of having no scar at all (or, at least, no external scar) leading to a family of techniques such as endoluminal surgery or natural orifice transluminal surgery (also known as NOTES). The idea underneath both of these techniques is to exploit the natural orifices and tracts to access the wanted surgical site by the

interior of the body. In endoluminal surgery, the operation is carried out in the interior of a tubular anatomical structure (lumen) such as urethra, colon or esophagus and no superficial incision is necessary since these structure are accessible through natural orifices. In the transluminal surgery, the lumen is used as a via to approach the site of the operation which is finally reached performing an incision in the wall of the tract that has been walked through.

These kinds of techniques are made possible by the utilization of flexible systems. Thanks to their compliance, they can reach the zone of interest in a really non-invasive way. Differently from the rigid instruments, they can enter through the natural orifices and walk through tortuous paths adapting to anatomical constraints and assuring almost atraumatic endoluminal movements. For these purposes, several devices are being developed which, often, are extension or adaptation of the already available technologies (e.g. endoscopes) whose main use was for diagnostic tasks (see Ch. 2)..

With respect to conventional surgical operations, the techniques using flexible instruments present a higher degree of difficulty. The relationship between the hand gestures and the actual instruments trajectories are not easily intelligible if the surgeon only relies on the endoscopic image and, furthermore, the large distance between operator and end-effectors, which is typical in flexible endoscopy, limits haptic feedback. Maneuverability also decreases due to the narrow space where camera and instruments are constrained. Moreover, positioning and field of view are maintained by a complex coordination of positioning-wheels and locking mechanism which are usually controlled by an assistant, while the surgeon operates the endoscopic instruments which move inside the channels of the endoscope. The proximity of instruments controlling handles and navigation/view control elements and the linkage of camera and instruments movements (due to the fact that the same flexible system houses the camera and the instruments) requires a solid team interaction which is not obvious. Additionally, mental workload will likely be increased by fluctuating visual frames of reference and angles of approach associated with NOTES procedures.

1.1 Motivation and Problem Statement

In this context, robotic assistance can provide large advantages in terms of manipulation easiness, ergonomics, accuracy in the gesture and repeatability [Dogangil 2010]. In the majority of the robotic systems for endoscopic (flexible or laparoscopic) surgery, the robot is thought within a teleoperation framework where the surgeon remotely controls the bedside slave system (cf. Ch. 2). In simple Master - Slave configuration, the robot is just another instrument that repeats the gestures performed by the physician with the specific human-machine interface. To upgrade the robot to be an assistant for the surgeon, autonomous movements control must be

contemplated in the medical robot scenario. The next challenge, then, is providing the robots with automatic control that can help following a particular trajectory, limit the permitted movement to the zone of interest, pre-positioning the instrument for a better incipit of a specific action during an operation or, later on, accomplishing some tasks autonomously. This could both increase the precision of the surgeon movements and considerably decrease the fatigue of the surgeon. Furthermore, the movements of a specific operation can be recorded and an a-posteriori analysis of them can be exploited for educational scopes.

To accomplish this project, the configuration of the robot must be known accurately to compute the appropriate control action. Nevertheless, obtaining the actual configuration and pose of these robotized flexible instruments is difficult. Mainly because of the transmission of the movement inside the flexible body, these systems have often a complex behavior with strong and variable non-linearities depending on several parameters. It is difficult, then, to determine a precise model and, consequently, model-based open-loop control strategies are not satisfactory. To obtain a good accuracy in instrument positioning, an external measurement seems to be necessary to close the loop.

1.2 Objectives

In this context, our main objective is to study the potentialities of using the embedded monocular vision to provide the 3D pose measurement of flexible surgical instruments.

Vision-based methods usually rely on particular visual features related to the 3D object whose pose has to be estimated and the precision of the pose estimation is indirectly affected by the accuracy attain in extracting these features. Therefore, the second objective of this thesis is the development of a feature extraction method that has to be effective in in-vivo environment and robust to the peculiar illumination conditions (strong central enlightenment and darker border zone) and possibly able to manage a partial loss of visual information.

In this work, the attention will be focused on those flexible systems provided by an embedded monocular vision system and at least one working channel where an instrument can slide. The flexible instrument in question is considered to be a conventional surgical instrument currently used in manually performed NOTES or endoluminal surgery characterized by a single bendable section, a small diameter. As a first approximation it is supposed to move within a free space and no contacts nor occlusion are considered.

1.3 Thesis Overview

The solutions proposed to attain the enunciated objectives is presented in details in this manuscript which is structured in 5 main chapters.

One introductory chapter (**chapter 2**) is dedicated to the description of the medical context underlying the role of flexible systems and the robotic solutions existing in this field. This kind of robots is usually inscribed in the *continuum robot* category and presents strong peculiarities both in the structure and drive. After a brief presentation of their characteristics and the main issues, state of the art works are presented concerning the measurements used for controlling continuum robots.

The four other chapters are arranged in two parts coinciding with two of the aspects of a visual-based pose estimation process: (I) choice and segmentation of visual features for pose estimation and (II) actual 3D pose inferring from a single 2D image visual features.

The beginning of **Part I** is dedicated to the detailed description of the flexible system taken as specimen, focusing on the visual perception of the flexible instrument by the endoscopic camera.

In **Chapter 3**, a preliminary theoretical study is carried out to investigate the suitability of some visual features for pose estimation. The suitability of the visual feature is defined based on the sensitivity of the 3D pose estimation with respect to the error in the segmentation process. Within this scope, the direct and inverse kinematics of a conventional single bending section flexible instrument are discussed assuming a constant curvature model. Furthermore, the geometric Jacobian (relating the variation of the robot degrees of freedom to the variation of a point of the robot) and the image Jacobian (relating the velocity of a 3D point with the velocity of its projection in the image) are computed considering, also, the effects of camera lens distortion.

The problem of extracting the visual features suggested by the theoretical study (i.e. the apparent corners of colored markers attached to the bendable section of the instrument) is detailed in **Chapter 4**. The possibility of using color information for feature segmentation is analyzed trying to manage those cases where color information is only partially available (e.g. overexposure regions due to strong frontal illumination). The main proposed idea is to detect the corners points as those points along the superior and inferior apparent borders where color transition (between two markers) occurs. A continuous representation of the instrument borders is proposed to achieve a subpixel accuracy in extracting these points. The results of in-vivo experiments are finally shown and commented.

In **part II** two novel methods are described to infer 3D information from 2D images for flexible instruments with the monocular endoscopic camera.

In **Chapter 5** we analyze the potentialities of estimating the 3D pose relying on the use of a parametrized kinematic model of the system. This problem can be formulated as an optimization problem intended to minimize the differences between the synthetic (built starting from camera and kinematic model) and the actual image.

The interest of considering additional degrees of freedom in the kinematic model is also addressed with the intention of modeling the mechanical play between the instrument and its housing working channel. These plays are essential for the instrument to slide inside the channel (and therefore somehow unavoidable) and influence the relative position and orientation of the instrument exiting the channel and, consequently, the appearance of the instrument.

In a robotic setup, the possibility of exploiting motor encoder data to reinforce the pose estimation is finally explored together with the effectiveness of such proposal in in-vivo environment.

To avoid all the possible errors deriving from an uncertain and potentially incomplete model, a model-free strategy has been evaluated in **Chapter 6**. The main intent is to learn "from experience" the function relating the 3D position of the instrument tip to the image features.

Two common learning methods for regression have been taken into consideration to study the suitability of such solution in pose estimation of flexible instruments and discuss the main issues usually deriving from high input dimensionality and poor or noisy (and therefore untrustworthy) training set.

Several simulation experiments led to the final learning method which is tested on a robotic flexible system in a laboratory environment.

1.4 Labex CAMI Context

This thesis is part of Labex (LABoratoire d'EXcellence, an ANR project) CAMI program "Investissement d'avenir" (french expression for "future investments"). This laboratory is actually a consortium of laboratories that share the common conviction that an integrated approach will strongly increase the quality of medical intervention and the common aspiration of diffusing CAMI (Computer Assisted Medical Intervention) technologies in routine clinical intervention.

CAMI LABEX Network intensively works in 5 areas of interests:

- *Augmented Perception.* Offering the operator the possibility to see beyond the immediately visible by innovative fusion of multi-modal data obtained by novel or classical sensors.

- *Augmented Decision.* Offering assistance to real-time decision-making through high-level planning and monitoring of the intervention.
- *Augmented Action.* Offering the operator a new dimension in intervention performance with miniaturized robots and solutions for augmented dexterity.
- *Augmented Learning.* Reducing the learning curve by offering User-centered learning strategies exploiting the educational potentialities of CAMI technologies.
- *Innovative Methods for medical benefit demonstration.* Developing and validate an adapted methodology for the demonstration of the Medical Benefit of CAMI techniques.

This thesis can be ascribed to the first listed area since the proposed solution can be considered as a "peculiar" sensor that could eventually allows gesture guidance or autonomous positioning decreasing mental and physical burden of physicians with consequent benefit also for the patients.

Publications

Peer Reviewed International Conferences

P. Cabras, D. Goyard, F. Nageotte, P. Zanne, C. Doignon. *Comparison of Methods for Estimating the Position of Actuated Instruments in Flexible Endoscopic Surgery.* IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2014.

Peer Reviewed French National Conferences

Paolo Cabras, David Goyard, Florent Nageotte, Philippe Zanne and Christophe Doignon, *3D Pose Estimation of Actuated Instruments in Flexible Endoscopic Surgery*, Surgetica 2014, December 2014.

Paolo Cabras, David Goyard, Florent Nageotte, Philippe Zanne and Christophe Doignon, *Positionnement 3-D d'un instrument flexible robotisé à l'aide d'une caméra monoculaire endoscopique: comparaison entre méthodes basées modèle et un apprentissage supervisé*, ORASIS, June 2015.

Chapter 2

State of the Art: Medical Flexible Systems and the Problem of their Control

Contents

2.1 Medical Context

Despite the great skepticism reserved to the first laparascopy [Litynski 1998], Minimally Invasive Surgery (MIS) is becoming the new paradigm for surgical operations. The strength behind this revolution has been the great benefits for the patient: small scars, reduced blood loss during the surgery, shorten recovery time and decrease post-operative pain convinced the patients to ask for MIS indirectly promoting the development of these techniques as they are known today. Since then, the surgery seems to focus the attention on the minimization of the invasiveness towards, ideally, a no-scar surgery.

In the last decade, flexible systems have acquired a fundamental role in this field due to their capacities of adapting to human body structures allowing the access of even remote surgical sites in the interior of the body through natural orifices and tracts.

In this direction, NOTES is considered as maybe the most revolutionary concept of surgery between the MIS techniques but it does not find yet a large acceptance among the physicians. In fact, despite the enthusiasm after the first pioneering works on animal [Kalloo 2004] and humans [Reddy 2004, Zorrón 2007, Marescaux 2007], it is still object of a controversial discussion. The transluminal approach appears to have tremendous potential, but several important issues, including the safety of this approach and whether it will provide significant patient benefit in terms of postoperative recovery compared with other procedures, must be resolved before the new technique is widely introduced into clinical use. As Podolsky stated: "Patients' desire for the cosmetic benefit is important, but the price cannot be increased risk or complication." [Podolsky 2010].

Dhumane et al. suspect that the benefits, at this point in its development, may be limited and the appropriate technology should be waited. They believe that the ultimate form of Minimally Invasive Surgery will be a hybrid form of Minimally Invasive Single Site surgery and Natural Orifice Transluminal Endoscopic Surgery, complimented by technological innovations from the fields of robotics and computer-assisted surgery [Dhumane 2011]. The necessity of an improved instrumentation is also underlined in the white paper on NOTES produced in 2006 by the working group of surgeons and gastroenterologists (known as Natural Orifice Surgery Consortium for Assessment and Research (NOSCAR)) promoted by the American Society for Gastrointestinal Endoscopy (ASGE) and the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES). In these documents several guidelines are proposed for research to make NOTES a viable clinical application for patient. These areas included development of a reliable closure technique for the internal incision, prevention of infection in addition to the creation of advanced endoscopic surgical tools [Rattner 2006, Swanström 2008]. This same necessity gave stimulus to the creation of advanced flexible systems to use for single port or NO-SCAR surgery which, in turn, can be either NOTES or endoluminal surgery.

2.1.1 Flexible Systems for NO - SCAR Surgery

The first systems were just a composition of already existing endoscopic and surgical instruments: this composite repurposed equipment were rudimentary for these applications and usually implied the participation of several surgeons in the operation and totally uncomfortable positions (Fig. 2.1).

Thus, since the first successful cases of NOTES the interest of having more specific and ergonomic instruments increased and led to redesign the endoscopic access device itself [Swanström 2008]. Most prototypes were modified based on the therapeutic endoscopes by augmenting the number and the diameter of the channels.

One of the first platforms to be used in NOTES on animal models was the "R-scope" by Olympus later called NOTES scope. It consists in a dual channel endoscope (DCE) with two bending segments. The primary segment is lockable

Figure 2.1: First flexible system for surgical intervention were a composition of existing tools which implied the participation of several surgeons during the operation (photo: IRCAD France).

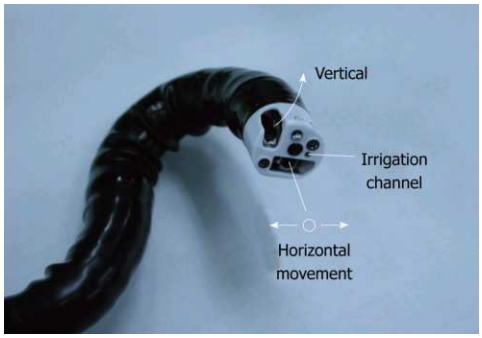


Figure 2.2: Tip of the R-Scope by Olympus. The two working channel has lifting gates that are orthogonally position wrt to each other allowing simultaneous lifting and dissection movements. The head also present two bending sections where the primary one is lockable (photo: http://www.wjgnet.com).

allowing a better angle of approach and more precise tissue manipulation with the maneuvering of the second segment. The peculiarity of this endoscope were the two working channels which has lifting gates orthogonally position with respect one another allowing simultaneous lifting (vertical) and dissection (horizontal) movements (Fig. 2.2). This endoscope improves the characteristic of a conventional DCE providing a better positioning and a small degree of triangulation of the instruments but, its characteristics were not optimal for NOTES such as the limited field of view or the complexity of maneuvering the two independent sections.

From a surgical point of view, though, these two movements are not sufficient since more dexterity is needed and triangulation must be assured to carry out a complete surgical intervention. With the term *triangulation* is meant the fact that the surgical tools should form a triangle pointing towards the organ so as to permit traction on tissues to facilitate dissection along normal anatomical planes. If, on the other hand, the instruments are inserted close to each other and almost parallel,

the applicable forces in the horizontal direction are weak and the collisions between the two instruments are almost unavoidable complicating the manipulation of the organs.

Another important aspect in MIS is the position of the camera wrt the surgical plane[1] The ideal solution would be to maintain the camera outside the surgical plane so as to allow visualization of the surgical plane and prevents the camera to physically interfering with the surgical plane itself.

A more surgery-aimed example by Olympus is the EndoSAMURAI designed to operate within the flexible-laparoscopic paradigm. This specialized endoscope is manipulated by a remote working station and is provided with a locking over-tube and two independent arms in the distal part of the scope. These arms have 5 degrees of freedom (DOF) and serve as conduit for different end-effectors. In addition to that, there is a third channel that can be used for auxiliary equipment or for suction/irrigation. The triangulation assured by these arms makes them suitable for tying sutures in addition to providing traction (Fig. 2.3). The locking overtube confers to this device more stability than the DCE but always presents the same image-perspective limitations since the arms are married to the camera. The methodology employed by the EndoSAMURAI is a "drive, park and move" methodology. Once the target is achieved, the user locks the scope position with the over-tube and proceed to the user-interface: the assistant should solely adjust the image (as in conventional laparoscopic) without worrying of the configuration of the endoscope. The idea beneath this was to avoid the necessity of great physicians coordination which often result difficult.

An alternative to endoSAMURAI is the flexible-laparoscopic platform from Boston Scientific called direct drive endoscopic system (DDES). The guide sheath houses 3 working channels and the user interface is an ergonomic rail-guided drive handles. By these handles the physician can move all the 7 dof of the system. The instruments can be inserted in the lumens and the depth of insertion is independent of the image. The scope is inserted in one of the channels and it can be freely rotated independently of the instruments and DDES, adjusting the horizon. Contrarily to the endoSAMURAI, it does not present a dedicated channel for irrigation/suction (which may not be adequate for intra-abdominal surgery) and the maintenance of the visualisation may require endoscope and DDES adjustments during manipulation of tissues (see Fig. 2.4).

A similar philosophy is the one of the IOP (Incision Operating platform) system by USGI Medical (Fig. 2.5). It is a flexible device with 4 lumens which are used as guides for the scope (N-scope by Olympus), specialized instruments or for high-flow carbon dioxide insufflation. If on one hand the triangulation is improved with respect to the DCE, on the other hand the in-line channel orientation still favors

[1]Plane formed by the access point of the instruments (trocars position in case of Laparoscopy or working channels end in case of flexible systems) and - ideally - the point of the organ that has to be operated.

(a) Tip of EndoSAMURAI (b) Master Console

Figure 2.3: EndoSAMURAI by Olympus endoscopic head with instruments and master consoles (right). A locking overtube make it more stable than DCE but always present the same image-perspective limitations since the camera is married with the instrument (photo: http://www.wjgnet.com).

(a) (b)

Figure 2.4: Boston Scientific's Direct Drive Endoscopic System. The sheath houses three channel where two instrument and a scope can be inserted. The two handles let the user control all the 7 DOFs of the system. No irrigation/suction channel is present but instrument movements are independent from the camera. The scope is not integrated in the system and, then, it can rotate and translate independently from DDES (photo: http://www.wjgnet.com).

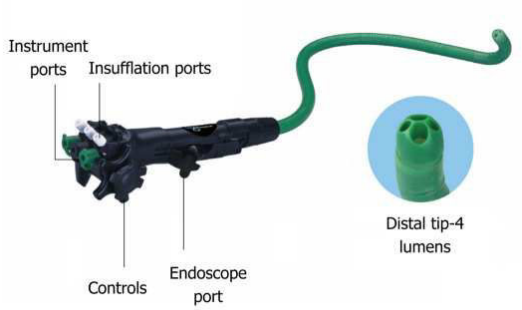


Figure 2.5: Incision Operating Platform by USGI Medical. It presents 4 lumens that can be used to house a scope in addition to specialized instruments or for carbon dioxide insufflation. The in-line channel orientation favors parallelism (photo: http://www.wjgnet.com).

parallelism. Work load for the IOP is high and requires skilled assistants as the primary operator shares responsibilities for instrument exchange, device orientation and scope positioning.

In parallel to the NOSCAR, the New European Surgical Academy (NESA) established in Berlin in 2004 the first European working group over NOS (Natural Orifice Surgery) which includes NOTES but also all the surgical procedures performed through all the body natural openings. After the first meeting they decided to focus their attention to vaginal/douglas route, which they considered more promising for its safety and cost-effectiveness. They claimed that this approach presented several advantages with respect to transgastrical way. Among them: a simple entry and straight view for upper abdominal procedures, the flexibility of the vaginal walls which allow inserting wide instruments without causing mechanical damage, an adequate working space (the diameter of the pouch of Douglas is suited to operative intervention) and an increased surgeon comfort which can perform the procedure seated [Ramesh 2013]. The NESA drew attention on the lack of adequate instrumentation in this context and develop a new device called TED (Transdouglas Endoscopic Device) conceived to adapt to the pelvic anatomy (Fig.2.6). Even though it is not a flexible instrument, it is interesting to mention it here since its application is no-scar surgery and it represents an alternative to purely flexible endoscopic and instruments.

TED is a multichannel articulated rigid endoscope which can assume two different operational configurations: S and U shaped. The S-shape was designed to reach the upper abdomen (for cholecystectomy, splenectomy, ...) bending first anteriorly and then posteriorly and the U-shape variant for lower abdomen operations (appendicectomy, hysterectomy, ...). The controllable instruments (pincer and scissors) are hidden in the head of the device and, once the TED is positioned inside the body, the arms with the instruments are deployed. They can be pivoted and partly

Manual Flexible Platforms

# of working channels	DOFs	Pros	Cons
R-Scope			
2	1 x 2 end effectors (EE)	• 2 orthogonal Lifting gates in the working channels → vertical and horizontal movements allowed • 2 bending segments - 1 lockable	• Instruments arms married to the camera
EndoSAMURAI			
3	5 x 2 EE	• Remote working station • Locking/Stiffening overtube • Suction/Irrigation channel	• No control on insertion depth
DDES			
3	7x2 EE	• Ergonomic Interface • Instrument Depth of insertion independent from the image	• Not self supporting
IOP			
4	depending from used instruments	• Effector can enter the operative field • Built in shaft - stiffening system	• in-line channels orientation • no irrigation/suction dedicated channel
TED			
4	2 for the body 3x2 instr.	• articulated rigid links → high forces are applicable • Good triangulation and positioning (for the specific operation)	• specific for one application (not flexible)
Anubis			
3	3 x 2 instruments 2 for endoscopic head	• Last part of instrument channels deviated → triangulation assured • suction/irrigation channel	• Not self supporting

Table 2.1: The main commercial manual flexible platforms for medicine.

(a) (b)

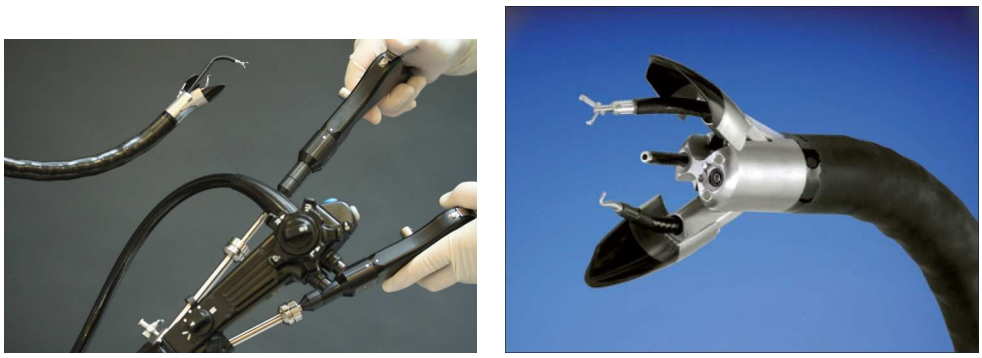
Figure 2.6: Transdouglas Endoscopic Device in the S-shape operational configuration. It is 3 rigid links device with 4 working channels. Once positioned, the arms with the instruments are deployed as shown in (b). Moreover, the head of TED houses also a camera, a light and a dedicated channel for flushing and sucking (image: http://www.nesacademy.org).

rotated and also moved forwards and backwards. This enables tissue manipulation with traction and counter-traction in all planes. Besides the instruments, the TED head houses a camera, a light and a channel dedicated to flushing and sucking. A further channel is available as working channel. In 2009 preclinical studies should have started [Stark 2008] but no further publication are available yet.

In 2009 from the synergy of Karl Storz and IRCAD - France [2] born the Anubis Scope (Fig. 2.7): a flexible endoscope specifically conceived for NOTES. It is larger than the conventional endoscope and it houses different channels: in addition to the dedicated channel for irrigation and suction, it has two main channels on each side of the camera and a third channel beneath it. The two main channels are used to insert two flexible instruments (usually a grasper and an electric tool) which can smoothly slide and rotate. These instruments are endowed with an additional degree of freedom: the distal part can be bent on a plane ($+/-120^o$) by a lever integrated on the handle of the instruments (cf. Fig. 2.7 (a)). Finally, the triangulation is assured by the fact that the final section of the main channels is slightly deviated (10^o) with respect to the camera axis. More precisely, the tip of the endoscope is composed by two jaws (Fig. 2.7) that, once closed, confer to it an arrow-like shape, which eases the insertion of the device through natural orifices and wall incisions.

However, as already pointed out in the introduction, these manual systems often lack of maneuverability and ergonomics, not to mention the difficulties relating to staff coordination when carrying out an operation. Robotics offered and is still offering several solutions to these needs easing the task of the surgeon and, consequently, decreasing physicians' physical and mental burden and increasing their and their patients' safety. Some of these examples are described in the next sections focusing on flexible robotic system after presenting a brief evolution review of medical robotics.

[2] Institut de Recherche contre les Cancers de l'Appareil Digestif [Institute of Digestive Cancer Research]



(a) (b)

Figure 2.7: Anubis Scope by Karl Storz. Wider than the conventional endoscope, its body houses two main channels (one at each side of the camera) where surgical instrument can be inserted and an additional channel beneath the camera. The triangulation is assured by the slight deviation of the main channels wrt the camera axis. Thanks to the knobs on the main shaft handle shown in (a), the head of the endoscope can be oriented. In (a) also the instrument handles are shown with the integrated lever that allows to bend the tip of the instruments.

2.1.2 Medical Robotics for MIS

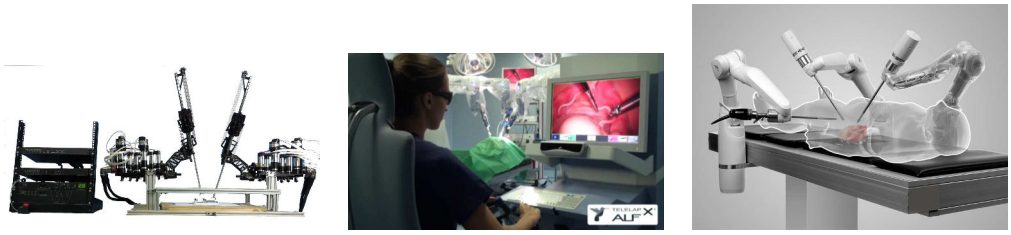
Differently from laparoscopy, which can be defined an already mature field, surgical operation using flexible instruments and devices is still an open field which is at the center of many researches. That is the reason why the most numerous examples of commercial medical robots are in the field of laparoscopy or medical operation with rigid instruments.

Nowadays, the most known medical robot for laparoscopic surgery is probably the DaVinci platform, but it represents the last step of a long pathway, which is worth mentioning here and which is still in vivid evolution presenting new less invasive solutions for surgery.

2.1.2.1 Medical Robotics Evolution

It was 1985 when the first robot was introduced in an operative room to perform biopsies [Kwoh 1988]. But in that case it was an industrial robot (PUMA 560) employed for medical purposes.

The first robot designed for medical application was ROBODOC which was actually capable of autonomous movements, assisting in Total Hip Arthroplasty (THA). In the same period (1994), Computer Motion got the FDA clearance for AESOP (Automated Endoscopic System for Optimal Positioning): a robotic arm holding an endoscopic camera which could be voice controlled by the surgeon. This robot represented the first try to solve the problem of team interaction and coordination even though voice control may be not always easy and practical within a surgery



(a) (b) (c)

Figure 2.8: Example of robots for performing laparoscopic surgery. The open-architecture robot Raven II in (a) (photo:www.washington.edu) and the two European robots: ALF-X (photo:alf-x.com) by NESA and SOFAR S.p.A in (b) and MIRO by DLR - Germany in (c) (photo:tilo-wuesthoff.de).

and operating room (OR) context.

An important breakthrough arrived with the first teleoperated surgical platforms for laparoscopy: Zeus by Computer Motion (finally incorporated in Intuitive Surgical) and DaVinci by Intuitive Surgical. The first was composed of a surgeon control console and of a patient cart where three robotic arms were mounted. The right and left 7-dof robotic arms replicate the movements of the right and left hands of the user while the third arm is an AESOP voice-controlled endoscope holder used for visualization. The surgeon is seated upright in front of a video screen and the instrument handles are adjusted to maximize comfort and dexterity. The idea underneath DaVinci is similar to Zeus: the system is composed by three robotic arms (one for the dual camera and two for the instruments) which can be controlled remotely at a console. One of the peculiarities is that the image is a 3D image which is displayed above the hands of the surgeon so that it gives the surgeon the illusion that the tips of the instruments are an extension of the control grips, thus giving the impression of being at the surgical site.

The Master-Slave paradigm seems to be one of the most diffuse paradigm for medical robotics for surgery: the robot is just a more precise and comfortable tool that is always controlled by the surgeon.

Since then, new robots for general laparoscopy were presented and conceived. Two further examples worth mentioning can be the Raven II and the European MIRO and ALF-X. The Raven is an open-architecture surgical robot for laparoscopic surgery research from Applied Dexterity. It has two cable-driven 7 DOF arms (6 DOF + grasp) and it is intended to facilitate collaborative research on advances in surgical robot [Kehoe 2014]. An upgrade of the Raven II is the Raven IV which is composed of 4 arms and two cameras allowing the collaboration of two surgeons interacting with the surgical site in teleoperation.

Together with the more known MIRO (DLR, Germany), the TELEALAP ALF-X is one of the newest european proposal for robotic laparoscopy [Gidaro 2012]. ALF-X was born from the collaboration of NESA with the Joint Research Center of the

(a) (b) (c)

Figure 2.9: Example of robots for single port surgery. The Sp system by DaVinci in (a) (photo: medgadget) and SPORT by Titan Medical in (b) (photo: twitter.com/simnovate) which derives from IREP project by Columbia University (c) (photo: linazeldovich.com).

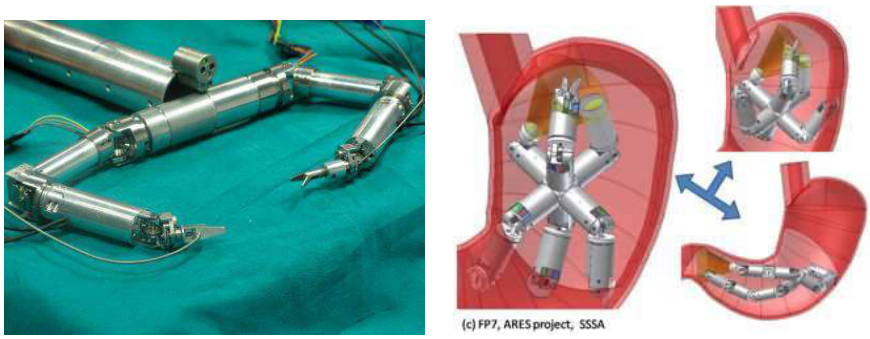
European Commission and SOFAR S.p.A. (Milan, Italy). The great peculiarities of this system are the tactile sensing and the eye-tracking system which not only allow to control the camera view just by looking in the wanted direction but also to activate the instruments by merely looking at them. Further advantages are the universality, wide range of application for every-day usage and cost-effectiveness. Preliminary good clinical results have already been achieved [Fanfani 2015, Gidaro 2014].

In the single port approach, two products are about to be launched on the market: Da Vinci's Sp system and Titan Medical Inc's SPORT. The former appears as a rigid tube where three instruments and a stereo camera are allocated. Once in the body the instruments exit the tube and thanks to their distal bendable sections can assure triangulation during the operation. The stereo camera can be controlled as well, providing a 3D image from an overhead point of view. It is meant to work with the Da Vinci Console and should be launched in late 2015.

The latter derives from IREP (Insertable Robotic Effector Platform) project by Columbia University who signed an exclusive agreement with Titan Medical. It is a rigid tubular structure that, once inserted in the body through a single incision can be deployed. In the unfold configuration three parts can be distinguished: two dexterous flexible arms and an upper view stereo camera. Each flexible instrument is composed of a two-segment continuum robot, a parallelogram mechanism for improved rigidity and a distal wrist for improved dexterity during suturing.

Other examples for single port surgery are ARAKNES SPRINT which is a dual-arm robot with six degrees of freedom per arm plus one tool or gripper [Piccigallo 2010] and, very similar to this, the Virtual Incision's new robot for general abdominal surgery.

A totally different paradigm in this field, is bringing the entire robot where the operation is needed. In this direction, the project ARES had the objective to develop modular miniature robots being able to re-assemble once got to the desired position (the main application was the gastrointestinal surgery). The idea is that the patient



(a) (b)

Figure 2.10: In (a) another example of robot for single port surgery (ARAKNES SPRINT, photo: sssa.bioroboticsinstitute.it). In (b), the concept of ARES project is shown: a miniature-modular robot able to reassemble once got to the desired site inside the human body (photo: Scuola Superiore Sant'Anna).

could swallow different pills (i.e. the different components of the robot) which will self-assemble thanks to magnetic anchoring. Each component has its particular role (image control, communication, diagnostics and others) inserted in a collaborative framework to carry out a specific task. As one can imagine seen the complexity of the concept, this clever idea has several implementation issues and it is still under investigation.

2.1.2.2 Flexible Robots in Medicine

Differently from robotic laparoscopy with rigid instruments, the market does not offer many examples of flexible robotics for surgery because of the great challenges that robotizing miniaturized flexible instruments entails. On the other hand, the research in this field is fervid indicating the great interests of both roboticians and physicians on this subject. Many different architectures have been proposed in the last years for medical applications and more specifically for NOTES or endoluminal surgery [Traeger 2014, Seneci 2014, Simaan 2004]. In the next paragraphs we will describe some representative examples to show the philosophy beneath 3 different categories of conception.

The *first category* could be defined as those robots being born by the assemblage and automatization of conventional endoscopes with the necessary instruments.

A well known system in this category is the MASTER system entirely developed in NTU (Nanyang Technological University) by the group leaded by prof. Phee (Fig. 2.11). For sake of intuitiveness in using the robot, they decided that the instruments should mimic the arm movements. Therefore, they conceived the attached instruments with 5 dof (plus one additional for manipulating the end effector) and tendon-sheath actuation is used to assure flexibility. The anthropomorphic data of

Robotic Flexible Instruments: Commercial Examples

Application	DOFs	Characteristics
SeseiX/Magellan		
Electrophysiology / Intravascular	1	• Master installed far from patient → less radiations exposure • 2 guides: 1 stiffer to get to the site of interest, 1 with larger bending range for the operation • Force sensor → tip of catheter always in contact with tissue
Amigo System		
Electrophysiology / Intravascular	1	• Master installed far from patient → less radiations exposure • Commercial Catheter can be docked in the system (no need of proprietary catheter)
NES		
Colonoscopy	16	• Sensor in the tip. • Physician controls the tip trajectory, the system measures tip position and instructs the 16 articulations to follow the leader.
FLEX		
General Endoscopic Surgery	30	• Self Supporting ← location and point of view can be chosen properly. • 2 side-channels for MANUAL instruments.

Table 2.2: Principal commercial flexible robotic platform for medicine.

Robotic Flexible Instruments: Research Experimental Platform		
MASTER		
Application	DOFs	Characteristics
General Endoscopic Surgery	5 (+1) x 2 instr.	• Master console controlled by surgeon arms. • An endoscopist must hold the camera.
STRAS		
General Endoscopic Surgery	10 3x2 instr.	• Robotization of Anubis. • Not self supporting. • All DOFs can be telemanipulated. • Instrument insertion independent from camera.
Heart Lander		
Heart Operations		• No lumen constraint needed. • Adheres to organ walls (inchworm-like locomotion). • No scope → external imagery system needed for guidance.
Bio-Inspired		
Abdominal Single Port Surgery	-	• Still of big dimensions (wrt STRAS for example). • Possibility to modify stiffness → alternatively compliant and high forces application.

Table 2.3: Principal flexible robotic platforms for medical application (Research).

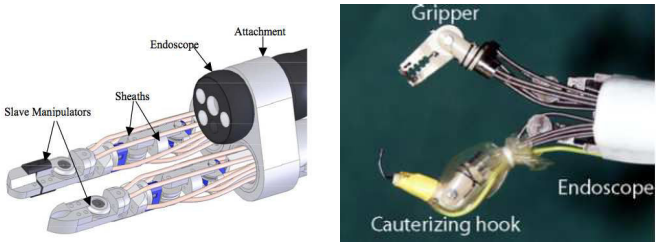
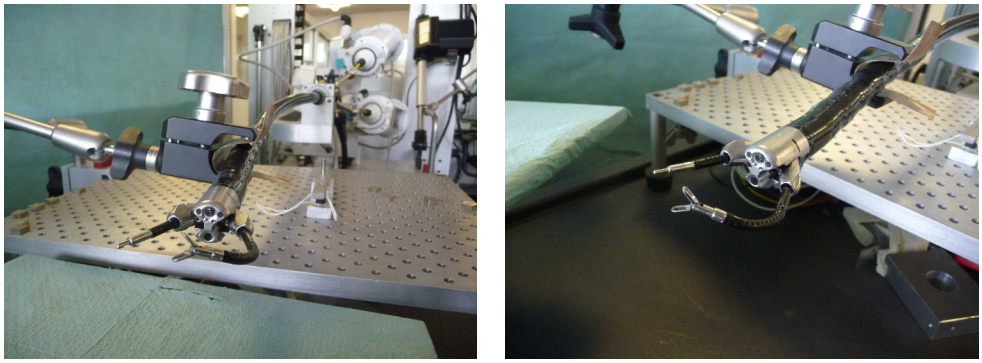


Figure 2.11: Scheme and photo of the tip of MASTER system. A gripper and a hook are located beneath the endoscopic camera and telemanipulated (photo from [Phee 2009]).



(a) (b)

Figure 2.12: STRAS system global view (a) and tip (b). It is the robotic version of the Anubis scope where all its DOFs can be tele-manipulated.

the surgeon's arms are measured by the master console and transmitted to the robot as inputs.

Another example in this category is the STRAS (Single access and Transluminal Robotic Assistant for Surgeons) robot (Fig. 2.12). Also conceived for a tele manipulation framework, the slave robot resulted from the robotisation of the Anubis platform by Karl Storz (which has been presented previously). The master interface is composed by two handles whose degrees of freedom are similar to those of the flexible instruments so as to assure an easy mapping between the two. The shape of this handle is conceived to be familiar to the surgeon and it is an adapted version of the flexible instruments of the Anubis platform. An additional joystick is added on the tip of the handle for controlling the endoscope movement during operation without the need of switching pedals or any assistant [De Donno 2013] (cf. Part I for more details).

University of Twente has also developed his solution consisting in 3 add-on steering modules: a tip steering module, a shaft manipulator and an instrument module. Together they allow to control the degrees of freedom (required to position and

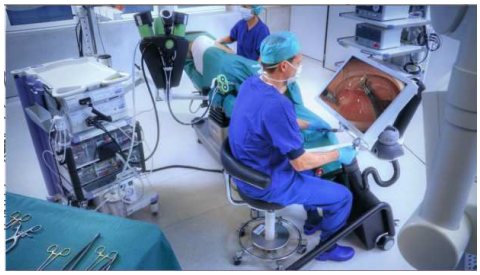
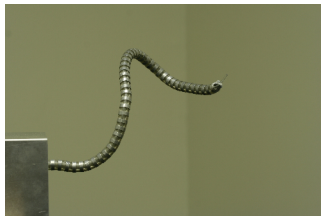


Figure 2.13: TeleFLEX system by the University of Twente. The (green) steering modules are attached to the conventional flexible system (e.g. Anubis) (photo: demcon.nl).

(a) FLEX system

(b) FLEX system



(c) HARP system

Figure 2.14: The FLEX system in (a) and (b) is composed by an endoscope and two working channels where specific instruments can be inserted (photos: medrobotics.com). The scope is self supporting and take inspiration from the HARP system in (c) (photo:carnegie mellon.)

stabilize the endoscopic tip and the instruments) with a master console (Fig.2.13).

The unique example of commercial flexible robot for surgery is *Flex* by Medrobotics (Raynham, Massachusetts), which received the CE mark in 2014 and started to be sold in Europe and only in 2015 the FDA clearance. This robotic endoscope (Fig. 2.14(a)) can be manipulated with a joystick and, thanks to its more than 30 dof, enhances the physician ability to reach hard to access areas of the inner body. Moreover, it is self supporting and can generate a unique shape in 3D following the user input. The necessity of working inside a natural lumen constraint is avoided and a better point of view or location can be chosen properly. The two side channels can house two flexible instruments (usually to grasp or cut) manipulated by the physician. Conversely, FLEX can also become limp to adapt to the surrounding anatomical structures. Originally conceived for heart surgery, a recent study showed its utility also in transoral surgery [Johnson 2013].

The original idea behind FLEX is HARP, which was developed in Carnegie Mellon by Choset and Degani [Degani 2006]. HARP (highly articulated robotic probe) was conceived to fuse in the same device the advantages of both rigid and

(a) Global view (b) Follow-the-leader endoscope

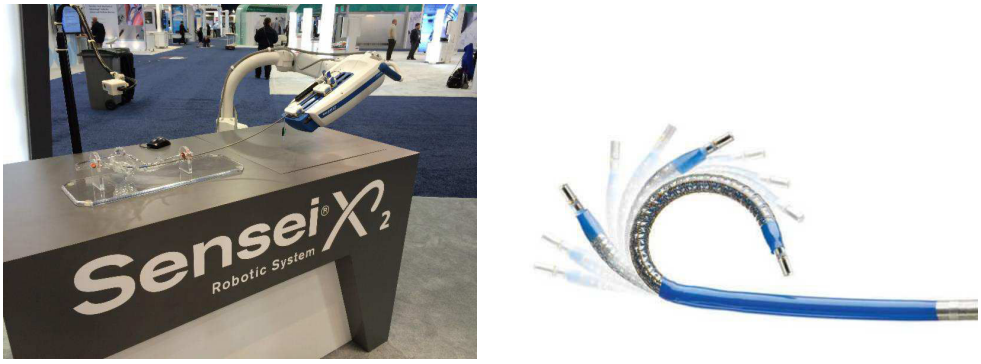
Figure 2.15: NES system used in colonoscopy (a) (photo: intechopen). A follow-the-leader strategy make the endoscope follow the path described by the tip (b) which is controlled by the physician (photo: nature.com).

flexible instruments (Fig. 2.14(c)). The probe was actuated by four cables and the distal link was relayed to a camera by a fiber optic endoscope attached to the outer mechanism. The HARP consisted in two concentric tubes that can alternate between rigid and limp states. The self sustainability is assured by alternating the motion and state of the inner and outer mechanisms. The same robot is at the base of the CardioARM developed by the startup Cardiorobotics located in Pittsburgh. It has 102 dof and can follow whatever 3D curve in the space (A Novel Highly Articulated Robotic Surgical System for Cardiac Ablation).

Enlarging the boundaries of this group including colonoscopy, NeoGuide presented the Navigator Endoscopy System (NES) which allows to avoid the looping problem which occurs when the forces of the scope stretches the colon (Fig. 2.15). Looping is associated with pain and need for sedation and consequent procedural difficulty and procedural time. The physician inserts the scope and controls the tip of the scope, the sensor attached to the tip measure its position and create a 3D map of the colon (actually the map of the path it made) and, based on this information, the system instructs the 16 articulating segments composing the scope to follow the path.

A *second category* is the one of robots for cardiovascular intervention. No endoscope is employed here since smaller structures are needed in order to walk along veins or through membranes. The guidance of these tools is usually performed using an external imagery system (X-ray or others).

The most popular in this category may be Sensei X for electrophysiology operations and the Magellan (both by Hansen Medical, Inc) for intravascular operations. A bedside robotic arm is holding the catheter which is teleoperated by the physi-



(a) (b)

Figure 2.16: Sensei X system and associated catheter. The force sensor integrated in the tip allows to maintain the tip in contact with the tissue increasing the efficacy of the therapy (photo: Hansen Medical).

cians acting on the master console. Sensei X catheter (Fig. 2.16) is composed by two guides with different stiffness and magnitude of deflection. In beating heart operations, the outer stiffer guide is used till reaching the atrium of heart when the inner guide is drawn to allow more articulated movements. Thanks to the accurate force sensor embedded on the tip, Sensei X is capable of keeping the tip in contact with the tissue or applying a desired pressure autonomously. The stability of the catheter and even the pressure it exercises on heart tissues are of great utility, above all in complex ablation procedures such as atrial fibrillation [Di Biase 2009]. The master console can be installed far from the slave cart preventing the physicians of being repeatedly exposed to noxious radiations. The basic structure is the same for Magellan for which just the catheter changes which are more adequate for intravascular procedures.

Another interesting proposal comes from Catheter Robotics, Inc. with its Amigo System (Fig. 2.17). It was designed to be easy to use and cost effective. Contrarily to Sensei X, it does not require proprietary catheter: the slave component is actually a simple mechanical interface to selected commercially available catheters whose handle can be placed in the robot docking station. A particular handle replicating the one of the conventional catheter allows the remote control of the catheter as if it were manipulated manually. Thanks to the remote control (up to 30.5 meters) the radiation exposure of the physician is reduced. Unfortunately, no force sensor is embedded in the tip.

In addition to the mentioned commercial examples of teleoperated catheters, one robot for heart surgery worth mentioning is the Heart Lander born by the collaboration of Carnegie Mellon and Pittsburgh Medical Center. Much bigger than a catheter, it can be inserted through a small incision below the xiphoid process of the sternum, avoiding any intervention to the lung. It adheres to the epicardium

(a) (b)

Figure 2.17: The Amigo System is a support where commercial catheter can be positioned. The particular handle in (b) is employed to control the movements of the catheter (photo: catheter robotics).

using suction, and navigates via inchworm-like locomotion, using flexible pushwires connected to motors located outside the patient.

The *last category*, which brings the flexible/compliant robot to its extreme declination, can be composed by bio-inspired robots whose shape and stiffness can be controlled. In the original idea of [Stilli 2014, Maghooa 2015], the body of the robot is a conic-shape air chamber where the TCP (Tool Center Point) is represented by the cone vertex. Three pairs of nylon tendons are attached on the surface of the cone and with proper actuation on them the robot can assume different shapes (see Fig. 2.18).

Another renown example is STIFF FLOP, a soft robotic arm that can squeeze through a standard 12mm diameter Trocar-port, reconfigure itself and stiffen by hydrostatic actuation to perform compliant force control tasks (Fig. 2.19). The project objective is to design a soft manipulator with a gripper at the tip, distributed sensing, biologically inspired actuation and control architectures, learning and developing cognition through interaction with a human instructor, and manipulating soft objects in complex and uncertain environments.

The idea within the surgical framework is to insert the uninflated robot inside the body and modify its stiffness regulating the internal air pressure or the shape actuating on cables. In fact, the flexible instrument present an intrinsic "contradiction": they need to be soft enough to be compliant with anatomical structure but hard enough for effective force transmission and weight resistance [Loeve 2010]. In the studied context, then, the possibility of changing this quality can be very

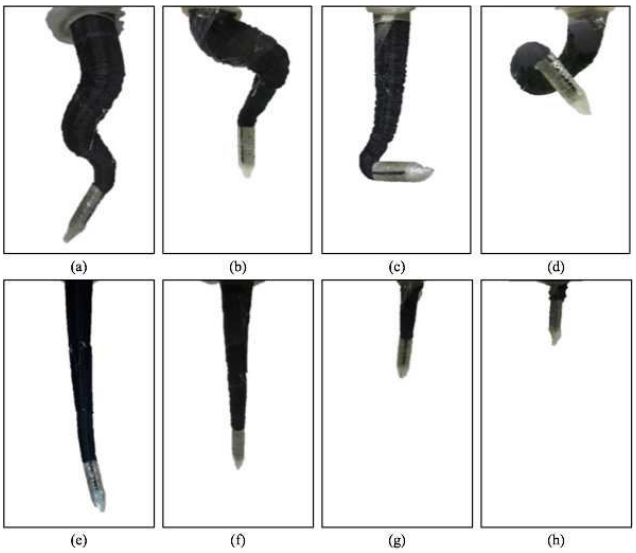


Figure 2.18: Examples of possible configuration of the bio-inspired manipulator: actuation of the middle and tip sections (a), the base and middle sections (b), the tip section (c), all sections (d), and elongation/shrinkage capability (e)-(h) (photo from [**?**]).

Figure 2.19: STIFF FLOP prototype. The robot can extend and compress and cable actuation allows to determine the shape of the two flexible sections (photo: National Museums Scotland - https://vimeo.com/97913172).

valuable .

All the cited robots differ from the conventional rigid link robots in several aspects concerning actuation, dynamic and shape variability. They are usually classified in a specific family called *continuum robots*. If the final aim is trying to assist the surgeon improving his/her skills by the means of the robot, the characteristics and behavior of continuum robots must be known, so that automatic control actions can be applied to these systems. In the next sections, continuum robots are presented (Sec. 2.2) and state-of-the-art solutions for the control are described, focusing on the surgical environment (Sec. 2.3).

2.2 Continuum Robots

From the classification of Robinson and Davies [Robinson 1999], a continuum robot does not contain rigid links or rotational joints and their structures can bend continuously along their length via elastic deformation. Continuum robots must not be confused with Serpentine Robots which combine very short rigid links with a large density of joints. These robots are not very common and the only commercial example are the snake-like arms by OC Robotics [oc 2015]. In this manuscript, though, serpentine robots will be considered as continuum robots in the case they are under-actuated or, more specifically, when the joints and links are passive and function only as a support structure (the backbone) which is deformed by external actuation.

Within continuum robots, a broad classification is proposed according to where the actuators are placed on the robot. They are named intrinsic if the actuators are on the structure of the robot and form part of the body of the mechanism, extrinsic if all the actuators are separated from the body and the movements is transmitted via a mechanical linkage and hybrid if a combination of the two strategies is used.

As it can be deduced also from the previous examples, extrinsic actuation seems to be more suitable for surgery since it allows obtaining smaller and more lightweight structures which are fundamental characteristics in this field.
In this class, two families can be distinguished according to how the robot shape is obtained. The first is formed by the concentric pre-bent tubes where the shape is obtained as result of the interaction of the different component of the robot. The other one includes those robots usually composed by a passive backbone whose shape is obtained by an external actuation e.g. using forces (which can be transferred using either cables or muscular hydrostats) or Shape Memory Alloy (SMA) actuators.

2.2.1 External Actuation

Tendons are maybe the most direct way of external actuation. Group of tendons are usually routed along the backbone and terminate at different levels along it:

all the tendons of a same group terminate at the same level dividing the structure into so-called *bending sections*. The forces applied at the base of the tendons are translated in torques at the termination points, hence, producing bending. The final shape of the backbone is then resulting by the different shape of the different sections. Tendon driven systems are usually conceived as extrinsic continuum robots and the actuators driving the cables are usually located at the base of the robot. One possibility of realization is using spring backbone [Mehling 2006, Zhao 2010] but the intrinsic natural compliance increases the difficulty in controlling them: control efforts intended for bending are lost in compression. A solution is using incompressible backbones using rigid vertebrae [Bardou 2010, Shang 2012], but this means loosing extension capability. In this design, the use of tendons allow high forces transmission but it implies finding some method to prevent backslash and slack.

Alternatively to cable driven instruments, SMA (Shape Memory Allow) can be used as external actuation for a continuum robot. In [Szewczyk 2001], two SMA springs, mounted in an antagonist configuration, are used to change the relative orientation of two consecutive backbone segments. The idea under SMA is that SMA actuators undergo a micro-structural transformation from their austenite phase to their martensite phase. Heating and cooling the material or applying an external stress, this transformation is activated and the actuator changes its shape and stiffness. However, they may not provide a large enough deflection range and their performances seem to be not sufficient yet when carrying weights or when submitted to strong traction.

A third actuation modality is using muscular hydrostats. In the Bionic Handling Assistant by FESTO [festo 2015], for example, the 3D printed structure is used as air chamber that can longitudinally extend or compress. In the OctArm by the Penn State research team [McMahan 2006], instead, "air muscle" actuators are used which consist of particular covering latex tube with a plastic mesh sheath conferring to the muscle a large strength. By combining three or more of these actuators and combining the pressure in each of the chambers extension/compression or bending of each section of the robot is achieved. A similar concept was used by Rosa et al. [Rosa 2011] to perform confocal endomicroscopy. Three balloons catheter, positioned with 120 deg between them, were used to orient the endomicroscopic laser probe in the space.

2.2.2 Concentric Tube Design

An alternative design is to use a backbone composed by concentric tubes that are free to rotate and translate one with respect to the other. The structure can extend/contract by longitudinal translation and rotate by rotational sliding of the tubes. Backbone shape modification is usually achieved by using pre-curved compliant tubes in combination with directly controllable relative translation and rota-

tion. This kind of actuation allows having thin design but, contrarily to the tendon driven system, the bending cannot be actively controlled. They encountered a high diffusion in medical field for their smaller profile and great compliance, even though they provide less forces with respect to tendon-driven mechanism. In this context they are also known as active cannulas and have been used to carry out various MIS operations permitting task oriented designs [Torres 2012] or MRI image-guided operation [Su 2012].

2.3 Control of Continuum Robots

The design of automatic controllers is a fundamental issue for any robotic system and become more complicated for underactuated systems such as the described continuum robots. Additional difficulties derive from the usual absence of adequate sensors along the structure, the fact that actuators do not coincide with space variables and, above all, the complexity of their dynamic models.

Slack and backlash are two main phenomena which complicate the achievement of a precise model of tendon driven systems. If the cables are not properly pretensioned, there is a delay between the start of the motor movement and the actual deformation of the body of the robot. The initial movement of the motor is used to tension the slack cable in order to transfer the motion to the robot section through the cable. Same phenomena can be observed when a change in the direction of the motor occurs. For example, if antagonistic cable are employed for achieving bending in a plane, when the direction of rotation is changed the cable that was passive is slack and the initial movement of the motor is dedicated to recover from this loss of tension.
Friction between cables and backbones vertebrae and between cables and the sheath is the other main factor of complexity. In presence of friction, the relationship between traction forces and the shape of the robot body becomes more and slack becomes more severe [Agrawal 2010a].

Active cannulas do not have totally predictable behavior neither and, also suffer of friction between the tube surfaces. This can lead to torsion and traction of the tubes making a precise modeling of the motor to end effector relationship difficult.

To overcome these problems, several control strategies have been adapted for continuum robots throughout the years. The employed strategies can be classified mainly in three categories: model-based open loop approaches (1), closed-loop control strategies where the position/shape measurement is performed using attached external sensors (2) or using vision-based methods (3). Camera is still an external sensor but, given the specificity of the information processing they require, it has been chosen to separate them from category (2).

2.3.1 Kinematic and Dynamic Model-Based Methods

Despite the difficulties in obtaining a precise model, many studies have been made to transfer kinematic and dynamic control from rigid link to continuum robots, investing strong efforts to obtain general kinematic [Chirikjian 1994, Jones 2006, Gravagne 2000] and dynamic [Chirikjian 1993, Mochiyama 2003] models.

In [Ivanescu 1995], Ivanescu et al. proposed a variable structure controller for a tentacle manipulator, whose structure was made of flexible materials in conjunction with active controllable electro-rheological fluids. A non-linear observer, based on a simplified discrete spatial model, was adopted to know the "shape" state variable on the whole length of the arm. To avoid the complexity of dynamic model, the same authors introduced in [Ivanescu 2004] an energy based control law relying only on those relations of the dynamic model that determines the energy stored in the system. In both cases, only simulation results were presented.

Inverse kinematic and dynamic models are also at the base of the work in [Kapadia 2011], where the first task space controller for continuum robot is presented. The trajectory is defined in terms of contact point (position, velocity and acceleration) between the Octarm robot [Tatlicioglu 2007] and the manipulated object, allowing, then, grasping and manipulation instead of only shape configuration tracking.

Other works in cable-driven continuum robots try to directly compensate the sources of error in the model of mechanical transmission of the movements. Usually, actuator and tip positions are measured and the hysteresis loop, due mainly to backlash and frictions, is determined and, then, used in a feedforward control scheme [Kesner 2010, Agrawal 2010b]. Furthermore, decoupled inverse kinematics has been described to allow for independent control of multiple sections [Camarillo 2009].

In medical field, model based approaches (with or without backlash compensation) have been extensively studied and applied for different designs and actuation modalities: active cannulas [Webster 2009, Dupont 2010] and tendon driven catheter or instrument [Camarillo 2008b, Agrawal 2010b, Xu 2006].

The works cited here are just a small panorama of all the model based solutions for the continuum robot control. However, it seems that this approach presents some inherent limitations deriving from the complexity of describing all the phenomena occurring in mechanical transmission. They still suffer from lack of accuracy due to unavoidable kinematic approximation and consequent modeling mismatches (at the time of implementation), actuation coupling between the different vertebrae not to mention the friction, extension and torsion of the actuation lines. Moreover, in the case where the continuum robot to control is inserted in a flexible shaft (such as for STRAS) the backlash behavior of the flexible instruments changes with the shape of the housing flexible shaft. Trying to model all these behaviors seems to result in *ad-hoc* solutions which are not smoothly transferable to other systems.

This suggests the necessity of having an external measurement of the robot tip

position or shape configuration to trustfully close the control loop. Some authors who proposed specific model based controllers, showed, in later works, the interest for combining model *a priori* knowledge with actual measurements [Agrawal 2012] or trying being independent from dynamic model [Yip 2014].

2.3.2 External Sensors

External sensing of continuum robot shape has also been demonstrated for example using electromagnetic (EM) field sensors. The electromagnetic sensor system is usually composed by a magnetic field generator and several (up to 5-7) sensor coils whose position and orientation can be retrieved when immersed in the generated magnetic field.

In [Agrawal 2012], the authors propose an adaptive robust control which combines EM sensors feedback with the actuator feedback. They obtain better results than using pure open-loop approach, but they also denote the necessity of improving the used dynamic model (of cable actuation) to make the results applicable to other flexible instruments.

In [Bardou 2010] and [Bardou 2012], the EM sensors are applied in a tendon-driven flexible system similar to STRAS: one coil attached at the endoscope head and another at the flexible instrument tip. Thanks to the sensed positions and taking into account the non-linearities, the entire shape of the flexible instrument is computed improving the control of such instrument. EM sensors were also adopted in [Penning 2011] to retrieve the tip location for completing a Cartesian control of the catheter prototype. Nevertheless, the reliability of their position and orientation measurements decreases if the sensor coil is outside the center of the generated magnetic field. To overcome this problem Reichl et al. [Reichl 2013] proposed to utilize a robotic arm to move the magnetic field generator so as to follow the sensor coil attached to a catheter, showing the feasibility of electromagnetic tracking in an extended workspace.

Despite this artifice, EM sensors still remain very sensitive to electromagnetic disturbances and are affected by metallic tools, which limits their usefulness in clinical applications. In addition, mounting an EM tracker in the distal part of a tool which is inserted in the human body is a challenging task. Moreover, some actuated flexible instruments achieve high curvature radius and the continual solicitation due to repetitive deflections may lead to break the filament of the sensor.

The same problem can occur integrating fiber optics into the flexible instruments for pose estimation purposes. It seems that small curvature radius cannot be achieved and that the bulky structure necessary as support is not yet suitable for surgical applications where the possibility of downscaling the instrument is fundamental [Roesthuis 2013].

An interesting alternative is shape sensing using vision by mono or multi camera systems, which can be external or embedded. In the next section, some examples

will be presented, taking into account the medical context and the objectives of the thesis which are focused in founding trustful measurements for controlling flexible instruments in flexible endoscopic surgery.

2.3.3 Pose estimation of continuum robots using vision

One of the first and most cited work where vision is employed to sense the shape of continuum robots is the one by Hannan and Walker [Hannan 2003], where they show their results using the multi-section Elephant's Trunk Manipulator. In this work, they use the robot as a planar robot moving on a plane perpendicular to the optical axis of the camera. The vertebrae of each section are detected in the image and the total shape of the robot is retrieved by fitting circumferences to the segmented vertebrae: one circumference for each set of vertebrae pertaining to the same robot section.

The same camera-robot configuration is employed in [Yip 2014], where a single bending section planar robot is controlled even in case of contact with obstacles. Here the camera is used to compute the displacement of a colored optical marker attached to the tip. After initialization, the estimation of the jacobian of the manipulator is updated according to the measured displacements from actuators and end-effector sensors.

To estimate the pose for any 3D configuration (not only planar movement) Camarillo et al. [Camarillo 2008a] investigated the potentiality of a voxel carving technique to retrieve the 3D shape of a single section bendable catheter. In this work, they use 3 cameras orthogonally placed and looking at the same scene surrounded by a green background to ease the detection.

A different multi camera approach is used in [Weber 2012] where micro cameras are mounted on the body of the flexible robot and their relative positions are computed by performing a bundle adjustment over tracked features visible by multiple cameras.

Unfortunately, these approaches would be difficult to directly transfer in surgical application since they rely on hypothesis hardly verifiable *in-vivo*: usually the background and point of view are not as favorable as in [Hannan 2003] and [Yip 2014] or multiple view techniques would mean more laparoscopes insertion in the case of [Camarillo 2008a]. The interesting work in [Yip 2014] should be extended for 3D movements assuring the visibility of the marker whatever the robot configuration is (even outside the plane). For [Weber 2012], the integration of cameras on the flexible instruments would be difficult and relative motion between camera and environment may complicate tracking and features correspondence.

Another possibility is to use stereo-vision to retrieve a more precise depth information. Stereo vision is used by Croom et al. to detect the shape of a concentric tube based on Self-Organizing-Maps [Croom 2010]. Also researchers developing the IREP robot have opted for embedded stereo vision to catch a more trustful depth

information. In [Reiter 2011, Reiter 2012], they proposed to learn the relationship between the visual information and the 3D position of the instrument. Some colored rings where attached around the flexible instruments to ease the detection and the computation of the visual information, which results from the composition of 2D descriptors on each of the stereo image.

Shape estimation of deformable instruments is a problem similar to that of estimating pose and shape of threads or tubes. In [Padoy 2012] a textured thread is tracked by modeling the thread as a B-spline directly in 3D and optimizing the 2D error between the actual detected thread and the reprojected B-spline. Here, stereo vision was adopted to discriminate ambiguous configurations. In [Caglioti 2006], the 3D pose of a tube is retrieved using solely one image and geometrical properties related to the apparent borders of the tube. The image of the tube is thought like the projection on the image plane of a series of circumferences. Flex and bitangent points are taken as "fiducials" points for creating the correspondence between upper and lower border points along all the apparent borders. The position of each couple of points and the tangents at these points are used to retrieve the position and orientation of the associated 3D circumference.

In the domain of endoscopy, where, usually, only monocular vision is available, the works in the literature become more rare. The only example, at the best of our knowledge, is the work by Reilink et al. [Reilink 2011, Reilink 2012]. They propose to use the monocular camera of ANUBIScope to estimate the pose of the flexible instruments inserted in the lateral channels. Supposing known the mechanical structure of the system, they retrieve the 3D tip position of the instrument from fitting the virtual model to the actual image. They actually present two methods: one is marker-less and the other one is relying on green markers attached on the instrument to ease features extraction.

Conclusion

Robotics solution for flexible endoscopic surgery are making inroad in the MIS scenario due to the increasing interest on no-scar or single-scar surgery techniques. One common possibility is to imagine the robotic system within a teleoperation framework where the surgeon closes the feedback loop correcting his/her movements according to what he/she observes in the endoscopic image.

To convert the robot into a real assistant for the physician (providing gesture guidance, autonomous positioning, ...) an automatic control strategy should be adopted. Due to the complexity in defining a complete and precise dynamic model and the dependency of such model to the particular system (or even configuration), open-loop control strategies are not the most adequate solution for retrieving real robot 3D pose.

The need of an external measurement arises, but the difficulty of attaching external sensors (e.g. EM sensors) and their fragility and possible incompatibility with other OR devices, bring to the idea of using only the embedded endoscopic camera to retrieve the 3D position of the flexible instruments.

Stereo-vision would permit to capture a precise depth information, but, since we are interested in conventional flexible endoscopic systems, which usually mount a monocular camera, some a-priori information must be considered to retrieve 3D pose from a 2D image. Pure geometrical and perspective based approach, though, seems to be very sensitive to noise (as will be explained next) and the use of a fix mechanical model would fail in case of mechanical plays, which would modify the mechanical model.

Thus, a new method is needed, which must be effective with conventional endoscopic systems configurations and robust even in *in-vivo* conditions. Two novel solutions are proposed to retrieve the 3D pose of the instrument from a single image. The first proposes to extend the mechanical model so as to take into account possible mechanical plays or model changes during the operation. Given the difficulty to satisfactorily model how the instrument is perceived by the camera, the second solution is a model-free approach to avoid problems related to model uncertainties. As the visual based technique accuracy is strictly related to the chosen feature and the precision with which these features are extracted, we decided to focus in both the aspects of the "pose from image" problem. Therefore, the first part of this thesis is dedicated to the selection of adequate features for 3D pose estimation and the segmentation of these features even in *in-vivo* scenario. The second part details the mentioned 3D pose estimation solutions.

Part I

Visual Features Selection and Extraction

Introduction

As described in the introduction. the main objective of this thesis is the pose estimation of flexible miniaturized instruments using monocular vision. However, as also remarked in [Azizian 2014], a good control or estimation require also a good choice and high precision in visual feature extraction. Therefore, both these objectives, feature extraction and pose estimation, are equally important and constitute the main contributions of this work. In this part, the focus is set on feature selection and extraction, whilst the second part will deal with the proper pose estimation problem.

In the previous chapter, several systems adopted in flexible endoscopic surgical procedures have been described and their controlling issues have been presented. Although a common thread can be individuated in the control and pose measurement strategies, the peculiarities that such systems present lead to the necessity of state the problem in details. Stereo vision instead of monocular, different points of view, absence of endoscopic camera (for catheter) are only a few of the reasons that make the problem to slightly change for each situation.

In this work, it has been decided to focus on devices for endoluminal or abdominal single port surgery or NOTES where their particular configuration of camera, lights and instruments are considered as general constraints worth to be studied and taken into consideration to validate possible solutions.

Thus, to comprehend the considered framework and the main contributions, it is fundamental to start with the description of the device taken as reference and present the main issues.

System Description

The device taken as reference is STRAS, a robotic system based on a short version of the manual Anubis platform developed by Karl Storz (Tüttlingen, Germany). The Anubis platform is a totally flexible system (in the sense that there are no discrete joints), initially developed for NOTES surgery. The short version consists of a main endoscope and two instruments, which can be used simultaneously. Instruments and main endoscope consist of a flexible passive shaft and are equipped with a bendable distal part, which is controlled from the proximal side using tendons. Each tendon runs inside a sheath inside the shaft of the endoscope (and instruments). The sheath is free to move inside the shaft so as to allow a good flexibility of the shaft itself and it is attached at the beginning of the bendable section of the endoscope (and instrument). At the base of this distal section the cables exit the shaft and are made to pass through dedicated supports attached to the support structure (backbone) before being attached to the tip of the bendable part.

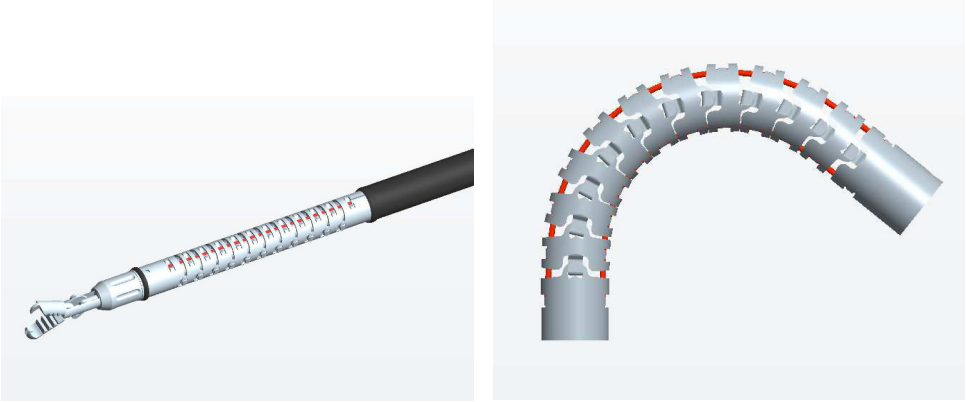


Figure 2.20: Mechanics of the bendable section of the instrument. A sequence of cilindrical vertebrae are linked together to form the backbone whose shape can be changed by two antagonist cables (in red) attached to the end of the last vertebra

The main endoscope has a diameter of 16 mm and it is equipped with a camera at the distal tip and a lighting system composed by two spots one on each side of the camera. The distal part of the endoscope can be deflected along two orthogonal directions and is actuated by two antagonistic pairs of tendons made of braided steel arranged in quadrature.

Three channels are available in the endoscope, which acts as a guide, for passing surgical instruments. One channel is located at the core of the shaft of the scope, while two channels (called lateral channels) are at the sides of the shaft. The instruments can be inserted in the lateral channels of the endoscope and thanks to the mechanical play between them and the channel, they can smoothly translate and rotate. The instruments have long flexible shafts (length 900 mm) and a bendable distal part (length 18.35 mm, diameter 3.5 mm). This part consists of hollow cylindrical-shaped vertebrae linked one another in two diametrically opposed points forming a rotational joint (see Fig. 2.20). A pair of antagonist tendons made of braided steel allow bending the distal part in a plane.

As the instruments are hollow, they can receive inserts equipped with distal tools. Tools can be mechanical (grasper, scissors) or electrical (knife, ball, hook).

STRAS Concept

The conceptual idea for STRAS was to design a teleoperated modular platform, which could be easily setup at the operating table side. It was also decided to keep as much of the original design of the Anubis platform in order to keep the sealing of the endoscope and the fluid management (water for irrigation and camera washing, air for insufflation and aspiration).

The main scope and the two instruments are held by a cart (see Fig. 2.21). This

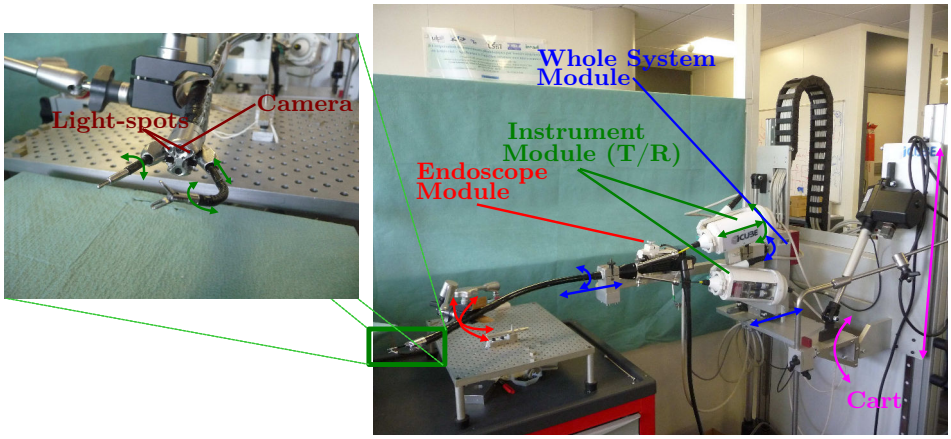


Figure 2.21: Overview of the system with the motorization of the different modules. Each color represents one module: endoscope, instruments, whole system (main endoscope + instruments) and cart. With the same colors are marked the movement that each motor of each module confer to the corresponding part of the system. Excluding the cart inclination and vertical translation (which are controlled by a wire remote-control), all the other DoF can be tele-operated and controlled by a master interface.

platform supporting the instruments and the endoscope has two motorized DoFs (up/down and inclination), which allows to position the whole system in a large variety of configurations depending on the access port to the operating area.

The entire slave system (main scope + instruments) can be translated and rotated, as one entity, around the entrance direction at the proximal side. Thanks to this combined translation of the endoscope and the instruments, the friction efforts of the instruments at the entrance of the channels of the endoscope decrease considerably.

All manual wheels and knobs on the instruments and scope have been replaced by small electrical motors (see scheme on Fig. 2.22) and the whole slave system can be controlled from a master console. The head of the endoscope has two DoFs and can be deflected along two orthogonal directions. Both instruments have 3 DOFs: they can be translated and rotated inside the channels of the main scope and their extremity can be bent in one direction. Moreover, mechanical instruments (such as graspers) can be opened and closed conferring an additional DOF.

The overall system has 12 motorised DoFs of whom 10 (those regarding translation, rotation and bending of endoscope and 2 instruments) are also teleoperable.

Modularity is another important concept beneath the design of STRAS slave system (see Fig. 2.21). The endoscope holds the motorization for its deflection and it can easily be attached to / detached from the cart. Two fixed modules contain the motorization for the translation and the rotation of the instruments. Instruments handles have been replaced by cylindrical bull-nose shaped casings containing the motorization for bending and grasper opening / closing. These modules can be

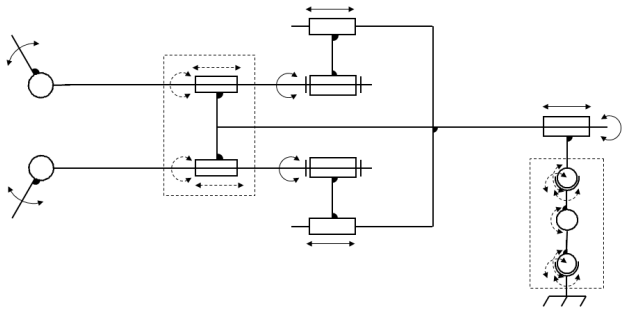


Figure 2.22: Motorization scheme of STRAS.

easily plugged onto the translation / rotation modules. This modular structure allows changing instruments during operations, while limiting the weight of the instrument modules to be manipulated and the required number of motors for a given set of surgical effectors.

Using visual perception in this context is challenging and the need of new approaches arises to extract the required visual information for pose estimation. In the next section, the main issues and the derived hypotheses are described and compared with state-of-the-art solutions.

Problem Statement

As the instruments bodies are uniformly black, the only element that presents visual features that can be related to 3D physical points is the attached surgical tool. Unfortunately, during the operation, the tool is often hidden since it continuously interacts with tissue for grasping or cutting. The focus, then, must be displaced to the body of the instrument bendable section which is textureless, though. It is not uncommon, then, to employ colored markers to overcome such problems and, in addition, to strengthen the features mainly for *in-vivo* environments where smoke or other fluids often corrupt the quality of the image making the features to appear less clear or fuzzy.

Additionally, the strong frontal lighting creates three zones in the image: a central zone where the light is very intense, an intermediate zone with soft lighting and a border zone where the image is almost dark. As one may expect, if the instrument is located in front of the camera, the intense light and its reflective material

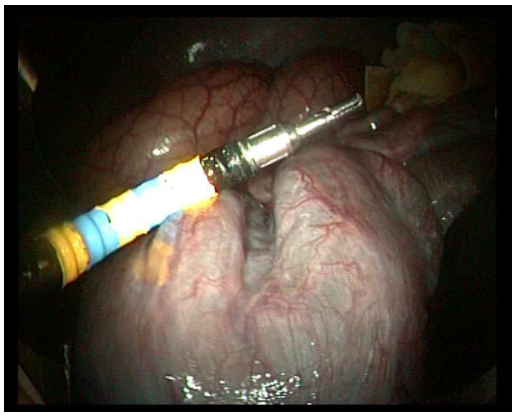


Figure 2.23: Example of *in-vivo* image grabbed by the endoscopic camera in the abdomen of a porcine model. In this example, the main characteristics of such kind of images are visible: distortion, strong central enlightment and consequent specularity effects in the middle of the instrument and mirror effect on the shiny surface of the organs.

cause strong specularity effects that modify the color appearance of the instrument, causing a wide white region in the middle of its bendable section (Fig. 2.23). This effect is not very attenuated even using a matte painting for the markers. Only when the instrument is no more exposed to this strong direct light, the specularity disappears.

Finally, the small elevation angle of the camera wrt to the instruments does not allow a clear vision of the surgical plane and increases the difficulty at perceiving the depth of the instruments. This particular point of view, indeed, impedes to clearly capture the shape of the instrument whose deflection is often perceivable only for the sheath (and consequently the markers') surface deformation while the outline stays almost unvaried for certain configurations.

In *in-vivo* environments further problems arise: the strong reflectance of the organ surfaces produces mirroring effects of the instrument, increasing the ambiguity of visual information (Fig. 2.23).

In the solution proposed here, all these remarks are taken into account, the only constraining hypotheses made are to consider that the instruments move in a free space and that no occlusions of the bendable section occur (neither occlusions due to external objects nor self-occlusions).

As shown in the precedent chapter, solutions regarding the pose estimation in such configuration are rare in the literature. However, this problem can be contextualised in a wider framework letting aside the *in-vivo* or surgical constraints. This framework is therefore the pose estimation of deformable objects using 2D images.

Depth Information from 2D images

The main issue in monocular vision is related to the retrieval of depth information which must be inferred from 2D visual measurements. The question is then which visual features contain shape information or how to combine features to recreate the 3D scene. Gibson, in his studies on animal/human visual perception, argued that humans obtain a lot of information on depth and environment structure by moving inside it. The *structure from motion* framework recalls this concept [Longuet-Higgins 1986] and aims to retrieve 3D information from a sequence of images. Usually, some features are tracked through the sequence and, based on them, the motion and the structure can be obtained by a projection model [Tomasi 1992] or using 2 or 3 views at a time [Hartley 2004]. A final bundle adjustment is made to find the best structure and movement explaining the scene observed in the sequence.

The outline of an object can be also used for shape retrieval even using uncalibrated camera. Different approaches have been developed [Cipolla 1992, Wong 2004, Utcke 2003] for object that can be described by circular cross-section (of different radius) where the axis goes through the center of the cross-section, and the cross-section is orthogonal to the axis which is usually straight and constant. They exploit revolution surface invariant to calibrate the camera and specific geometrical/projective properties to retrieve the shape from the surface outline [Cipolla 1992, Wong 2004] or the outline and two cross-section [Utcke 2003]. For depth determination, the dimension of at least one section [Wong 2004] or the outline deformation from several known point of view [Cipolla 1992] must be given.

Other authors proposed interactive methods for depth estimation with uncalibrated cameras. In [Sturm 1999], the user is meant to provide 3D constraints in terms of coplanarity, perpendicularity and parallelism which are then used to calibrate the image and perform the 3D reconstruction. In [Criminisi 1999], instead, the proposed solution needs a reference plane and a reference direction not parallel to the plane to compute respectively the vanishing line and the vanishing point. With this information, 3D affine measurement can be computed in terms of distances between planes parallel to the reference plane, area and length ratio on any plane parallel to the reference plane. Given the value of a reference distance or height, the real 3D position of other points can be computed and a 3D model can be created from these points.

For automatic structure determination using only one image, *Shape-from-Shading* (SfS) is another interesting approach. The original idea by Horn [Horn 1970] was to obtain the shape of a smooth opaque object from one view, given the knowledge of the surface photometry (which must be uniform), the position of the light source and other additional information such as the position of the observer with respect to the object and the light source.

Alternatively to shading, if all the possible deformations of the object pertain to a limited set, another solution is to determine the shape as a linear combination

of the elements of that set. In [Blanz 1999] this concept of *morphable models* is applied to faces, where the set of basic shapes is composed by 3D scanned head of 200 people including color information. Obviously, this process is applicable only to a specific category whose database is available.

If no information on the illumination source or no prior information is available on the kind of deformation, none of the previous framework would work. An alternative can be Shape-from-Template provided that a template is known. For template is meant a prior knowledge over the appearance (texture of the object), the shape (the resting position) and the material of the object (implying the dynamic of the deformation). Feature such as SIFT are extracted both in the template and the actual image and a warping is done based on these features. After that, the possible deformation of the object that would generate the actual image is computed taking into account the deformation law of the template according to its material. A strong hypothesis in this process is the isometry of the object even though more complex deformation law can be taken into consideration [Bartoli 2015].

Taking into consideration the problem we want to address (see *problem statement* section), in structure from motion the observed objects in the scene are supposed to be rigid which is not the case here. Non-Rigid Shape-from-Motion [Chhatkuli 2014] can be considered, but it is still a complicate and open problem. In the aim of performing automatic control, interactive methods are evidently unsuitable and shape from shading can lead to imprecise solutions due to the fact that material reflectance and illumination models are not easy to obtain for this kind of systems and, especially, in *in-vivo* environment. It is true, though, that do exist shape from shading methods that need less a-priori knowledge [Barron 2015]. Shading is, then, parametrized as a function of shape and reflectance which are considered unknown. To compensate this information loss, plausible constraints (about surface smoothness, uniformity of paint and illumination) are added to the original SfS problem to find the final solution. Unfortunately, in our problem, the strong saturation due to specular component of light make the most of instrument white, loosing, in this manner, important structure information.

On the other hand, the use of morphable models would mean to create a sufficiently dense database considering all the possible deformations of the instruments. Doing so, with the will of obtaining high precision, there can be the risk of creating a self-exhaustive set that would convert the approach almost to a look-up table making morphable model to lose their "interpolating role". In addition, the instruments in question are textureless and some new features must be found for the correspondence. Furthermore, the authors need a rough interactive alignment of the model with the actual photo and suppose some strong a-priori: camera distance, surface shininess and light direction remain fixed to the value estimated by the user.

Finally, neither an easy adaptation of shape from template is possible in the studied case due to its strong dependency on clear texture for initial warping, which the considered instrument does not present.

However, assuming that the intrinsic parameters of the camera are computed and some *a priori* information about the environment is known, neither the user interaction nor texture or lighting knowledge are needed for pose estimation. It is quite easy for a rigid object (with easy recognizable *physical* features) [Lepetit 2005], but further considerations must be added for solids of revolution (such as cylinders or sphere) [Doignon 2007] or freely deformable objects such as threads or flexible tubes (as described in Sec. 2.3). In [Caglioti 2006] a completely geometrical based method is used to estimate the position of a flexible tube in the space. After border extraction, upper and lower correspondence is settled based on bitangent and flex points. Each couple of points (upper and lower border points) is considered to be two points on the tube section, knowing, then, the 2D tangent to these points (computed numerically) gives information on the vanishing point which is used to compute the position and orientation of each section and, finally, composing all these sections, the entire tube. This approach, though, shows its weakness in those cases where the borders appear as almost parallel lines which is often the case with the ending part of the bending section of the considered surgical instruments. In this case, in fact, strong noise on the detected points could bring to a wrong estimation of tangents and, in turn, to the corresponding section orientation.

In the light of what discussed so far, a dual necessity arises:

1. Find which are the most suitable visual features for pose estimation.
2. Find a manner to obtain such features in the enunciated conditions while, as a first approximation, considering the instrument moving in free space.

These two problems are addressed in the next chapters: in section 3.5 the choice of a suitable image features for pose estimation is argued on simulation data and a novel method to obtain the selected feature is presented in chapter 4.

The suitable feature will be defined as that visual feature which, considering a given 2D error in its extraction, allows having a smaller estimation error on the 3D position of the surgical Tool Center Point (TCP). To carry out such study, the relationship between the robot configuration and image features and between the image feature and 3D tip must be known. This means knowing:

- the **kinematic model** of the instrument wrt the camera, that relates the values of its DOFs to its 3D shape.
- the **camera model** for being able to create the image relative to a specific 3D configuration of the instrument.
- an extended **Image Jacobian** that describes the relationship between the instrument joint velocities and feature movement in the image. The inverse of this Jacobian provides the relationship between the feature movements (or uncertainty) in the image and the instrument DOFs variations and, indirectly, on the 3D tip position variations.

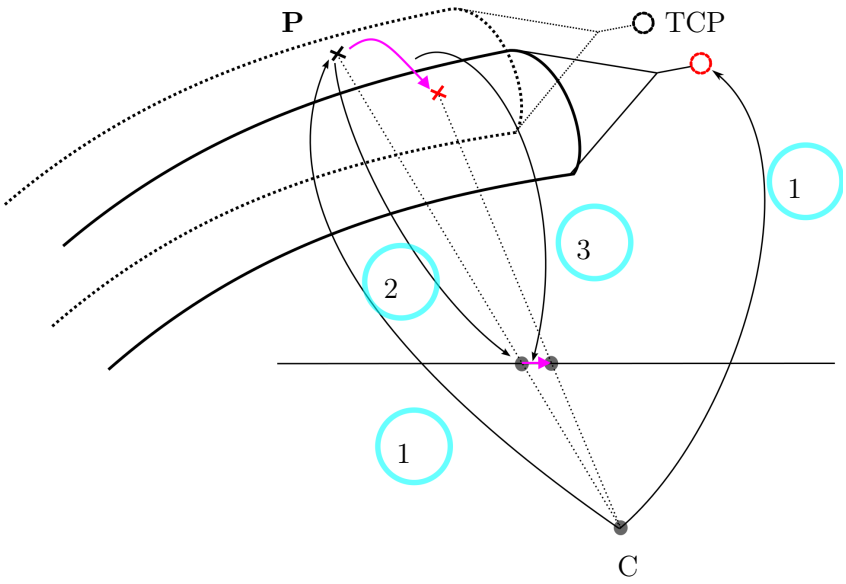


Figure 2.24: Scheme of the relationships between instrument pose and image projection. Number one indicates the kinematic model wrt the camera relating a point on the body of the instrument or the TCP 3D position to the world reference frame (i.e. the camera origin). Number two: with the camera model the projection of the 3D point of the image is computed. Finally (number three) the image Jacobian relates the 3D point velocity to the 2D image variation. Composing 1 and 2, an *extended* image Jacobian can be computed relating the robot DOFs velocity with the apparent velocity of the corresponding features in the image.

The next chapter (3), then, will clarify and formalize all these elements which will be fundamental tools all along this manuscript and not only for this preliminary study (chapter 4).

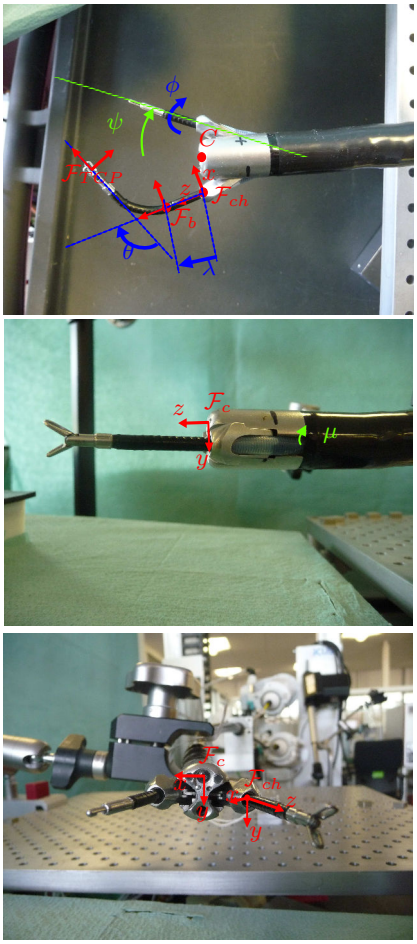
Chapter 3

System Modeling: Basis for Feature Selection

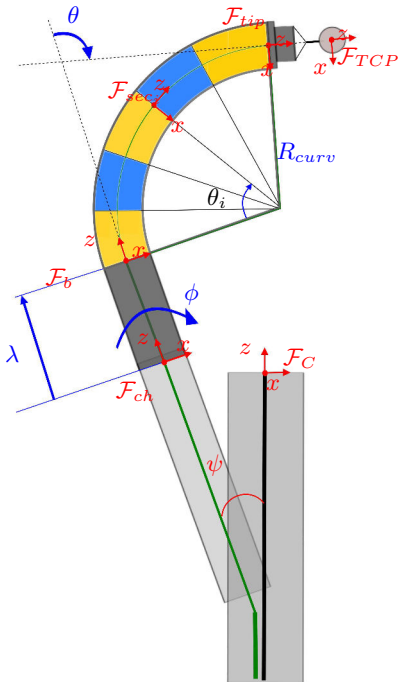
Contents

3.1 Geometric and Kinematic Model

For computing the model of the instrument, the attention will be focused on the tip of the endoscope and the camera-to-end-effector kinematic chain. This can be divided into two parts: one consists of the properly-said kinematic model defined by the controllable DoFs (marked in blue in Fig. 3.1(b)) of the instruments and the other represents the geometrical relationship between the channel exit and the endoscopic camera, i.e. the position and orientation of $\mathcal{F}_{ch}$ wrt $\mathcal{F}_c$ in Fig. 3.1(a). For the former part, the 3 DOFs of the instrument are considered (translation λ and rotation ϕ with respect to their axis and deflection θ) and for the latter the position of the exit of the channel (x_{ch}, y_{ch}) housing the instrument and its orientation (angles (ψ, μ) marked in green in Fig. 3.1(a)) with respect to the endoscopic camera optical axis are taken into account. The instrument can slightly move inside the channel and to entirely describe the instrument position wrt to the camera these parameters are needed. The only two restrictive yet reasonable hypotheses are that the instrument bends in a plane and that it conserves a constant curvature all along the bendable section.



(a) Tip of the Isiscope



(b) Scheme of the tip of the Isiscope

Figure 3.1: In (a) the picture of the tip of the endoscope with the instruments and in (b) the corresponding scheme. For sake of easiness and completeness in the representation, the scheme shows the instrument in the particular configuration with $\phi = 0$, this implies that the bending plane of the instrument is parallel to the one of the (x, z) camera. In (b) it can be seen that the estimated position of the end of the instrument duct does not coincide with its actual end, in fact it has been chosen to consider the end of the duct on the same plane (x, y) of the camera, to eliminate a sort of redundancy between λ and $z-$coordinate of the channel.

To formalize what discussed so far, the reference coordinate frames showed in Fig. 3.1 are defined. $\mathcal{F}_c$ indicates the camera frame, the frame associated to the end of the instrument channel is called $\mathcal{F}_{ch}$, whereas the one associated to the beginning of the bendable part is $\mathcal{F}_b$. The torus-like bendable part can be discretized in sections whose reference frame are $\mathcal{F}_{sec_i}$. For the last section, the one corresponding to the tip of the instrument, $\mathcal{F}_{tip}$ is used. Depending on the attached surgical tool, the distance from the tip to the surgical Tool Center Point (TCP) can vary and the frame associated to it ($\mathcal{F}_{TCP}$) would result in a rigid translation of $\mathcal{F}_{tip}$ along its z-axis.

Without losing sight of the main objective of using the camera for pose estimation, the camera frame will be considered as the world reference frame and all the other frames will be expressed with respect to it. Thus, $\mathcal{F}_{ch}$ can be obtained from $\mathcal{F}_c$ by a 3D translation and two subsequent elementary rotations: one of angle ψ around the camera $y-$axis and one of angle μ around the resulting $x-$axis.

The vector defining the mentioned translation is known from the CAD model of the instrument provided by the manufacturer and it will be denoted as $t^c_{c,ch}$ meaning the translation from the origin of $\mathcal{F}_c$ to the origin of $\mathcal{F}_{ch}$ (the two ordered subscript elements) expressed in the camera frame $\mathcal{F}_c$ (the superscript element). A similar notation will be used for the rotation matrix: R^c_{ch} express the orientation of $\mathcal{F}_{ch}$ with respect to $\mathcal{F}_c$ coordinate system. The frame of the base of the bendable section ($\mathcal{F}_b$), in turn, can be defined from $\mathcal{F}_{ch}$ by a translation along $\mathcal{F}_{ch}$ $z-$ axis with λ (first instrument DOF) and a rotation around the same axis of an angle ϕ (second instrument DOF).

The affine map components between the camera and the base of the bendable part can, then, be expressed as follows:

$$\mathbf{t}^c_{c,b} = \mathbf{t}^c_{c,ch} + \lambda\, \mathbf{z}^c_b \tag{3.1}$$

$$= \begin{bmatrix} x_{ch} \\ y_{ch} \\ 0 \end{bmatrix} + \lambda \begin{bmatrix} \sin\psi\cos\mu \\ \sin\mu \\ \cos\psi\cos\mu \end{bmatrix} \tag{3.2}$$

$$R^c_b = R_y(\psi)\, R_x(\mu)\, R_z(\phi) \tag{3.3}$$

where $\mathbf{z}^c_b$ is the third column of R^c_b, i.e. the unit length vector (also called versor) of the $z-$axis of the channel frame (whose direction coincides with the one of the base frame), whereas $R_x(.)$, $R_y(.)$ and $R_z(.)$ are the matrix form of rotations around x, y and z axis with angles μ, ψ and ϕ, respectively. Finally, it has been chosen to consider the end of the instrument duct fixed and on the same plane (x, y) of the camera, so as to eliminate a sort of redundancy between λ and $z-$coordinate of the channel. Consequently, the third component of $\mathbf{t}^c_{c,ch}$ is equal to zero.

Assuming a constant curvature along the bending section and the fact that the instrument axis pertains to the (z, y) plane of $\mathcal{F}_b$, the relation between the base

frame and $\mathcal{F}_{tip}$ (Fig. 3.1(b)) can be easily decomposed as a rotation around $\mathcal{F}_b$ $y-$axis with angle θ and a translation defined by the vector:

$$\mathbf{t}^b_{b,tip} = \begin{bmatrix} R_{curv}(1 - cos\theta) \\ 0 \\ R_{curv} sin\theta \end{bmatrix} \tag{3.4}$$

The curvature radius R_{curv} is gathered from the definition of radians: $R_{curv} = L/\theta$ where L is the length of the arc corresponding to θ. In this case L is the length of the whole bending section of the instrument (arc connecting $\mathcal{F}_b$ to $\mathcal{F}_{tip}$ in Fig. 3.1(b)) and θ the bending angle.

Finally, the orientation and translation of the tip with respect to $\mathcal{F}_c$ results to be:

$$R^c_{tip} = R^c_b R_y(\theta) \tag{3.5}$$

and

$$\mathbf{t}^c_{c,tip} = \mathbf{t}^c_{c,b} + \mathbf{R}^c_b \mathbf{t}^b_{b,tip} \tag{3.6}$$

where $R^{tip}_b = [R_y(\theta)]^T$.

Camera to i-th section

The geometrical transformation between the camera coordinate frame and the frame $\mathcal{F}_{sec_i}$ associated to a possible section of the bendable part of the instrument is obtained from $\mathcal{F}_b$ as

$$R^c_{sec_i} = R^c_b R_y(\theta_i) \tag{3.7}$$

$$t^c_{c,sec_i} = t^c_{c,b} + R^c_b(t^b_{b,sec_i}) \tag{3.8}$$

where

$$t^b_{b,sec_i} = \begin{bmatrix} L_i/\theta_i(1 - cos\theta_i) \\ 0 \\ L_i/\theta_i sin\theta_i \end{bmatrix}, \tag{3.9}$$

and L_i are the arc length defined by the angle θ_i associated to the section in question. Thus, L_i/θ_i is the curvature radius R_{curv}.

Every point $\mathbf{P}_i$ of the surface of the torus can be expressed as a point pertaining on the circumference or radius r (radius of the instrument) individuated by a specific section ($\mathcal{F}_{sec_i}$) perpendicular to the instrument axis. Knowing the coordinate of $\mathbf{P}^{sec_i}_i$ with respect to $\mathcal{F}_{sec_i}$, the coordinates on the surface wrt to camera frame can be computed straightforward using (3.8):

$$\mathbf{P}^c_i = R^c_{sec_i} \mathbf{P}^{sec_i}_i + t^c_{c,sec_i}$$

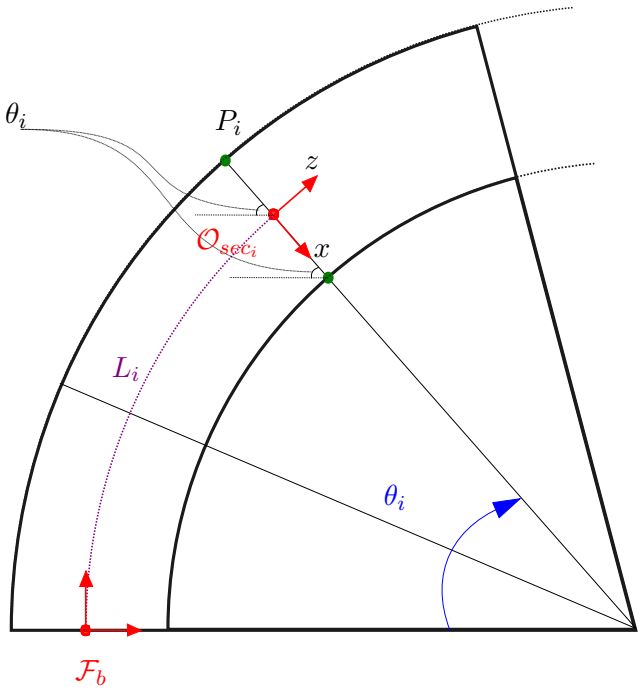


Figure 3.2: Scheme showing the geometrical relations between the different elements in a section.

Referring to Fig. 3.2, a relation between the base and the points of the sections can be defined. In fact

$$t^b_{O_{sec_i},\mathbf{P_i}} = \begin{bmatrix} P^{sec}_{i_x} cos\theta_i \\ P_{i_y} \\ -P^{sec}_{i_x} sin\theta_i \end{bmatrix} \tag{3.10}$$

and, joining the two results (3.8) and (3.9) :

$$t^b_{b,\mathbf{P_i}} = t^b_{b,sec_i} + t^b_{O_{sec_i},\mathbf{P_i}}$$

$$t^b_{b,\mathbf{P_i}} = \begin{bmatrix} R_{curv}(1 - cos\theta_i) \\ 0 \\ R_{curv} sin\theta_i \end{bmatrix} + \begin{bmatrix} P^{sec}_{i_x} cos\theta_i \\ P_{i_y} \\ -P^{sec}_{i_x} sin\theta_i \end{bmatrix}. \tag{3.11}$$

Once obtained the expression in the base frame it is easy to find the one with respect to the camera frame, which is, for the translation component:

$$t^c_{c,\mathbf{P_i}} = t^c_{c,b} + R^c_b \, t^b_{b,\mathbf{P_i}} \tag{3.12}$$

and for the rotation component:

$$R^c_{sec_i} = R^c_b \, R^b_{sec_i}. \tag{3.13}$$

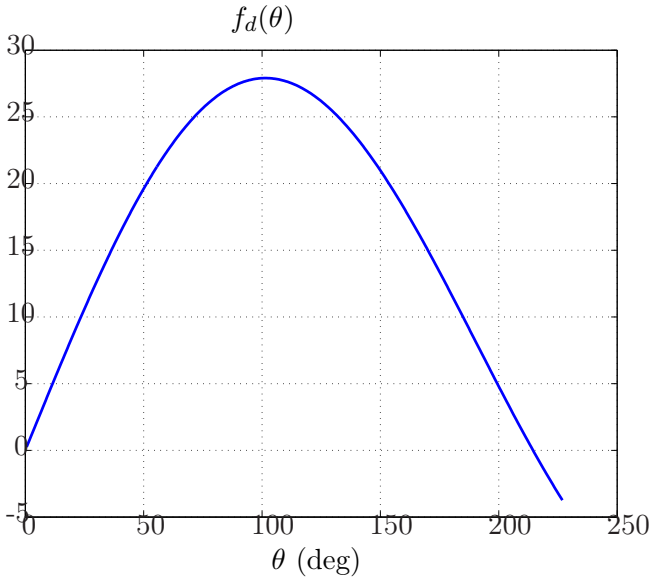


Figure 3.3: Graphic of the numerically computed relationship between the deflection angle θ and the distance of the TCP wrt to instrument rotation axis z_b.

Camera to TCP

The extension for the TCP is straightforward. In fact, its reference coordinate frame shares the same orientation as the frame associated to the tip (expressed by the rotation matrix in 3.13) and the position is obtained as a translation along the tip z-axis. The resulting orientation and translation wrt the camera are:

$$\begin{aligned} R^c_{TCP} &= R^c_{tip} \\ \mathbf{t}^c_{c,TCP} &= \mathbf{t}^c_{c,tip} + R^c_b \, R^b_{tip} \begin{bmatrix} 0 \\ 0 \\ l_{tool} \end{bmatrix} \end{aligned} \tag{3.14}$$

where l_{tool} is the distance between the end of the bendable part (i.e. the instrument tip) and the surgical tool center point (TCP).

3.2 Inverse Position Kinematic Model

No closed form solution is available for this problem and numerical optimization approach such as Gauss Newton algorithm using the inverse of the Jacobian would return only one solution. However, for obtaining all possible solutions the following procedure can be adopted.

First, the desired Cartesian position of the TCP is expressed with respect to $\mathcal{F}_{ch}$

with coordinates $\mathbf{P}^* = (X^*, Y^*, Z^*)$:

$$t^{ch}_{ch,TCP} = \begin{bmatrix} 0 \\ 0 \\ \lambda \end{bmatrix} + R^{ch}_b t^b_{b,tip}$$

where $R^{ch}_b = R_z(\phi)$. Therefore, the distance of the point from the instrument rotation axis can be expressed as:

$$f_d(\theta) = \sqrt{X^{*2} + Y^{*2}} = \left| \frac{L}{\theta}(1 - cos\theta) l_{tool} \sin(\theta) \right| \tag{3.15}$$

Fig. 3.3 shows the relationship $f_d(\theta)$ between angle θ and the distance of the TCP from the instrument rotation axis, the maximum is reached for θ_{sing} which is approximatively 101.77°. Looking at this figure (obtained with $L = 18.35$ and $l_{tool} = 15.8$), it can be seen that the previous equation generally has two solutions (θ_1 and θ_2), one on each side of θ_{sing} plus their opposite negative solutions. In fact, if a point in the workspace (without taking into account orientation) is attained by a specific deflection θ_1 and a specific rotation angle ϕ_1, it can be attained also by deflection equal to $-\theta_1$ and rotation $\phi_1 + \pi$. Finding the deflection angles is equivalent to solve:

$$\theta_{1,2} = \arg\min_{\theta} \sqrt{X^{*2} + Y^{*2}} - f_d(\theta).$$

This problem can be solved numerically. One solution is obtained initialising the optimization next to $\theta = 0$ and, for the second, near to $\theta = \theta_{max}$ where θ_{max} is the maximum deflection that the instrument can achieve (equal to 120 °in this case). The corresponding rotation angles ϕ is computed directly from $\mathbf{P}^*_1$ coordinates as

$$\phi_1 = \text{atan}\,2(Y^*, X^*);$$

Finally, the translation parameters λ_1 (λ_2) are computed directly from the direct kinematic model relations:

$$\lambda_{1,2} = Z^* - \frac{L}{\theta_{1,2}} \sin\theta_{1,2} - l_{tool} \cos\theta_{1,2}.$$

For the scope of the preliminary study, only one solution will be considered such that θ is positive and smaller than θ_{sing}.

3.3 Camera Projection Model

To relate the 3D shape of the instrument with visual features, a model of the camera is needed so as to know how the 3D points are projected onto the image plane. The pinhole model is a simple way to describe this mathematical relationship. In this model the camera aperture is described as an infinitesimal point where no lenses are used to collect the light and no distortion or blurring are taken into account.

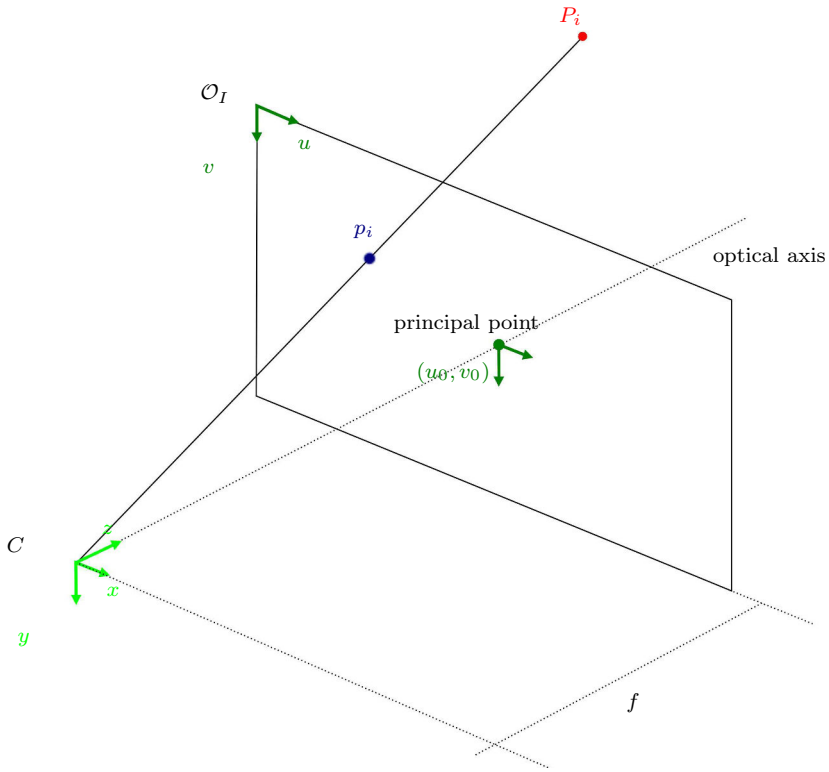


Figure 3.4: Scheme of the adopted pinhole model and image formation. For easing the representation of the image formation the plane is considered positioned at $+f$ with respect the optical center reference coordinate frame.

The rays of light coming from the environment travel along a single path through the pinhole and intersect the image plane where its upside-down image is formed. The reference frame associated to the camera is usually a right-handed coordinate system where the x-axis is taken as the horizontal direction and the y-axis as the vertical (downward) direction. This means that the optical axis (gaze direction) is the positive z-axis. The image plane is parallel to the camera sensor plane (x, y) and positioned at $-f$ along the camera z-axis where f is the apparent focal distance of the camera. A totally equivalent representation is chosen in Fig. 3.4 which does not pretend to be a physical representation of a camera. The fact of positioning the image plane at $+f$ is just for easing the representation since it avoids the formation of an upside down image. The intersection of the optical axis with this plane individuates the *principal point* of the image plane marked as (u_0, v_0).

The camera realizes the projection of a point $\mathbf{P}_i \in \mathbb{R}^3$ of the Euclidean space into a point $\mathbf{p}_i \in \mathbb{I}^2$ of the image plane whose homogeneous coordinates $(u_i, v_i, 1)$ are expressed in pixels. Assuming the point is expressed with respect to the camera frame, the projection in metric units of such point (X_i, Y_i, Z_i) onto the image plane can be written as:

$$\begin{bmatrix} \tilde{u}_i \\ \tilde{v}_i \end{bmatrix} = \frac{f}{Z_i} \begin{bmatrix} X_i \\ Y_i \end{bmatrix}. \tag{3.16}$$

The corresponding image coordinates (u_i, v_i) in pixel units are obtained from $(\tilde{u}_i, \tilde{v}_i)$ with the following relation:

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} = \begin{bmatrix} D_u s_u \tilde{u}_i - D_u \cot(\varphi) \tilde{v}_i \\ D_v \tilde{v}_i / \sin(\varphi) \end{bmatrix} + \begin{bmatrix} u_0 \\ v_0 \end{bmatrix} \tag{3.17}$$

where D_u and D_v are needed to change from metric units to pixels, s_u is the possible scale factor between the vertical and horizontal sizes of the sensor and φ is the angle between the image reference frame axis (usually $\varphi = \pi/2$). The addition of $[u_0, v_0]^T$ moves the principal point to the top left corner. This principal point can be taken as origin of the image coordinate system.

These expressions can be summarized in matricial form as:

$$\mathbf{p}_i = \frac{1}{Z_i} \mathbf{K} \left[\mathbf{I}_{3x3} \mathbf{O}_{3x1} \right] \mathbf{P}_i = \mathbf{K} \left[\mathbf{I}_{3x3} \mathbf{O}_{3x1} \right] \tilde{\mathbf{m}}_i \tag{3.18}$$

where

$$\tilde{\mathbf{m}}_i = \frac{1}{Z_i} \mathbf{P}_i = \begin{bmatrix} \tilde{x}_i \\ \tilde{y}_i \\ 1 \end{bmatrix}$$

$$K = \begin{bmatrix} \alpha_x & -\alpha_x \cot(\varphi) & u_0 \\ 0 & \alpha_y / \sin(\varphi) & v_0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3.19}$$

$\alpha_x = D_u s_u f$ and $\alpha_y = D_v f$ are the apparent focal distances expressed in horizontal and vertical pixels respectively.

To increase the field of view of the surgeon, flexible endoscopies usually mount a wide angle lens which return a highly distorted image both tangentially and, above all, radially. It is denominated radial distortion when the projected point suffers a deviation along the line connecting it to the principal point, whose value depends on its distance with the principal point and it arises because the "thin lens" hypothesis considered in the pin-hole model is not valid anymore.

Another source of image distortion is the fact that centers of curvature of lens surfaces are not always strictly collinear. This distortion type, called decentering distortion, has both a radial and tangential component [Slama 1980, Heikkila 1997].

The model just presented then becomes insufficient to describe the whole phenomena of image formation, therefore a more accurate model must be developed which derives from the combination of the pinhole model with radial and tangential distortions, obtaining:

$$\mathbf{m}_i = \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = \begin{bmatrix} \tilde{x}_i + \delta x_i^{(r)} + \delta x_i^{(t)} \\ \tilde{y}_i + \delta y_i^{(r)} + \delta y_i^{(t)} \\ 1 \end{bmatrix} \tag{3.20}$$

The radial distortion contribution is [Heikkila 1997]:

$$\begin{bmatrix} \delta x_i^{(r)} \\ \delta y_i^{(r)} \end{bmatrix} = \begin{bmatrix} \tilde{x}_i (k_1 r_i^2 + k_2 r_i^4 + k_5 r_i^6 + \cdots) \\ \tilde{y}_i (k_1 r_i^2 + k_2 r_i^4 + k_5 r_i^6 + \cdots) \end{bmatrix} \tag{3.21}$$

and the tangential contribution is:

$$\begin{bmatrix} \delta x_i^{(t)} \\ \delta y_i^{(t)} \end{bmatrix} = \begin{bmatrix} 2k_3 \tilde{x}_i \tilde{y}_i + k_4 (r_i^2 + 2\tilde{x}_i^2) \\ k_3 (r_i^2 + 2\tilde{y}_i^2) + 2k_4 \tilde{x}_i \tilde{y}_i \end{bmatrix} \tag{3.22}$$

where $r_i = \sqrt{\tilde{x}_i^2 + \tilde{y}_i^2}$ is the distance of the normalized points from the principal point.

The parameters k_1, k_2, k_5 are the coefficients for radial distortion (sixth order is usually sufficient) and k_3, k_4 are the ones for tangential distortion. To completely define the projection transformation, these parameters must be known together with the so called *intrinsic parameter* α_x, α_y and (u_0, v_0). It is fundamental, then, a preliminary calibration step of the camera to estimate that parameters. For this purpose, the methods of Doignon [Doignon 1999], Zhang [Zhang 1999] and Hiekkilä [Heikkila 1997] can be used. The last two have been implemented by Yves Bouguet and are available as a Matlab Toolbox [Bouguet 2013] which has been employed here.

Several views of the same known calibration grid are grabbed and camera-grid relative pose (so called extrinsic parameters), distortion and intrinsic parameters

are optimized to obtain the least reprojection error defined as the 2D error between the detected grid corners and the reprojected corners using the estimated model parameters.

Knowing the distortion parameters, the inverse of the distortion transformation can be computed and applied to the image obtaining its *rectified* version. Once rectified, the pinhole model is a good approximation of the camera behavior and the perspective projection relation are valid and can be possibly utilized to infer 3D information from image measurements.

3.4 Extended Image Jacobian

The last step is to relate the 3D velocity of given 3D points with the image motion of their projections. The transformation matrix describing such relationship is called Image Jacobian:

$$\dot{\mathbf{p}}_i = \mathbf{J}_{I_i} \dot{\mathbf{P}}_i \tag{3.23}$$

where $\mathbf{P}_i$ is a point defined in the task space (camera frame: $\mathbf{P}_i = t^c_{c,P_i}$) and $\dot{\mathbf{p}}_i$ is the feature parameter rate of change in the feature (image) space.

Since camera and endoscope guides are linked together, the relative movement of a point $\mathbf{P}_i$ on the instrument surface (wrt the camera) depends solely on the movements of the flexible instrument itself:

$$\dot{\mathbf{P}}_\mathbf{i} = \mathbf{J}_{g_i} \dot{\mathbf{r}}. \tag{3.24}$$

$\mathbf{J}_{g_i}$ is the geometrical Jacobian relating the joint velocities ($\dot{\mathbf{r}}$) to the velocity of the chosen point on the instrument.

Thus, defining the 3D point $\mathbf{P_i} = [X_i, Y_i, Z_i]^T$ and its relative 2D visual feature point $\mathbf{p}_i = [x_i, y_i, 1]^T$ and considering equation (3.24), the relationship described in (3.23) can be written as:

$$\dot{\mathbf{p}}_i = \mathbf{J}_{I_i} \mathbf{J}_{g_i} \dot{\mathbf{r}} \tag{3.25}$$

where $\mathbf{r} = [\lambda \, \phi \, \theta]^T$ is a specific instrument configuration.

The product between $\mathbf{J}_{I_i}$ and $\mathbf{J}_{g_i}$ can be considered an *extended* image Jacobian $\mathbf{L}_i$:

$$\mathbf{L}_i = \mathbf{J}_{I_i} \mathbf{J}_{g_i} \tag{3.26}$$

which relates the joint velocity of the instrument with the image motion of a point of the instrument itself.

3.4.1 Image Jacobian

Considering only the 3D translational effects on the 2D features, the computation of the image Jacobian is equivalent to find the gradient:

$$\mathbf{J}_{I_i} = \left[\frac{\partial \tilde{\mathbf{m}}_i}{\partial X_i}, \frac{\partial \tilde{\mathbf{m}}_i}{\partial Y_i}, \frac{\partial \tilde{\mathbf{m}}_i}{\partial Z_i} \right].$$

The final expression, then, can be derived from 3.18 and 3.19 resulting in:

$$\mathbf{J}_{I_i} = K \begin{bmatrix} \frac{1}{Z_i} & 0 & \frac{-X_i}{Z_i^2} \\ 0 & \frac{1}{Z_i} & \frac{-Y_i}{Z_i^2} \\ 0 & 0 & 0 \end{bmatrix} = K \begin{bmatrix} \frac{1}{Z_i} & 0 & \frac{-\tilde{x}_i}{Z_i} \\ 0 & \frac{1}{Z_i} & \frac{-\tilde{y}_i}{Z_i} \\ 0 & 0 & 0 \end{bmatrix} \tag{3.27}$$

3.4.2 Distortion Effect

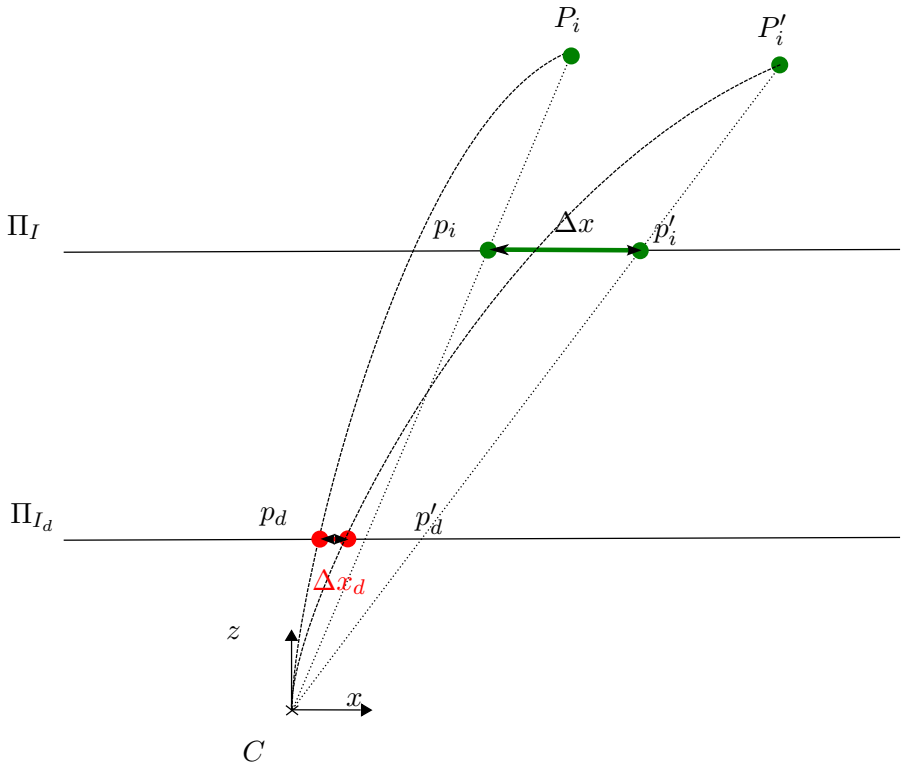


Figure 3.5: Scheme for the computation of the Distortion Gain. The plane Π represents the image plane as if no distortion ocurred whereas in image plane Π' the image formation is done taking into account radial and tangential distortions. The Distortion Gain is computed as the relative image displacement seen on the two image planes.

To take into account the distortion effect an equivalent gain is computed which expresses a magnification factor between the displacement on the image plane without distortion with respect to the same displacements on the image plane with distortion. In fig 3.5, the principle is schematized for the x coordinate on the plane (x, z). The ray of light passing through the optical center intersects the ideal (= no distortion effects) image plane at a point p_i which can be computed with the projective relationships described in the precedent section with equations (3.16) and (3.17) (to obtain the result in pixel units).

For the effect of wide angle lenses, though, the ray of light is "bent" before intersecting the real image plane (Π_{I_d}). As showed in precedent section, the projection of P_i onto this plane can be computed with (3.18) considering $\mathbf{m}$ (from eq. (3.20)) instead of $\tilde{\mathbf{m}}$:

$$\mathbf{p}_d = \mathbf{K}\left[\mathbf{I}_{3x3}\mathbf{O}_{3x1}\right]\mathbf{m}_i. \tag{3.28}$$

The distorted projection in metric units $\mathbf{m}$ can be computed through (3.21) and (3.22) which, in turn, require the knowledge of the distortion parameters usually available after a the calibration stage.

Let us, now, suppose that P_I moves outside the projection line achieving the new position P'_I and let us indicate the perceived corresponding motion on the image plane Π_I with $(\Delta x, \Delta y)$. The corresponding new position (p'_d) on the real image plane Π_{I_d} can be computed relying on eq. (3.28). Afterwards, the displacement vector $(\Delta x_d, \Delta y_d)$ describing the translation between p_d and p'_d can be easily retrieved. The equivalent distortion gain is then defined as:

$$\mathbf{G}_d = \begin{bmatrix} \frac{\Delta x_d}{\Delta x} \\ \frac{\Delta y_d}{\Delta y} \\ 1 \end{bmatrix} \tag{3.29}$$

and, post-multiplying it by J_I, the new Image Jacobian J_{I_d} taking into account distortion effects is obtained as:

$$J_{I_d} = \mathbf{G}_d^T \mathbf{J}_I. \tag{3.30}$$

3.4.3 Geometric Jacobian

As explained before, the geometric Jacobian is the matrix transformation which describes the relationship between the joint velocity and the end-effector velocity. Since the considered robot has 3 DOFs the Jacobian can be thought as a block matrix where each block is the contribution of each joint action onto the 3D position of the chosen instrument point.

$$\dot{\mathbf{P}}_i = \left[\begin{array}{c|c|c} A_\lambda & A_\phi & A_\theta \end{array}\right]_i \begin{bmatrix} \dot{\lambda} & \dot{\phi} & \dot{\theta} \end{bmatrix}^T \tag{3.31}$$

where $A_{r_k} = \partial \mathbf{P_i} / \partial r_k$, with r_k being alternatively λ, ϕ or θ. The final expression, then, would result by the computation of these three blocks separately.

A variation of λ affects the position of P_i according to the channel direction. Indeed, when computing $\partial t^c_{c,\mathbf{P_i}} / \partial \lambda$ the only element depending on λ is t^c_b i.e. the position of the base with respect to the camera finally giving

$$A_\lambda = \frac{\partial t^c_{c,\mathbf{P_i}}}{\partial \lambda} = \frac{\partial t^c_{c,b}}{\partial \lambda} = \begin{bmatrix} \sin\psi\cos\mu \\ -\sin\mu \\ \cos\psi\cos\mu \end{bmatrix} \tag{3.32}$$

which is the channel z-axis direction which correspond to z_c^b (the 3rd column of R_b^c).

The rotation around the channel axis can be considered as a revolute joint and, as for any revolute joint, the effect of this joint on another point of the kinematic chain is equal to the cross product between the angular velocity screw and the vector joining the center of rotation and the point of interest. Since the rotation is around the $\mathcal{F}_b$ $z-$axis the angular velocity screw will correspond to the third column of R_b^c and the radius is $P_i^c - t_{c,b}^c$ always with respect to the camera frame, obtaining:

$$A_\phi = -[(P_i^c - t_{c,b}^c) \times z_b^c)]. \tag{3.33}$$

The last block concerns the variation of θ. Since

$$\begin{aligned} \mathbf{t}_{c,P_i}^c &= \mathbf{t}_{b,c}^c + R_b^c \mathbf{P_i}^b \\ &= \mathbf{t}_{b,c}^c + R_b^c \left(\begin{bmatrix} L_i/\theta_i(1-\cos\theta_i) \\ 0 \\ L_i/\theta_i \sin\theta_i \end{bmatrix} + \begin{bmatrix} P_{i_x}^{sec}\cos\theta_i \\ P_{i_y} \\ -P_{i_x}^{sec}\sin\theta_i \end{bmatrix} \right) \end{aligned} \tag{3.34}$$

the only non constant element with respect to θ is $\mathbf{P_i}^b$. The variation of $\mathbf{P_i}$ coordinates with respect to the camera can be computed pre-multiplying the same variation expressed with respect to the base of the bendable part by R_b^c:

$$A_\theta = \frac{\partial \mathbf{t}_{c,P_i}^c}{\partial \theta} = R_b^c \frac{\partial \mathbf{P}_i^b}{\partial \theta}$$

with

$$\frac{\partial \mathbf{P_i}^b}{\partial \theta} = \begin{bmatrix} k_\theta L/\theta \sin\theta_i - P_{i_x}^{sec} k_\theta \sin\theta_i - L/\theta^2(1-\cos\theta_i) \\ 0 \\ k_\theta L/\theta \cos\theta_i - P_{i_x}^{sec} k_\theta \cos\theta_i - L/\theta^2 \sin\theta_i \end{bmatrix}$$

where $\theta_i = \theta k_\theta$ being $k_\theta = (i-1)/(n_{sec}-1)$ where n_{sec} is the number of section and $i = 1, \cdots, n_{sec}$ is the considered section.

3.4.3.1 Point on the axis and TCP

The computation of the different blocks for a point on the axis of the instrument bendable segment or for the TCP is totally similar to the one described previously. Actually, the block A_λ and A_ϕ are equal to the previous ones (substituting P_i with a point in the axis), whereas A_θ needs to be recomputed or derived from the precedent without considering instrument thickness (i.e. $P_i^{sec} = [0\,0\,0]^T$) and a further transformation in the case of the TCP. Finally, for a general point on the axis

$$A_\theta = R_b^c \begin{bmatrix} k_\theta L/\theta \sin\theta_i - L/\theta^2(1-\cos\theta_i) \\ 0 \\ k_\theta L/\theta \cos\theta_i - L/\theta^2 \sin\theta_i \end{bmatrix}$$

and for the TCP

$$A_\theta = R_b^c \begin{bmatrix} L/\theta \sin\theta - L/\theta^2(1-\cos\theta) + l_{tool}\cos\theta \\ 0 \\ k_\theta L/\theta \cos\theta - L/\theta^2 \sin\theta - l_{tool}\sin\theta \end{bmatrix}$$

Both 3.27 and 3.31 still require the knowledge of the 3-D point depth. The only issue is that the considered points do not have a direct physical sense in 3D but are only apparent features depending on the pose of the instrument with respect to the camera. Their 3D position can be computed for a specified robot configuration $\mathbf{r}$ (see Appendix A for details) giving to the extended Jacobian only a local meaning. The dependency of the apparent points on the pose of the instrument with respect to the camera affects also their velocity. This effect has to be taken into account in the geometrical 3D Jacobian and further details are given in Appendix A.

All these relationships between robot (i. e. the robotized instrument) configuration and the image are of fundamental importance for any control or pose estimation approach based on image measurements. This result will have a duple utility: studying the pose estimation problem in simulated data for choosing a proper visual feature to use for such purpose and also providing a model of instrument visual perception that will be exploited in the first proposed solution (Sec. 5).

3.5 Theoretical Study for Features Selection

The precision of vision based methods is inherently linked to the intrinsic parameters of the camera and with the accuracy of the image features location. To know which precision can be achieved with such methods and, based on that, which visual features to adopt, a preliminary study has been carried out. It consists in quantifying the error on the pose estimation when a misplacement of one pixel is done on the image.

For sampled test positions, the Inverse Kinematic Model (see section 3.2) of the instrument is used to compute the corresponding Degrees of Freedom (DoFs) of the instrument. Then $J_{g_{tip}}$ the Jacobian from the DoFs variations to the 3D position changes of the TCP of the instrument and J_I the Jacobian from the DoFs to the image features are computed.

The local variance of the estimated TCP position is computed as

$$\Sigma_{TCP} = (J_{g_{tip}} J_{I_d}^T \Sigma_{2D} J_{I_d})^{-1} J_{g_{tip}}^T$$

with

$$\Sigma_{2D} = I_{2n_{feat}}$$

assuming a uniform independent one pixel covariance error on each coordinate of the feature location.

Three cases are now studied: (1) Considering markers centroids and 3DOFs (the instrument is totally constrained in the channel, case reported in [Reilink 2012]), (2) Considering instrument borders, (3) Considering the apparent corners of some visual markers which divide the body of the instrument in equally spaced sections. For sake of result visualization, the focus has been centered on the estimation of the TCP coordinates in the central (x, z) plane of the instrument work-space, which is defined by the origin of $\mathcal{F}_{ch}$ and two vectors parallel to the x and z coordinate of $\mathcal{F}_C$.

A first analysis can be done on the outline of the variance on the defined principal (x, z) instrument plane (Fig. 3.6). A sort of symmetry can be observed with respect to the instrument channel axis and, focusing on those configurations where $\phi \in [-\pi/2, \pi/2]$ and $\theta > 0$ (in the work-space on the right of the blue line), its tendency shows a loss of accuracy on the tip position estimation for instrument positions farther from the endoscopic camera. This is quite an expected behavior due to the perspective projection which makes the size of the object in the image to decrease as the distance from the camera increases. In fact, when the object is far away, the resolution on the object given by the image is lower, since one pixel represents a wider region on the object itself. Thus, a larger movement or deformation would be necessary so that to be visible in the image, at a pixel scale.

A second level analysis is to compare the different descriptors in terms of mean and maximum errors on the TCP position estimation. For each robot configuration the direction with the largest variance in the TCP position estimation are taken and sorted (Fig. 3.7(a)). It can be straightforwardly seen that the use of centroids only does not permit a proper discrimination of different configurations, whereas the other considered descriptor seems to provide better results.

Features based on the borders only does not seem to be a robust choice, either. This can be explained by the fact that more than one optimal solution can be found where the estimated borders are inside the actual apparent border of the instrument: in this case, in fact, the image residual is zero even if the estimated border does not really represent the real apparent border. Two points on each apparent border could be enough to know the *scale* to correctly identify the configuration. The use of the term *scale* here is not the most adequate since the instruments are deformable and the relative distances between points vary. Knowing some physical points on the border, actually, gives an information on the shape of the instrument: for example, if the beginning and the end of the instrument positions are known and visible in the image, this information can be used to improve the definition of image residuals, by constraining the estimated border to recover the totality of the actual border.

To strengthen this shape information, one can think to mark more than two points on the instruments that can be easily detected in the image and that carry shape information for example using the apparent corners of colored markers. Using markers painted on the whole surface of the instrument is, at the same time, a mean to ensure the visibility of the feature point for any rotation of the instrument

(which would be not the case if dots are painted in specific positions). This seems to significantly help the pose estimation as can be seen when zooming on the worst cases of each approach (Fig. 3.7(b)).

Among the available corners, the ones associated to the end of the bendable part seem to carry an important part of the information about the pose. In fact, observing the results obtained without the consideration of these corners (green line in Fig. 3.7(b)) the imprecision is higher with respect the other cases. One can therefore think to give more confidence to these points to improve the robustness of the estimation. However, an exaggerate confidence on these points' locations may bring to higher errors (such that obtained with a weight of 10 on the tip, traced green line).

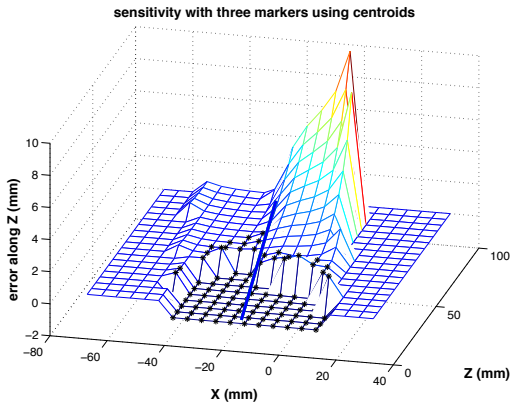
The discussion conducted so far is just a local analysis of sensitivity and it does not allow to discuss if the problem is a ill-posed problem i.e. if two or more configuration have the same visual representation (whatever the selected features are). Also the set of possible features is reduced and other features or composition of such features can be imagined; in the next sections, when talking about "best features" for pose estimation purpose is always referred to these 3 cases.

In light of this discussion, the best solution is taking into account reference points on the apparent borders which have some physical meaning and for whom a 3D relationship can be determined.

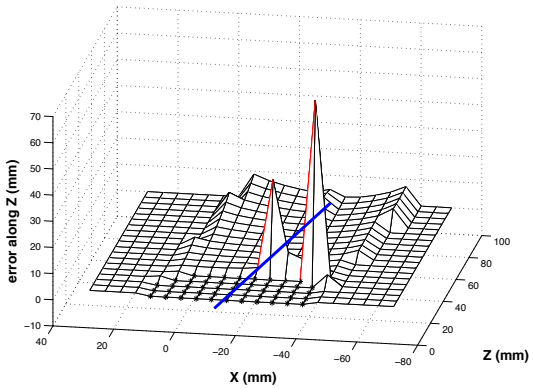
To increase the robustness, several points can be chosen over the border and to relate the image feature points and the 3D object positioning, a ruler-like pattern is painted on the instrument alternating blue and yellow markers (Fig. 3.8). This does not only creates salient object related features, it also provides correspondences between upper and lower border points which can be used as trustful features for the pose estimation (cf chapter 5).

The choice of using markers is quite extended in surgical robotics, where robustness of feature extraction and tracking is really challenging especially *in-vivo* scenarios where the illumination is non-uniform and varying, and where smoke, blood or other fluids may degrade the quality of the image. The choice of the colors allows to augment the discriminance between the instrument and almost any internal tissues and also to get good inter-markers distinction as they are complementary colors and, thus, their chromaticity values occupy opposite positions in most of the color representation spaces.

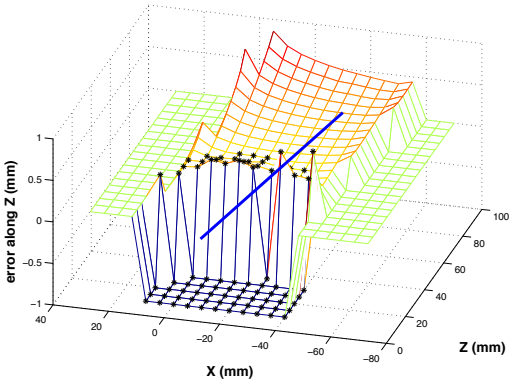
In the next chapter, the algorithm developed to extract the mentioned feature is presented with the studies and theoretical basis over which it is built.



(a) Only 3 centroids

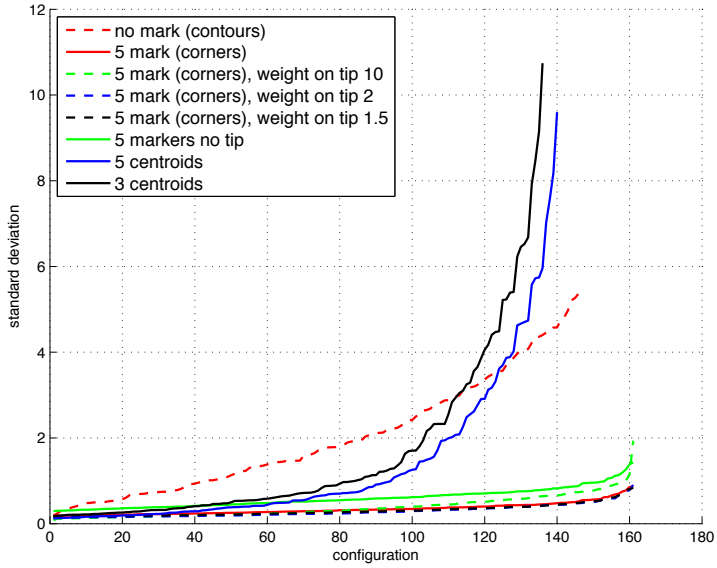


(b) Only apparent borders

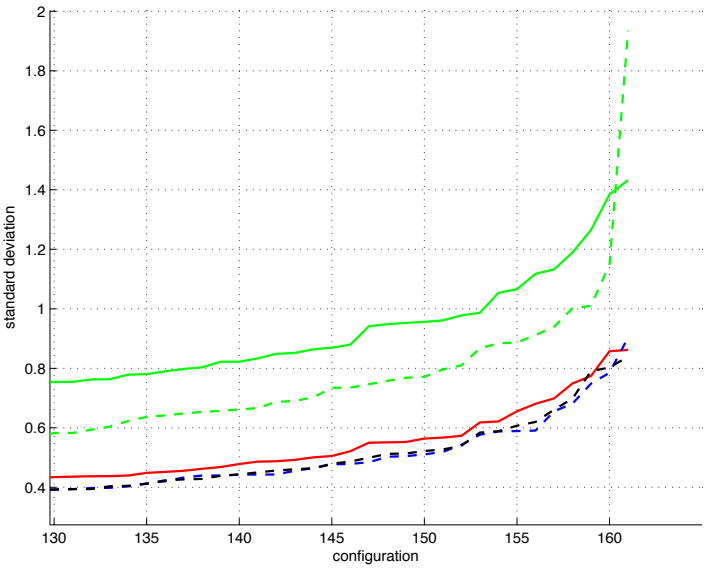


(c) Markers corners

Figure 3.6: Result of the variance propagation in the representative plane parallel to (x, z) of $\mathcal{F}_C$ and passing through the origin of $\mathcal{F}_{ch}$. The common expected tendency is a variance growing with the distance to the camera. The depicted blue line is representing the direction of the housing channel.



(a) sorted worst direction



(b) zoom on worst cases

Figure 3.7: For each robot configuration the direction with the largest variance in the tip position estimation are taken and sorted. The result for different chosen descriptors is shown in (a). In (b), a zoom on the worst cases is done: the best solution seems to be the one considering the corners of some markers (5 markers of equal dimensions in this case).

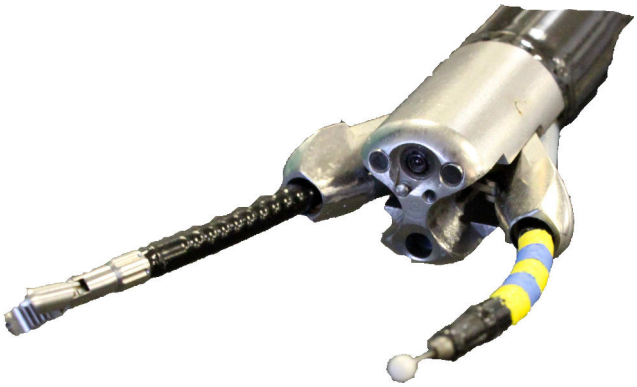


Figure 3.8: Anubis tip with the camera, the flexible surgical instruments with the visual markers.

Chapter 4

Extraction of Features for Pose Estimation

Contents

As evinced from the precedent chapter, adequate features for pose estimation are the corners of colored markers attached to the bendable part of the instrument. As already pointed out, the choice of locating the markers on the bendable part is due to the fact that this is the most visible part during an operation. Indeed, the surgical tool is continuously interacting with the tissues and, therefore, hidden by them, whereas the instrument section located upstream wrt to the bendable section is inside the channel most of the time.

The extraction of the apparent corners of the markers is quite challenging. The conventional approaches (e.g. Harris, maximum curvature of the borders, ...) are not efficient because the object has not sharp vertexes and the images are of low quality and highly distorted. The apparent corners can be defined as the points in correspondence with the color transitions on the upper (or lower) instrument apparent contours.

To exploit this property, we propose to first roughly segment the instrument in the image so as to to define a region of interest where the contours are detected. In order to reject falsely segmented areas and retrieve the whole instrument structure,

an interpretation step is performed (see section 4.2), which label the segmented regions according to the marker they pertain. Finally, a continuous definition of the upper and lower ordered border points must be obtained so as to be able to detect the color transitions with a subpixel accuracy. This process also allows to label the extracted corners according to the markers they pertains, which makes the matching with their 3D counterparts possible.

This chapter describes in detail the method used for the extraction of the wanted features. After a discussion over color segmentation for the adopted markers, the two keyparts of the algorithm are detailed in sections 4.2 and 4.3. Finally, the algorithm implementation is described before showing the results in both controlled and *in-vivo* environment.

4.1 Color model and segmentation

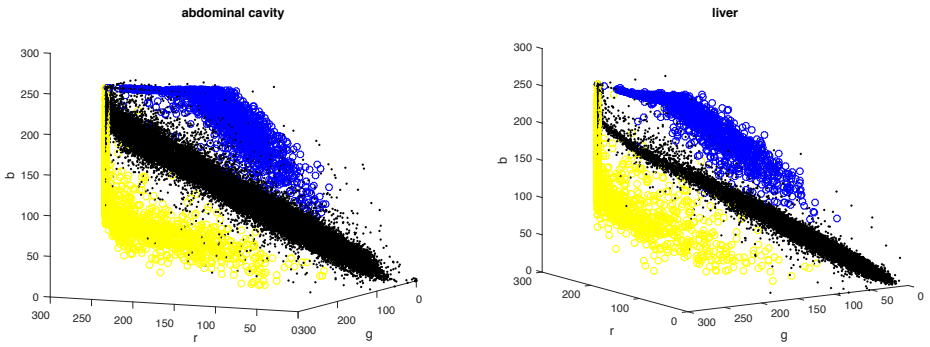
Thanks to the adoption of colored markers, the color itself can be used for the first broad segmentation to retrieve the instrument ROI. In Fig. 4.1 the distribution of yellow, blue and background colors are represented in the RGB (8-bit color representation) space for two common scenarios within the abdominal cavity: intestine and liver. Analyzing these distributions, it can be seen that the chosen colors for the markers are distinguishable from both the backgrounds. Furthermore, blue and yellow distributions show a large inter-distance between them confirming also visually the complementarity of the two chosen colors.

However, the limits between the three categories (yellow, blue and background - depicted in black in the figure) are not always clear: for example, in the zone of the RGB space representing whiter hues (i. e. R, G and B values next to 255). This is may due to the strong illumination which degrades the color perception having almost a whitening action and conferring to the same marker different hues including white (see sec. 4.1.1).

In endoscopy, the camera sensor is usually a mono-CCD with Bayes filter and the acquired image is in RGB format. This color space, though, is not ideal for color-based segmentation since its channels intrinsically contain luminance information which makes color perception (and consequent RGB values) sensitive to light [Doignon 2004]. For retrieving a true color information, the chroma component, which is independent from luminance variation, must be extracted.

In the L^*a^*b color space, the chrominance is separated from lightness information (L) and is represented by two channels a and b (Fig. 4.2): along axis a are represented the red - green colors and along b blue - yellow colors.

This color space provides a pure chrominance representation and, on theory, carries all the needed color information on the same coordinate b. Moreover, blue and yellow are represented on the extrema of this channel and therefore they should be easily distinguishable using this representation making this color space very suitable



(a) (b)

Figure 4.1: The distribution of yellow, blue and background colors are represented in the RGB (8-bit color representation) space for two common scenarios within the abdominal cavity: intestine (a) and liver (b).

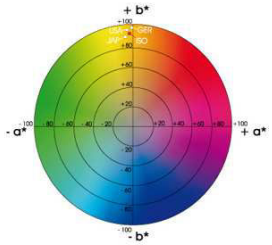


Figure 4.2: Chromatic representation for L*a*b space. Along axis a are represented the red - green colors and along b blue - yellow opponent colors.

for the aforementioned color segmentation.

However, as it can be deducted from Fig. 4.3, the blue paint is not a pure blue and, consequently, only one channel is not enough to represent the entire variability. The color "blue" seems to contain some green components (negative values along the a axis) and this fact obliges the use of the entire bidimensional chrominance space for having the real color representation. Despite the fact that blue is not pure, the two colors still cover two separate areas in the $a * b$ space, which only partially overlap the background gamut in the whiter zone (the same behavior present in the RGB space).

Given the quite large inter-distance between the different distributions, a model for each apparent color content of the markers can be determined, so as to use it for detecting the same color in further video frames. Observing the distribution of the two color clusters, it seems that the density of the point cloud decreases with the distance from the center of the cluster, that can be considered the nominal (or real) paint color. This fact suggests that the probability of one pixel to pertain to one of these two class can be modeled like a Gaussian centered on the nominal value (the barycenter of the point cloud) and whose shape is determined by the variance of the specific color samples. The obtained (blue and yellow) Gaussian Models (GM) (represented as a confidence ellipse in Fig. 4.3) can be used with the pixels $a * b$ values of a new image frame providing, as a result, a probability image where the pixel intensity represent the probability of that pixel of being yellow marker (or blue marker).

According to the dataset used in this work, the centers (c) and the variance matrixes ($\mathbf{C}$) of the fitted Gaussians are the following (respectively for yellow -yw- and blue -bl-):

$$c_{yw} = [-6.3190, 54.7758]^T, \quad \mathbf{C}_{yw} = \begin{bmatrix} 103.3870 & 0 \\ 0 & 246.0974 \end{bmatrix}$$

$$c_{bl} = [-18.4591, -13.6889]^T, \quad \mathbf{C}_{bl} = \begin{bmatrix} 34.3313 & 0 \\ 0 & 52.7258 \end{bmatrix}.$$

In the literature, assigning a probability value instead of a "hard" label is known as "soft" segmentation. In our case, the combination of a "soft" segmentation with a dynamic threshold determination based on image information demonstrated to bring more repeatable and robust results than a conventional "hard" segmentation where a specific label is assigned to each pixel [Chen 2007].

The soft segmentation step, in fact, impedes to have *a priori* exclusion of some pixels that can result to be useful according to a higher-level classification or segmentation criteria. The error due to "hard" decision, instead, cannot be changed and bring difficulties for image analysis.

A hard classifier such as ADABOOST [Schapire 1990] was also experimented but the described drawbacks make it less suitable for such application. This learning

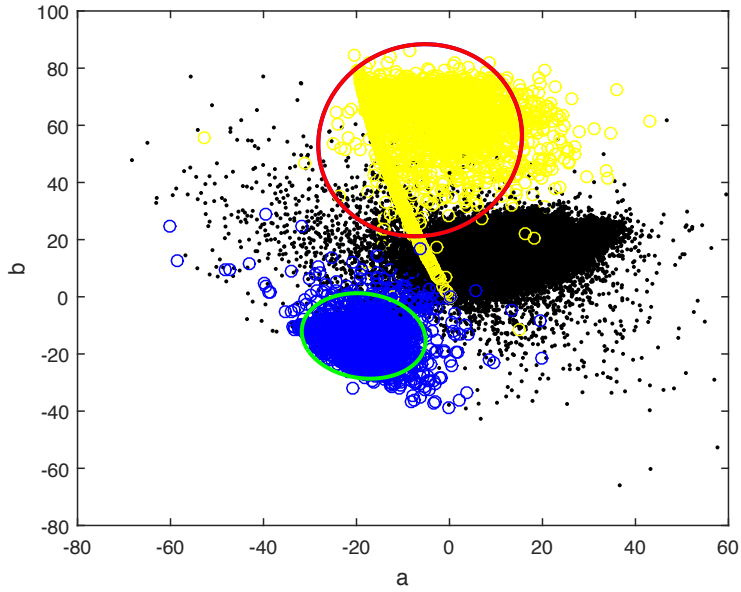


Figure 4.3: Color distribution in $L * a * b$ space. The yellow and blue values are depicted by the same colors whereas black is used for the background. The two ellipses represents the iso-contours of the Gaussian distributions corresponding to a confidence of 90%.

algorithm learns the boundary which best divides (with the greatest clearance) the background from the foreground according solely to the training set. If a case was not "seen " during the training, the classification could fails make it necessary to learn the markers color in different scenarios according to the operation which is carried out.

On the other hand, in a probabilistic approach, if the color is slightly different from the model the response would be weaker but thanks to a dynamic threshold such information can be recovered (and not negatively labeled from the off). However, a minimum threshold should be settled anyway to be able to determine when the instrument is absent.

A color-based segmentation, though, loose its effectiveness if the captured color changes during the video sequence or in cluttered environment where other object can have the same color. In these cases, solutions have been proposed either updating the color model [Raja 1998] or considering the structure of the object itself or border information [Moreno-noguer 2003].

However, the most critical issue for color-based segmentation is, obviously, when the color information is corrupted or totally absent. As already pointed out at the beginning of this section, the strong illumination degrades the color information even causing strong specular reflectance on the body of the instrument removing any color information.

4.1.1 Effect of Specular Reflectance

When the specular component of the reflected light hitting the camera sensor is of high intensity, the sensor saturates and a white region is formed in the image at the portion of surface responsible of that specular reflection.

As described before, the light is emitted by two punctual sources very close to the camera which provides high intensity in the center of the observed scene letting the border regions darker. The material of the instrument is very reflective, but even using matte painting for the markers, the high intensity of the light illuminating the center of the scene corrupts the instrument image showing wide white zones on the colored markers.

However, assuming that the bendable part of the instrument is a torus, its apparent contours correspond to physical points where the normals to the instrument surface are perpendicular to the line of view. Thus, the rays of light hitting that portion of surface cannot be reflected directly to the camera sensor which will collect only the diffuse reflection component for that portion. It can be stated, then, that the white zones deriving from direct reflectance are only present in the "middle" of the imaged instrument when it is exposed to direct high intensity illumination. As a result, each marker is partitioned into, at most, two yellow/blue parts.

This decomposition of the instrument body and the issue of distinguishing the instrument region by other regions with the same colors (false positive) brings to the need of a specific classification stage. It has the role of "understanding" each detected region recognizing which of them really pertain to the instrument and, in that case, to what marker they pertain.

4.2 Regions Interpretation

Looking at the aspect of the instrument and considering only the color, we decided to described the instrument as a multi-part object formed by the *ordered* sequence of the colored markers from the base to the tip. However, due to light saturation, the appearance of these parts may change due to the division of a single marker in multiple regions.

A post-processing stage, then, is needed to classify the real object parts with the right order (based on a topological criterion) and taking into account possible parts variations.

Global optimal solution methods such as Graph-Cuts or hierarchical structure do not fit exactly the problem due the complexity (or impossibility) of defining (and optimize) a global cost function that takes into account order relationship. Furthermore, the shape and the appearance of the instrument is subjected to changes during the operations and, for this additional reason, pictorial structure [Felzenszwalb 2005] based methods cannot be straightforwardly applied to our case and would return high score false positives due to the effect of specularities.

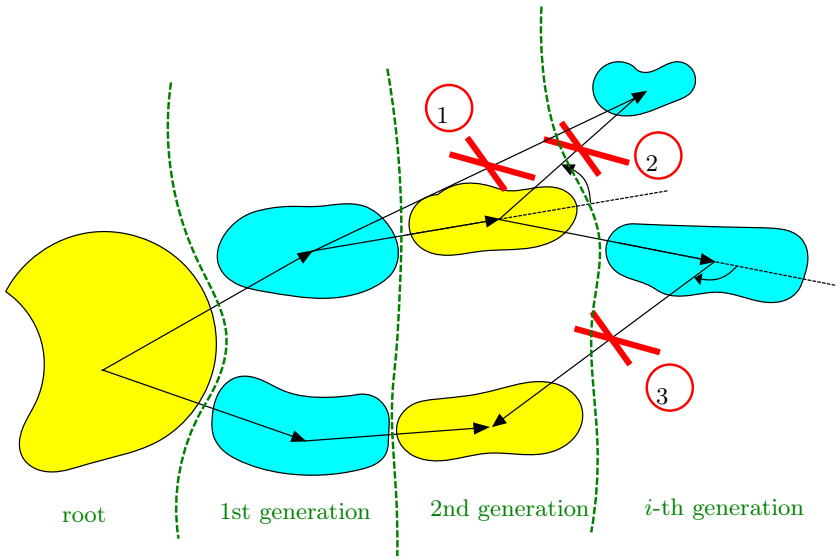


Figure 4.4: Tree building with the specified connectivity. In case 1 the branch is not created since it does not fulfill condition 1, whereas cases 2 and 3 does not respect condition 2 since the orientation of the possible branch strongly differs from the preceding outline. This prevents connecting outliers (case 2) and forming internal loop which correspond to physically unfeasible configurations (case 3).

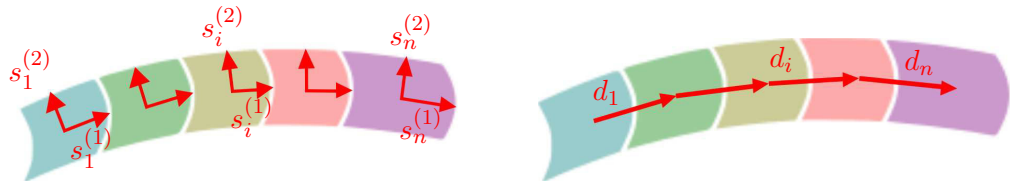
We propose here to simplify the problem and solve it locally, in two steps: firstly, a tree is created according to the regions topology and, secondly, that tree is processed till obtaining a directed path from the base (the root) to the tip marker (the unique leaf).

To build the tree, a neighboring connectivity (4- or 8-connected) is not sufficient to determine the correct relationship between the regions and, consequently, the right order. Therefore, a new connectivity is defined, based on two structural characteristics of the bendable section:

1. The markers are adjacent to one another alternating between blue and yellow
2. The centroids of the different markers in the image should lay on an hyperplane whose tangent at each point should not present strong discontinuities.

More specifically, starting from the root, a new node is connected to its parent only if it satisfies these two properties, i.e. assuring the alternation of colors and avoiding strong discontinuities on the orientation of subsequent branches (Fig. 4.4). Building the tree this way allows to order the regions from the base to the tip and to have, on each generation (cfg. Fig. 4.4), all the candidate regions associated to the same marker.

Moreover, for each generation, the nodes are considered 2 by 2 and a *putative node* is created which represents the region resulting by the possible fusion of the two considered regions, eventually providing the tree in the first line of Fig. 4.6.



(a) (b)

Figure 4.5: The appearance of each marker can be represented by two factors: one shape factor and a topological factor. The first is defined as the length and width of each marker according to its principal axis (left) and the second as the distance from the considered marker and its parent (right).

The final aim, though, is to obtain the directed path composed by the markers of the bendable part of the instrument. For pruning the tree, a time consistency criterion is used, assuming that from one frame to the following, the shape of the markers substantially stays unchanged.

To exploit this consideration, two marker descriptors are proposed to be taken into account:

- a *shape factor* depending on the second order moments of the region (computed along its principal component directions);
- a *topological factor* depending on the distance of the considered node (real or putative) from its parent node (=region).

These two factors are modeled as stochastic variables. Defining a as the event of being a particular marker A, s the event of having a particular shape S (shape factor) and d the event of having a specific distance D from the parent node (topological factor), the shape/topological frame-to-frame similarity can be defined in terms of:

$$p(A|S,D).$$

As a first approximation, we can consider both shape and topological events as Gaussian processes which can be defined as:

$$p(S^{(1)}|A) \sim \frac{1}{2\pi\sigma_{s_a}^{(1)}} exp(-\frac{(s_i-\mu_{s_a}^{(1)})^2}{2[\sigma_{s_a}^{(1)}]^2}) \tag{4.1}$$

$$p(S^{(2)}|A) \sim \frac{1}{2\pi\sigma_{s_a}^{(2)}} exp(-\frac{(s_i-\mu_{s_a}^{(2)})^2}{2[\sigma_{s_a}^{(2)}]^2}) \tag{4.2}$$

for the shape (Fig. 4.5(a)) and

$$p(D|A) \sim \frac{1}{2\pi\sigma_{d_a}} exp(-\frac{(d_i-\mu_{d_a})^2}{2\sigma_{d_a}^2}) \tag{4.3}$$

for the topology (Fig. 4.5(a)), where μ and σ are respectively the mean and the standard deviation of the observed values of, respectively:

- the euclidean distance from the father region centroid $(\mu_{d_a}, \sigma_{d_a})$
- the second order central moment computed along the principal components directions of the specific marker $(\mu_{s_a}^{(1)}, \sigma_{s_a}^{(1)})$ and $(\mu_{s_a}^{(2)}, \sigma_{s_a}^{(2)})$.

These quantities are specific for each marker from 1 to 5. To take into account the variation of size and shape during instrument movement, these parameters are updated at each frame considered the observations on the last 10 frames.

Except for the root, the joint probability value of each region (putative or real) of being marker given the shape and topological observation can be considered as the reward associated to each branch of the tree:

$$p(A|S,D) \sim \frac{1}{2}(p(S^{(1)}|A) + p(S^{(2)}|A))\, p(D|A). \tag{4.4}$$

The tree pruning, then, can be solved generation by generation (see Fig. 4.6) by selecting the branch with the highest probability. If the chosen branch is associated to a putative node, then, the two involved nodes are deleted and substituted by the putative node where the centroid and shape factor are computed for the region resulting from the merging of the two involved regions. Subsequently, the tree structure and the branches values are updated.

If, in one generation, there are more than 2 candidates then the process is iterated till no better (i.e. with higher probability) solution can be found.

At the end of this process, the tree will be the sequence of the 5 centroids coordinates ordered from the base to the tip of the bendable section (see second step of Fig. 4.7 with the corresponding labeled image).

4.3 Bézier Curve Representation of the Borders

Once the ROI and the skeleton (composed connecting the marker centroids) have been determined, the idea is to exploit that information to retrieve the superior and inferior apparent borders of the instrument so as to look for the color transition along them (see Sec. 4.4).

The conventional edge detection methods usually return a cloud of points which usually must be post-processed to eliminate outliers and select only the wanted edges.

Parametric curve fitting can be used to smooth contours extracted from local maximum gradient extraction, thus providing sub-pixel precision, and helping rejecting outliers. According to several test on synthetic data, we decided to use quadratic Bézier curves which resulted to be sufficient to represent the apparent borders of the projection of a torus, at least from the particular point of view of the considered endoscopic system. Therefore, we propose here a complete framework for robustly fitting Bézier curve to bidimensional data using an M-estimator.

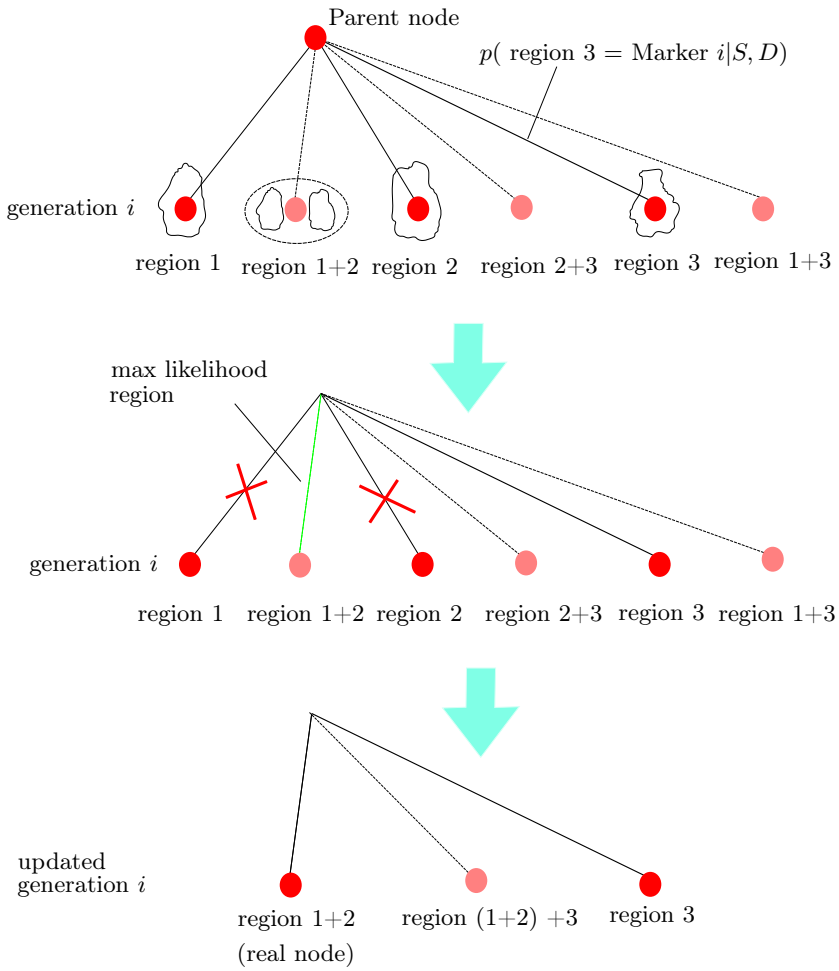


Figure 4.6: Scheme for the tree updating process. At each generation, the branch to whom the maximum likelihood is associated is selected (highlighted in green). If, as in the represented case, the most likely branch is that of a putative node (i.e. the node representing a fusion of two regions. The nodes of the corresponding two regions are deleted and the putative node turn into a real node. If the generation presents more than two regions, the new nodes and branch values are computed and the process is repeated on the updated tree (third line)

A Bézier curve of degree n is a parametric curve defined as follows:

$$B(t) = \sum_{i=0}^{n} b_{i,n}\beta_{\mathbf{i}} \tag{4.5}$$

where the polynomials

$$b_{i,n} = \begin{pmatrix} n \\ i \end{pmatrix} t^i(1-t)^{n-i},\ i = 0, 1, \cdot, n \tag{4.6}$$

are the Bernstein basis polynomials of degree n. The points $\beta_{\mathbf{i}}$ are the so called control points of the curve. The polygon formed by connecting the control points from $\beta_{\mathbf{0}}$ to $\beta_{\mathbf{n}}$ is called the control polygon and the resulting curve is contained in the convex hull convex hull of the control polygon.

The best fitting Bézier curve can be defined in several manners depending on the objective function. The objective function must be chosen properly and, especially in computer vision problems, with particular attention to the robustness it confers to the estimation. Different measures of robustness are proposed in the literature, but the most common are the *breakdown point* and the *influence function*. The former represent the minimum fraction of outlying data that can cause an estimate to diverge arbitrarily far from the true estimate. The latter is the change in an estimate caused by the insertion of outlying data as a function of the distance of the data from the (not corrupted) estimate.

The ideal situation for a robust fitting is having a breakdown point of 0.5 and an influence function which tends to zero with increasing distance. The breakdown point cannot achieve higher values because, when more than the half of the data are "outliers", it is impossible to determine which ones are the true data to fit.

The commonly used estimators become, then, insufficient to get a robust fitting, according to these two indicators. Indeed, in the case of least mean square (LMS), a single bad point can bring the solution far from the real fitting (i.e. breakdown point equal to 0) and the influence of a point is proportional with its distance. An alternative (not ideal though) is using least absolute deviation as fitting quality function which presents a bounded influence function (it is the derivative of a piece-wise linear function).

A robust solution in computer vision is represented by *M-Estimators*, which are a generalization of Maximum Likelihood Expectation and Least Squares. We propose, here, a possible formalization for computing the "best" fitting Bézier curve using M-estimators.

4.3.1 M-estimator for Bézier Curve Fitting

Let $X = \{x_i\}$ be a set of data points and let $\mathbf{a}$ a k-dimensional vector of the parameters to estimate. The M-estimate of $\mathbf{a}$ is

$$\hat{\mathbf{a}} = \arg\min_{\beta} \sum_{x_i \in X} \rho(r_{i,a}/\sigma_i) \tag{4.7}$$

where $r_{i,\mathbf{a}}$ is the error distance or residual function relative to x_i and $\rho(q)$ is a robust loss function that grows subquadratically and is monotonically nondecreasing with increasing $|q|$. In addition, σ_i^2 is the variance (scale) of the scalar value $r_{i,\mathbf{a}}$.

The solution in (4.7) is solved by finding $\mathbf{a}$ such that

$$\sum_{x_i \in X} \psi(r_{i,\mathbf{a}}/\sigma_i) \frac{dr_{i,\mathbf{a}}}{d\mathbf{a}} \frac{1}{\sigma_i} = 0 \tag{4.8}$$

where $\psi(q)$ is the derivative of $\rho(q)$. The same problem can be translated in solving

$$\sum_{x_i \in X} w(r_{i,\mathbf{a}}/\sigma_i) \frac{dr_{i,\mathbf{a}}}{d\mathbf{a}} \frac{1}{\sigma_i} r_{i,\mathbf{a}} = 0 \tag{4.9}$$

where w is a particular weighting function such that $w(q)q = \psi(q)$. This leads to a process known as "Iterative Re-weighted Least Squares" (IRLS), which alternates steps of computing the weights $w_i = w(r_{i,\beta}, /\sigma_i)$ using the current estimate of the parameters and solving (4.9) to compute a new $\mathbf{a}$ with given weights.

Following the M-estimator approach, an objective function can be conceived to fit the Bézier curve on the extracted borders:

$$\sum_{i_b=0}^{n_b} \rho(B_x(t_{b_i}, \beta_\mathbf{x}) - p_{b_i}^{(x)}) + \sum_{i_b=0}^{n_b} \rho(B_y(t_{b_i}, \beta_\mathbf{y}) - p_{b_i}^{(y)}) \tag{4.10}$$

where β_x and β_y are the vector of the x and y coordinates of the control points:

$$\beta_x = [\beta_0^{(x)}, \beta_1^{(x)}, \beta_2^{(x)}]$$

$$\beta_y = [\beta_0^{(y)}, \beta_1^{(y)}, \beta_2^{(y)}]$$

and $\mathbf{p_{b_i}} = (p_{b_i}^{(x)}, p_{b_i}^{(x)})^T$ are the superior (inferior) detected border points, and t_b is the parameter of the Bézier curve such that $B(t_{b_i}, \beta)$ is equal to p_{b_i} for some β that must be found. Comparing (4.10) with (4.7), the vector of parameters $\mathbf{a}$ is represented by the coordinates of the Bézier control points $\mathbf{a} = [\beta_{0_x} \beta_{1_x} \beta_{2_x} \beta_{0_y} \beta_{1_y} \beta_{2_y}]^T$ and the residuals are expressed as the distances along each coordinates of corresponding points $r_{i,\mathbf{a}} = [B_x(t_{i_b}, \beta_\mathbf{x}) - p_{b_i}^{(x)}, B_y(t_{i_b}, \beta_\mathbf{y}) - p_{b_i}^{(y)}]^T$.

As previously stated, this could be solved by means of the IRLS algorithm but thanks to the adopted formalism with Bézier curves, a closed form solution can be found for (4.9) once the weights are known.

A matrix form solution is presented here to find the best fit according to the M-estimator framework.

Rewriting the Bézier curve (4.5) as

$$\beta_0 + t(-2\beta_0 + 2\beta_1) + t^2(\beta_0 - 2\beta_1 + \beta_2)$$

it can be expressed, for each coordinate (x or y), in a matrix form as:

$$B_x(t) = T\, C\, \beta_\mathbf{x}$$

where C is the matrix of the Bézier coefficients:

$$C = \begin{bmatrix} 1 & -2 & 1 \\ -2 & 2 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

and each line i of T is composed by $[t_{b_i}^2 \, t_{b_i} 1]$ corresponding to each detected point i.

For a quadratic Bézier curve ($n = 2$), the lines defining the control polygon are the tangents to the Bézier curve in the first and last control points. As the Bézier curve must fit the border points, it is plausible to state that ideally the control polygon must be tangent to the apparent borders at its terminal points. This property indirectly suggests that the first and last control point must lie on the terminal parts of the apparent superior and inferior contour of the object.

This characteristic allows to define the different t_b as the curvilinear coordinate over the curve connecting all the points in the given order. Its value will be 0 for the first (superior/inferior) border point and 1 for the last. A proper way to compute these values is assigning:

$$t_{i_b} = \frac{|d_i - d_{i-1}|}{\sum_{i=1}^{n_b} |d_i - d_{i-1}|}$$

where

$$\begin{cases} d_1 &= 0 \\ d_i &= \sum_{j=2}^{i} \left\| \mathbf{p_b}_j - \mathbf{p_b}_{j-1} \right\|_2 \end{cases}$$

Thus, at each iteration and for each coordinate x and y, the β solving the least square problem is defined as:

$$\beta_x^{(j+1)} = \arg\min_{\beta_x} \sum_{i=1}^{n} w^{(j)} \left| p_{b_i}^{(x)} - T_i C \beta_{\mathbf{x}} \right|^2 = \left((TC)^T W^{(j)} (TC) \right)^{-1} (TC)^T W^{(j)} \mathbf{p_b^{(x)}}$$

where $\mathbf{p_b^{(x)}}$ is the vector containing all the x-coordinates of the detected border points and W is a diagonal matrix gathering the weights computed at step j for each border point. The problem has a closed form solution and the new weights can be computed after each IRLS iteration.

Considering that the whole $\mathbf{a} = [\beta_{0_x} \beta_{1_x} \beta_{2_x} \beta_{0_y} \beta_{1_y} \beta_{2_y}]^T$ is taken into account and permitting an over-definition of the variables, the whole problem can be defined (and solved) in a matrix way by composing the matrices defined so far. Thus, new T_i is $[t_{i_b}^2 \, t_{i_b} 1 \, 0 \, 0 \, 0]$ for residual along x-coordinate and $T_i = [0 \, 0 \, 0 \, t_{i_b}^2 \, t_{i_b} 1]$ for ones along y; the matrix of coefficient

$$C = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ -2 & 2 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & -2 & 2 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

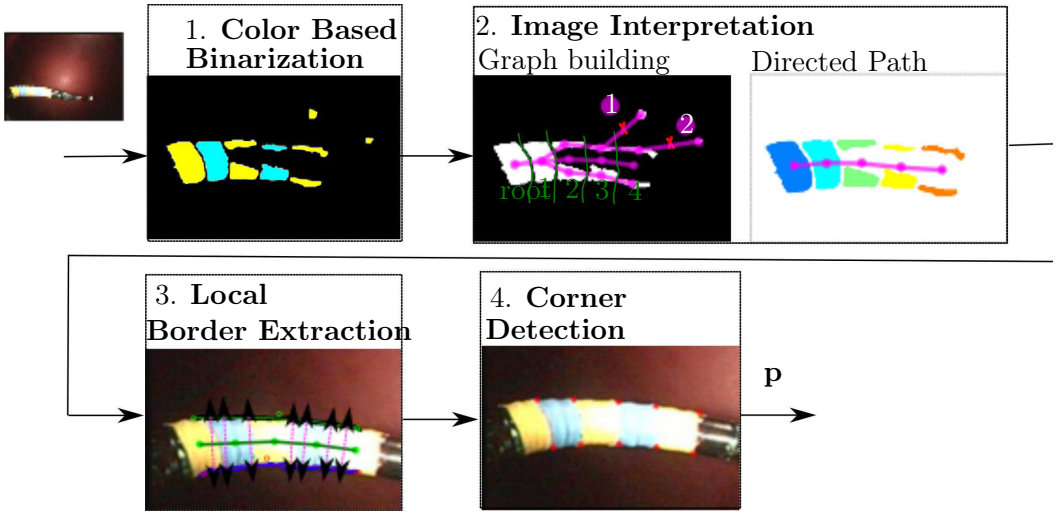


Figure 4.7: Steps of the feature extraction. (1) GM are used to select blue and yellow regions. (2) Creation of the tree using the specified connectivity: Branch 2 is not created since it does not fulfill condition 1, whereas branch 1 does not respect condition 2 because the orientation of this branch strongly differs from the preceding outline. The tree is finally processed to get a directed path. (3) The skeleton is used for searching border points. These are fitted with Bézier curves along which color discontinuities are searched and considered as corners (4).

At each iteration, then, the searched β is

$$\beta^{(j+1)} = ((TC)^T W^{(j)} (TC))^{-1} (TC)^T W^{(j)} \mathbf{a}.$$

Thanks to this formalization and solution by M-estimator, a continuous representation of the border can be computed which is able to filter the outliers due to image noise of miss-detection (see Sec 4.5).

4.4 Overview of the whole process

4.4.1 First Stage: Candidate Regions

For any new frame the L^*a^*b representation of each pixel is retrieved and its probability of being either yellow or blue is computed according to the built Gaussian Models.

Subsequently, the image is labelled (yellow / blue / background) according to two thresholds (one for blue and one for yellow) iteratively computed to assure a reasonable number of candidate regions according to the considerations on light saturation and markers division: at most 8 yellow and 6 blue regions. The initial thresholds are increased (or decreased) if the number of binarized regions is higher (or lower) of the admitted quantity.

Figure 4.8: The green line represent the estimated instrument entering direction. The first yellow marker encountered walking along this line is considered to be the root (i.e. the first marker) of the tree.

A further control is performed to check if each connected region actually pertains to a single marker. The convex hull of each binary region must contain, in prevalence, one color only (it may contain background or neighbor marker little portion). If it is not the case it means that two different markers resulted in only one region which must be divided by an erosion.

4.4.2 Second Stage: Image Interpretation

For building the tree, the root node must be firstly detected. Actually, it can be defined as the first yellow marker encountered walking through the projection in the image of the instrument tube which precedes the bendable part (cfr. Fig. 4.8). Indeed, in STRAS the axis of the passive part of the instruments wrt the camera is constrained by the mechanical structure and its projection in the image can be coarsely estimated. This defines the entering direction of the instrument in the image which may vary during the operation for example when the tip of the instrument is pushing and organ. To be robust to these possible entering direction variations, the root position at step i is used to define the direction along which searching the root at step $i + 1$.

Moreover, the marker associated to the base of the bendable section (the root of the tree) is generally well detected because it lies in a part of the workspace that does not receive much direct illumination.

The tree is then built and processed till obtaining the sequence of the 5 centroids coordinates ordered from the base to the tip of the bendable section.

4.4.3 Third Stage: Apparent Borders extraction

At this stage, a local approach is adopted to extract the upper and lower apparent borders of the instrument. The skeleton obtained by connecting the ordered centroids is evenly sampled and gradient magnitude local maxima are searched along the normal to the skeleton. The candidate contours points are selected according to

their gradient direction (along the normal to the skeleton) and their position (near the boundaries of the labelled image obtained in the previous step).

More in details, for each ordered couple $(\mathbf{c_i}, \mathbf{c_{i+1}})$ of subsequent centroids, the segment connecting them is uniformly sampled and for each sampled point, a line $g_{i,j}$ is built, which passes through this point and which is perpendicular to $\mathbf{c_{i+1}} - \mathbf{c_i}$ (see step three in Fig. 4.7). The instrument edges can then be defined as composed by the points satisfying the following three properties:

- Being a local maxima in the gradient magnitude (computed along x and y of each RGB channel) along $g_{i,j}$.
- The orientation of the gradient of the candidate point should be close to the orientation of $g_{i,j}$ relative orientation between the phase of the gradient of the candidate point and the direction of $g_{i,j}$ should be within a little interval around zero.
- The candidate point should be contained within the crown defined around the final binary image obtained in the previous step. Such crown can be obtained by subtracting the binary erosion of the image obtained in step two (Fig. 4.7) to its dilation: this results in a thick crown where most probably the borders are.

Along each $g_{i,j}$ the outline of the gradient is analysed and the local maxima are recorded both in upper and lower directions.

Once collected all the upper and lower border candidate points, only those respecting the three aforementioned properties are kept. If some ambiguity is still present, the furthest pixel from the $\mathbf{c_{i+1}} - \mathbf{c_i}$ segment is chosen. In fact, when walking along g_j from the centroid to the outer zone, two peaks in the gradient can be encountered if the marker in question is affected by saturation: from white (saturated) central zone to yellow (blue) marker and from yellow (blue) to the background color. But only the external transition is the one describing the apparent border.

4.4.3.1 Bézier Fitting

In fitting the borders, the outliers have small residual magnitude, therefore a strong rejection of outliers has been adopted. This can be achieved with the so-called "hard-descenders" whose loss functions quickly tend to zero outside a certain threshold. One example is the Beaton and Tukey loss function [Beaton 1974] (Fig. 4.9) but, unfortunately, this usually means that the objective function is non convex, implying that the IRLS method can converge to a local minimum. Furthermore, using Beaton and Tukey function still implies a breakdown of 0 due to possible *leverage points* [Stewart 1999]: outliers positioned far from the remainder of the data in the independent variables as well as far from the fit to the uncorrupted points. However, thanks to a good initialization and, especially, to the formalization with Bézier

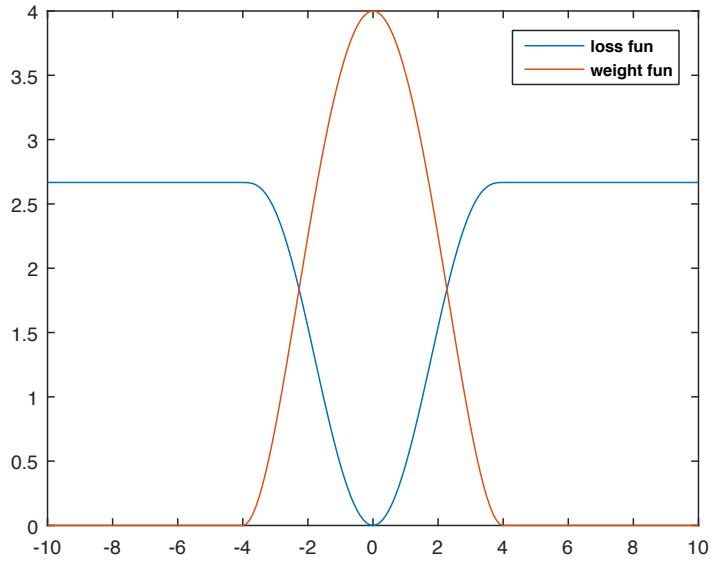


Figure 4.9: In blue the loss function known as Beaton and Tukey and in red the corresponding weight function.

curves, non-sense local minima can be avoided and higher breakdown values can be achieved. In fact, the order of the extracted edge points and the topological information (where the border begin and where it ends) give further information that have been integrated in the optimization thanks to how the t parameter of Bézier curve is computed.

According to Beaton and Tukey loss function, the updated weights can be calculated as:

$$w^{(j+1)} = \begin{cases} \left[1 - \left(\frac{r/\sigma}{c}\right)^2\right]^2, & |r| \leq c \\ 0, & |r| > c \end{cases}$$

If the scale σ of the residuals is not known *a priori*, it can be estimated (and, if needed, re-estimated after some iterations) in several manners. The most adopted and less sensitive to noise is taking the median of the residuals. Optimal c must be chosen according to the application (even though suggested values exist). Here the value $c = 1.5$ pixels seems to be the most appropriate.

4.4.4 Fourth Stage: Corner Localization

To finally extract the markers corners, the color transitions along the Bézier curves should be found. Since blue and yellow are complementary colors, the boundaries between two consecutive markers appears on the image as a gradual transition from yellow to blue (or vice-versa), provided the absence of overexposure (where color information disappear).

As a first approach, one may think to use the probability images resulted by the application of GM and search the local minima along the mentioned curves. Although, the pixels pertaining to the intermarkers frontiers does not fit any of the GM and, therefore, the resulting probability image present wide low value regions in correspondence of these frontiers. Thus, searching the local minimum on this wide flat valley would have no sense and surely lead to corners mislocation.

For this reason, in this step, we consider again the $L*a*b$ image. As yellow and blue are defined in opposite position along the b axis of the $L*a*b$ color representation, the same gradual color transition between two markers should be observable in the b channel of the image. In other words, in correspondence of yellow marker there would be a high b value which smoothly decreases in the blue zone. Considering the origin of the b axis as the frontier between the yellowish and bluish colors, the intermarker frontier points (i.e. the marker corners) can be defined as the zero-cross points of b values along the superior or inferior borders.

To help the detection of such point avoiding possible outliers, the binary masks of the yellow and blue markers (Fig. 4.10(c)) are used to limit the region where to search the color transition points (Fig. 4.10(d)). Finally, the outline of the absolute value of the b signal along the border (blue line in Fig. 4.10(d)) is retrieved (Fig. 4.10(e)) and the global minimum is computed (which is equivalent to the zero-cross point) leading to the final result in Fig. 4.10(f).

If the interpolated line does not overlay the transition (for example when a more external border is detected) the corners are retrieved in another way. By realizing a weighted interpolation between the two Bézier borders, one obtains a set of similar curves inside the instrument image which cross all the markers (Fig. 4.11(a)). The previously defined blue-yellow transitions are then searched on these lines (Fig. 4.11(c)), hence providing a sampled version of these transition limits (Fig. 4.11(d)). An average tangent direction of the upper-most (resp. lower-most) points of each blue-yellow limit is then computed to estimate the direction of the limit near the borders (green line in Fig. 4.11(d)). The intersection with the upper (resp. lower) border then gives the upper (resp. lower) corner point for each transition.

For the corners corresponding to the tip, the same procedure described in the third stage is employed. Since no skeleton information is available in the last segment, the $g_{5,j}$ are built using the same inter-distance and direction of $g_{4,j}$. This is repeated till detecting the absence of yellow information. The last point of the superior and inferior border points are taken as the final corners.

4.5 Results and Comments

Using Bézier curves has, in fact, a triple purpose: obtaining a continuous representation of the borders with consequent sub-pixel precision, filtering the outliers of the local gradient-based border detection and complete the possible missing infor-

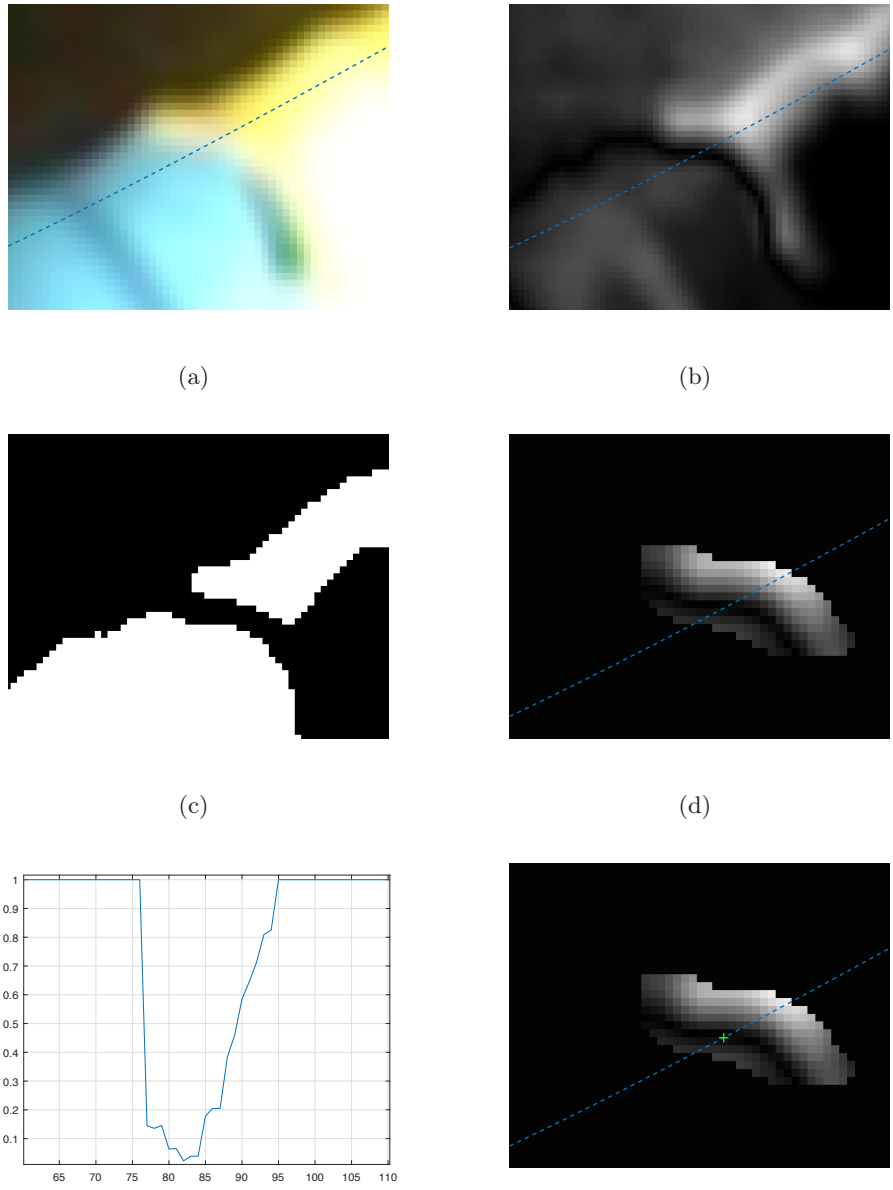
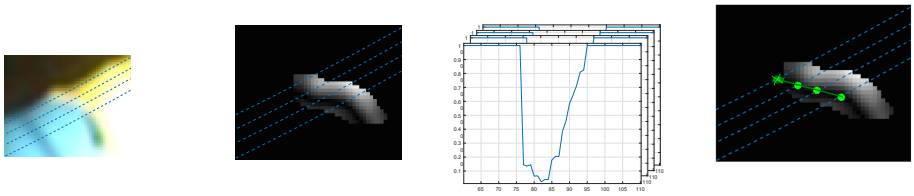


Figure 4.10: Detection of transition points for a particular region in (a). In (b) the absolute value of the b channel of this region is shown. The space between two consecutive segmented markers (shown in (a)) is used to limit the searching zone of the corner as in (d). The chrominance values on the marked profile are shown in (e) where the global minimum is taken as the limit between the two markers obtaining the result in (f).



(a) (b) (c) (d)

Figure 4.11: Detection of transition points for a specfic region when a more external upper boarder is detected. In (b) other Bézier curves are computed interpolating the upper and lower borders and the blue/yellow profile along them are computed in (c). The intersections between yellow and blue profile are computed over these lines. An average direction is computed (green line in (d)) and the intersection of this straight line with the outer border is considered to be the corner location.

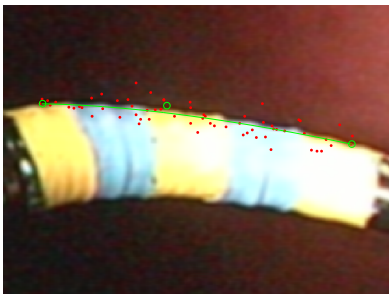
mation. These last two properties are not that noticeable in controlled environment but they become more useful in *in-vivo* environment. To study these capabilities, two artificial cases are artificially generated: the first where 60% randomly chosen detected border points were randomly misplaced within a range of $\sqrt{15^2 + 15^2}$ pixels and the other where some randomly chosen misplaced border points (60% of them) were deleted. The results are shown respectively in Fig. 4.12(a) and Fig. 4.12(b).

The analysis of the breakdown point shows, somehow, the plausibility of the constraints on the control points and the effectiveness of the whole algorithm. Even though a strong noise (50 pixels) is added to the 30% of the detected points, the IRLS solution (cyan curve in Fig. 4.12(c)) is not affected. This is thanks to the synergy between the outliers rejection capacity of the Beaton and Tukey loss function with an added robustness (in terms of breakdown point) thanks to the Bézier formalization and implicit constraint of considering the first and last control points to be near to the first and last detected border points.

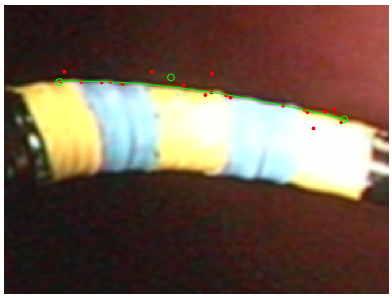
The same figure also shows the comparison of this method with a classical least mean square solution highlighting the good quality of the proposed solution even with highly corrupted information.

However, this last constraint is the bottom line of such formalization. When the initial and ending points of the detected borders are strongly misplaced, the algorithm needs the other border information to be correct and dense to succeed (Fig.4.13(a) and (b)).

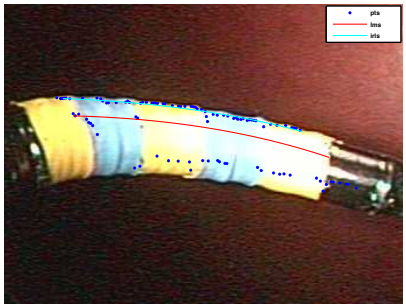
The same process has been used in *in-vivo* cases showing good results in free space without occlusion and demonstrate to be quite robust to strong specularity effects (Fig. 4.14) and mirroring effect on the organs (Fig. 4.14(c) and 4.14(d)). As it does not take into consideration any template of the object nor multiframe analysis, the evident limit of this approach is occlusion (Fig. 4.14(f) and 4.14(e)) which should be taken into consideration in future work (see 7.2).



(a)

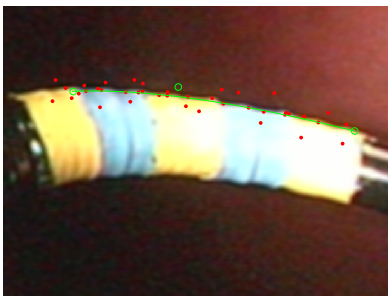


(b)

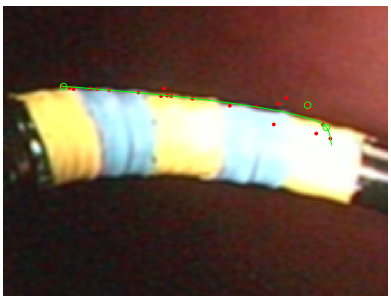


(c)

Figure 4.12: Bézier curves efficiency in presence of outliers or non dense information. In (a) a random error was added to 60% of the detected border points (showed in red). In (b) only 40% of the points (with randomly added noise) are considered in the Bézier curve fitting. The green line is the optimal curve and the green rings are the corresponding control points. In (c) the IRLS and conventional LMS are compared when strong outliers are present in the border detection.

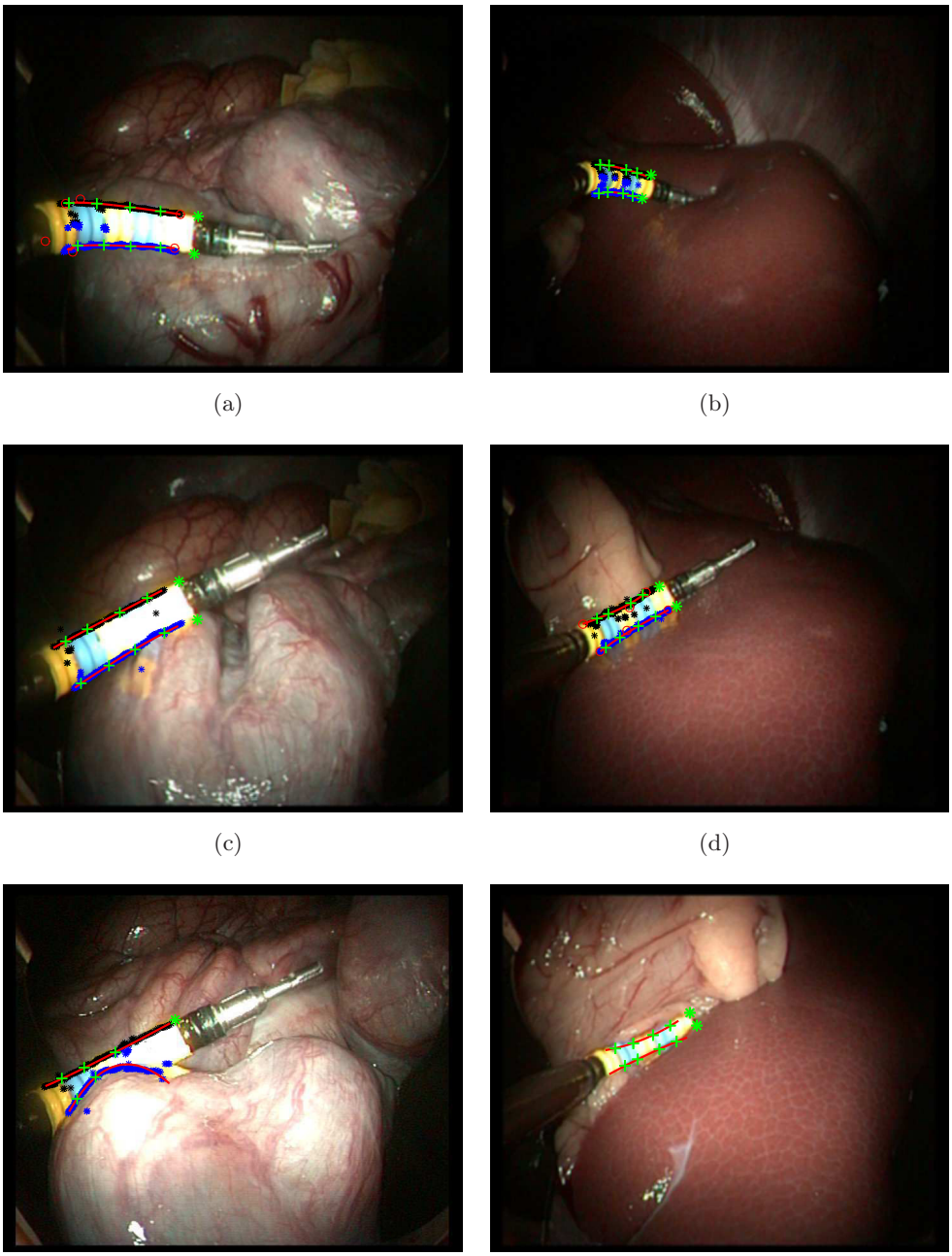


(a)



(b)

Figure 4.13: Bézier curve seems to be more sensitive to noise if all the initial and ending points of the apparent border are strongly misplaced (a). A higher density (even though with noisy points) is requested to mitigate this effect and have a good representation (b) .



(a) (b)

(c) (d)

(e) (f)

Figure 4.14: In-vivo results of the presented method. The green dot is the reprojection of the center of the grasper. In (a), despite the strong specularities the instrument structure is recovered. In (c) and (d), thanks to the local border extraction and the Beaton Tukey loss function the outliers are eliminated and the real instrument is correctly discriminated from its reflection on the organ. In (e) and (f) the segmentation process partly fails for the lower border because the instrument is hidden.

Part II

3D Pose Estimation of the Flexible Instrument

Introduction

Once chosen the suitable features, two opposite approaches have been studied to determine strengths and weaknesses of each of them in the purpose of pose estimation.

Firstly, a *model-based* approach is described which considers some *a priori* information about the mechanical structure of the system and the shape of the instrument to compute the extrinsic parameters of the endoscopic camera and, consequently, the entire pose of the instrument.

Secondly, a *learning based* method is presented which considers the image to pose system like a black box where the relationship between two quantities can be learned provided that they are measurable and a set of this input-output measurement is available.

Chapter 5

Model-Based Pose Estimation

Contents

For the model-based approaches, the model of the object geometry is supposed to be known and the features of interest are usually geometric (vertexes, edges, distances between vertexes, ...). A complementary problem, then, is to find the correspondence between the 3-D model features with all or part of the corresponding 2-D image features. This process is referred to as interpretation or correspondence problem [Doignon 2007]. Once this correspondence has been solved, the best pose is defined in term of a chosen similarity criterion between the projected 3-D features and the 2-D features.

Many works have been presented on this topic with simple shape rigid objects [Doignon 2007]. Fewer solutions, instead, have been proposed regarding articulated [Hel-Or 1994] and shape varying objects [Croom 2010, Padoy 2012] even though, nowadays, this topic has acquired a superior interest also because of the increasing use of flexible tools in many applications.

In our work, the pose estimation problem is seen as the dual problem of an image based visual servoing [Marchand 2002]. Assuming that the kinematic and camera models are known, for a given instrument configuration, a synthetic (virtual) image of the flexible instrument can be computed. The aim, then, will be to find the pose that minimizes the visual measurement error between the virtual and actual feature points. This has not a closed-form solution but can be solved iteratively by moving

the virtual camera towards the real one along the (locally) best direction given by the minimization process.

In part I the correspondence problem has already been addressed and in this chapter the attention will be seeking a solution as an optimization problem.

As a first hypothesis, the geometrical mechanical model of the system is considered to be known (Sec. 5.1). However, since the system is designed with a play between the channel and the instrument so as to allow the instrument to smoothly slide inside it, a second model is used so as to improve the pose accuracy (in sec. 5.2). This extended model takes into account the described variability on the parameters defining the instrument position and orientation.

With a single view, the accuracy of the extrinsic parameters estimates relies on the ability to extract the perspective effect in a reliable fashion from the imaged 3-D objects. Sometimes, the noise or the partial view of the instrument (and consequently partial visual information) can affect the correctness of the estimation because the visual information is insufficient or ambiguous. To overcome such issues, multi-frame continuity and coherence with the motor data can be taken into account (sec. 5.4).

The approach is firstly validated with synthetic data and then tested on our robotic experimental cell and on a manual version in *in-vivo* scenario.

For the rest of this chapter, the following notation is used: $\mathbf{P_i}$ indicates any 3D point coordinates of the i-th section in the camera frame, $\mathbf{P_{i,u}}$ and $\mathbf{P_{i,l}}$ are the 3D upper and lower points of the i-th section, where $i = 1$ is the base and $i = 6$ is the end of the bending part. Furthermore, $\mathbf{P} = [\mathbf{P_{1,u}} \cdots \mathbf{P_{6,u}}, \mathbf{P_{1,l}} \cdots \mathbf{P_{6,l}}]^T$, $\mathbf{P_u} = [\mathbf{P_{1,u}} \cdots \mathbf{P_{6,u}}]$ and $\mathbf{P_l} = [\mathbf{P_{1,l}} \cdots \mathbf{P_{6,l}}]^T$. The same notation but with a little $\mathbf{p}$ will be used to express the coordinates in pixels of the same points projected in the image. The hat symbol $(\hat{\cdot})$ is used for the estimated values.

5.1 Fixed Mechanical Model Parameters

As announced, a first solution is to consider that the mechanical structure parameters are totally known and that the only movements of the instrument are due to its 3 DOFs (λ, ϕ, θ). This approach is inspired of that of [Reilink 2012], where they adopt a visual measurement error based on the centroid of 4 markers attached to the instrument (3 on the bendable part and 1 on the instrument).

Translating this very approach to our case (no markers on the tip and corner related features), the problem would be to find:

$$\mathbf{r} = \arg\min \chi^2$$

where the considered cost function χ^2 can be written, for N image points, as :

$$\chi^2 = \frac{1}{2} \sum_{i=1}^{N} w_i \left[\mathbf{p_i} - \hat{\mathbf{p}}_\mathbf{i}(\hat{\mathbf{r}})\right]^2 . \tag{5.1}$$

For each feature point, the following well known relation can be written:

$$\dot{\mathbf{p}}_\mathbf{i}(\mathbf{r}) = \mathbf{L}_\mathbf{i}(\mathbf{r})\dot{\mathbf{r}}, \tag{5.2}$$

where $\mathbf{L_i}$ is the extended image Jacobian as defined in the precedent chapter. Once stacked all the $\mathbf{L_i}$ into what can be called the *global extended jacobian* matrix $\mathbf{L}$, the optimization problem can be solved iteratively by Gauss-Newton approach as proposed in [Reilink 2012]. From given initial conditions, the correction to apply to $\mathbf{r}$ at iteration n to decrease the visual measurement error is calculated as follow:

$$\delta\mathbf{r} = c\,\mathbf{L}_W^\dagger\mathbf{e} = c\,(\hat{\mathbf{L}}^T\mathbf{W}^T\mathbf{W}\hat{\mathbf{L}})^{-1}\mathbf{W}^T\mathbf{W}\mathbf{e} \tag{5.3}$$

where $\mathbf{e}$ is the visual measurement error (a column vector containing the 2D errors $\mathbf{p_i} - \hat{\mathbf{p}}_\mathbf{i}(\hat{\mathbf{r}})$), $\mathbf{L_W}^\dagger$ is the weighted Moore-Penrose pseudo inverse of an approximation of the global extended jacobian matrix ($\hat{\mathbf{L}}$), c is a fix gain and $\mathbf{W}$ is a weighting matrix. The interest of having a weight associated to the measurement error is the possibility to weight the confidence associated to some feature wrt to others. In [Reilink 2012], they empirically determined that the information associated to the tip worth higher confidence (somehow confirmed by our preliminary sensitivity study) and, therefore, $\mathbf{W}$ is a diagonal matrix with all ones except for those entries corresponding to the coordinates of the upper and lower points of the tip where it is set to the value 4.

In simulation, where the model used to create the reference image is the same as the one used for the optimization process, this approach has very nice results either using marker centroids (as in [Reilink 2012]) or the corner based features proposed in chapter 4.

The limits of this approach show up when applied in real contexts: observing the results in Fig. 5.1(g), it can be noticed that the image corresponding to the best configuration (in the terms of minimizing the cost function) is not coinciding with the original endoscopic image. To confirm that the obtained solution were not local minima, several starting configurations have been considered. Taking into account the last row case (cf Fig. 5.1), using the different initial conditions listed in table 5.1 the optimization process leads to exactly the same result both in terms of residual error (root mean square error computed considering the distances between extracted corners and estimated corners) and robot configuration. This suggests that the found solution is not a local minimum[1].

As one can expect, the bad convergence in the image is confirmed by high errors in the estimation of the 3D position of the tip of the instruments: in a test sequence of 116 images where the Ground Truth (GT) was made with a electro-magnetic (EM) tracker attached to the end of the bendable part, the resulting RMS deviations are 3.16, 4.54 and 5.45 mm respectively for x (horizontal), y (vertical) and z (depth) coordinates in the camera frame. These visual and numerical results suggest that

[1] Alternatively, it could be a local minimum with a wide attraction basin.

Initial Conditions ($\lambda\,[mm]$, $\phi\,[deg]$, $\theta\,[deg]$)	2D error [pixels]	Estimated Pose $[\lambda, \phi, \theta]$
$[40, 0, 50]$ $[10, 10, 0]$ $[20, -90, 90]$ $[30, 90, 30]$ $[50, -20, 15]$	13.12	[24.78 79.66 5.44]

Table 5.1: This table shows, in its first column, the configurations used to initialize the optimization algorithm (described in section 5.1) for a specific case (cf. last row of Fig. 5.1). The initial condition, in this case, seems not to affect the final solution both in term of image residuals (2nd column) and estimated configuration.

some of the hypotheses made are not reliable either on the geometric or on the camera model.

The low reprojection error obtained during the camera calibration (RMS error of [0.48 0.49] pixels over 15 different points of view) suggests that the large error in the image space cannot be linked only to a poor camera calibration. The main reason, then, seems to be an incorrect geometrical model and, consequently, we decided to extend this model by incorporating new "spurious" DOFs corresponding to those mechanical parameters whose real values are difficult to measure correctly or can vary during the manipulation due to the described mechanical plays. These DOFs, that will be called *mechanical* DOFs, are the position of the instrument channel exit ($[x_{ch}, y_{ch}]$) and its orientation (angles (ψ, μ)), both wrt to the camera.

5.2 Variable mechanical parameters

With the will of including the four *mechanical* DOFs, a new vector of DOFs can be defined as $\mathbf{r}_e = [x_{ch}, y_{ch}, \psi, \mu, \lambda, \phi, \theta]^T$ (cf. Fig. 3.1(a) and (b)). The z_{ch} coordinate is considered to be constant and equal to the nominal value. In fact, considering the possible movements along z_{ch} would be redundant with variation of the λ DOF: variations of z_{ch} (*ceteris paribus*) only regulates how much the instrument exit its housing channel which results to be similar to the effect achieved with variation of λ.

When the number of parameters to estimate increases, the probability of run into local minima increases as well. In addition to that, even when there are more matches than unknown free parameters, it is often the case that some of the matches have some relationships which lead to an ill-conditioned problem. In this case, the considered parameters are not totally independent in terms of effect on the image features: one needs only to consider, as example, x_{ch} and ψ parameters which both produce an horizontal apparent displacement in the image.

Within this context, Gauss-Newton's method as described above can present

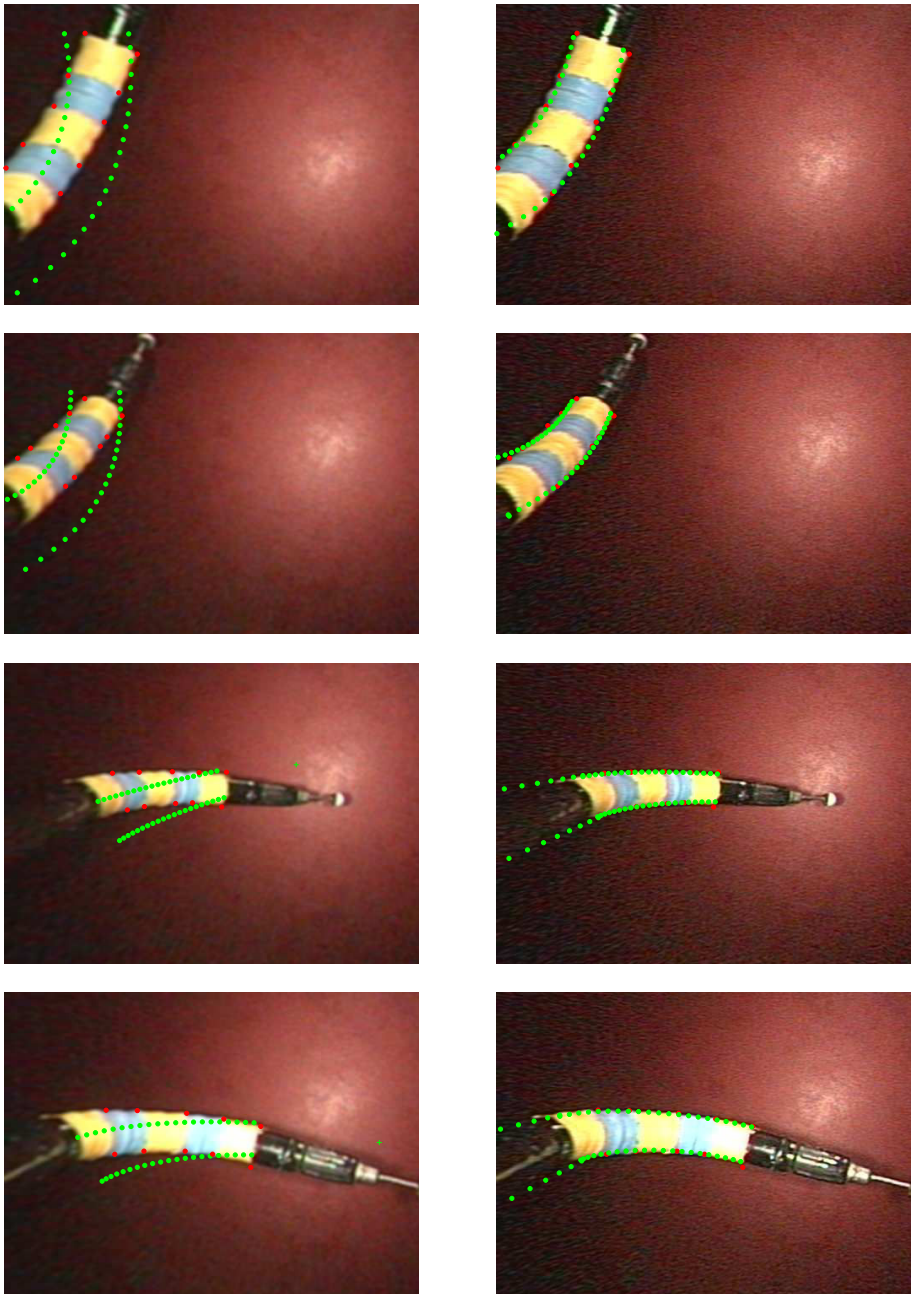


Figure 5.1: Reprojected instrument borders (green spots) after the optimization process. The reference image features are the markers apparent corners (red spots). On the left the result considering the theoretical CAD parameters. On the right the solution with some tolerances on the mechanical parameters.

some problems of convergence. Thus, to assure convergence and speed up the optimization process, a Levenberg-Marquardt (LM) approach has been chosen instead of the described Gauss-Newton.

The formalization is exactly the same with the only difference that the proposed optimal step is a combination of the Gauss-Newton solution with a gradient steepest descent step. The combinatorial weights $\ni$ are varying according to the cost value at each iteration: if, with the computed step, the cost χ^2 decreases $\ni$ is divided by a factor greater than 1, whilst, on the contrary situation, the weight related to the gradient descent is increased. The equation (5.3) becomes:

$$\delta\mathbf{r} = c\,(\hat{\mathbf{L}}^T\mathbf{W}^T\mathbf{W}\hat{\mathbf{L}} + \nu\mathbf{I})^{-1}\hat{\mathbf{L}}^T\mathbf{W}^T\mathbf{W}\mathbf{e}. \tag{5.4}$$

5.2.1 Managing the mechanical parameters

In addition to the adoption of LM, with the aim to improve the results and limit local minima due to partial redundancy on some parameters, prior constraints can be introduced to define the default to use in case of absence of further information [Lowe 1991]. Lowe proposes to add these constraints to the previous system of equations (5.2), obtaining[2]:

$$\begin{bmatrix}\mathbf{I}\\ \mathbf{L}\end{bmatrix}\dot{\mathbf{r}}_\mathbf{e} = \begin{bmatrix}\mathbf{d}\\ \mathbf{e}\end{bmatrix}$$

where the identity matrix $\mathbf{I}$ add one row for each parameter whose displacement can be specified *a priori* and the vector $\mathbf{d}$ collects all the desired default displacement values of each parameter. The problem in such solution is that there is no specification on the trade-offs between meeting the default values or the actual external constraints coming from the image. The solution proposed in [Lowe 1991] is to weight each row so that each element on the right-hand side has the same standard deviation:

$$\begin{bmatrix}\mathbf{W}\\ \mathbf{L}\end{bmatrix}\dot{\mathbf{r}}_\mathbf{e} = \begin{bmatrix}\mathbf{W}\mathbf{d}\\ \mathbf{e}\end{bmatrix}$$

where $\mathbf{W}$ is a diagonal matrix whose elements are the inverse of the standard deviation of each parameter. Doing so, during the optimization step each component will contribute in proportion to the number of standard deviation from its nominal value.

For the pose computation of the flexible instrument, no prior information is available for the 3 DOFs of the instruments, but it can be reasonable to assume that the actual values of the mechanical parameters are close to their nominal values defined by the CAD model furnished by the manufacturer ($x_{ch} = -13.3$, $y_{ch} = 6.2$, $\psi = 10°$ and $\mu = 0°$).

[2]The notation here is simplified and details will be given later on. $\mathbf{L}$ is computed from (5.8) where 4 blocks are added corresponding to the new DOFs.

As already mentioned, the variation of the mechanical parameters is mainly due to the necessary mechanical play between instrument and housing channel. Constant weights as in [Lowe 1991], then, would not model this phenomena properly: a mechanical play lets almost a free movement in a small interval around the resting position whereas the constraints get stronger as it gets far from its resting state (described by the nominal value furnished by the manufacturer).

A peculiar weighting function, then, is used in the cost function (cf. eq. (5.1)) to maintain the estimation of these parameters ($\mathbf{r}_m = [x_{ch}, y_{ch}, \psi, \mu]^T$) close to their nominal values $\mathbf{r}_m^* = [x_{ch}^*, y_{ch}^*, \psi^*, \mu^*]^T$. To replicate the mechanical play description, this weighting function should have almost a "dead-zone" around the parameters' nominal values and an increasing steepness as getting far from them. Furthermore, it must be $\mathcal{C}^1$ in its whole domain ($\mathbb{R}$ in this case) so as to be used in the optimization process.

The function we propose is the modulus of a cubic function where dead-band and steepness are parametrized:

$$\rho(u) = \frac{k_w}{3a^3}|u|^3,\ a \in \mathbb{R}^+,\ k_w \in \mathbb{R}^+ \tag{5.5}$$

Its corresponding weighting function is:

$$\zeta(u) = \sqrt{\frac{k_w}{a^3}}|u|,\ a \in \mathbb{R}^+,\ k_w \in \mathbb{R}^+ \tag{5.6}$$

where k_w determines the steepness of the function outside the dead-zone whose width is defined by parameter a. Going back to the Lowe's framework, it would result in:

$$\left[\begin{array}{cccc|c} \zeta(x_{ch}^* - \hat{x}_{ch}) & 0 & 0 & 0 & \\ 0 & \zeta(y_{ch}^* - \hat{y}_{ch}) & 0 & 0 & \mathbf{O}_{4\times 3} \\ 0 & 0 & \zeta(\psi^* - \hat{\psi}) & 0 & \\ 0 & 0 & 0 & \zeta(\mu^* - \hat{\mu}) & \\ \hline & & \mathbf{L} & & \end{array}\right] \dot{\mathbf{r}}_\mathbf{e} = \begin{bmatrix} \zeta(\mathbf{r}_m^* - \mathbf{r}_m)^T\,(\mathbf{r}_m^* - \mathbf{r}_m) \\ \mathbf{e_{2d}} \end{bmatrix}$$

where $\zeta(\mathbf{r}_m^* - \mathbf{r}_m) = [\zeta(x_{ch}^* - \hat{x}_{ch}), \zeta(y_{ch}^* - \hat{y}_{ch}), \zeta(\psi^* - \hat{\psi}), \zeta(\mu^* - \hat{\mu})]$.

From the point of view of the optimization process, the similarity criterion also changes so as to take into account the weighted distances of the estimated mechanical parameters from their nominal values. The new cost function to minimize is:

$$\chi^2 = \frac{1}{2}\sum_{i=1}^{N} w_i\,[\mathbf{p_i} - \hat{\mathbf{p}}_\mathbf{i}]^2 + \rho_{x_{ch}}(x_{ch}^* - \hat{x}_{ch}) + \rho_{y_{ch}}(y_{ch}^* - \hat{y}_{ch}) + \rho_\psi(\psi^* - \hat{\psi}) + \rho_\mu(\mu^* - \hat{\mu}) \tag{5.7}$$

and the new visual measurement error vector is re-defined as:

$$\mathbf{e} = \begin{bmatrix} (\mathbf{r_m}^* - \hat{\mathbf{r}}_\mathbf{m})^T & \mathbf{e_{2d}} \end{bmatrix}^T$$

where $\mathbf{e_{2d}}$ is the row vector containing the errors committed on the visual features (upper and lower corners points) along each image coordinate:

$$\begin{aligned}\mathbf{e_{2d}} = \; & [(\mathbf{p}_{u,1} - \hat{\mathbf{p}}_{u,1})_x \quad (\mathbf{p}_{u,1} - \hat{\mathbf{p}}_{u,1})_y \quad \cdots \quad (\mathbf{p}_{u,n} - \hat{\mathbf{p}}_{u,n})_x \quad (\mathbf{p}_{u,n} - \hat{\mathbf{p}}_{u,n})_y \\ & (\mathbf{p}_{l,1} - \hat{\mathbf{p}}_{l,1})_x \quad (\mathbf{p}_{l,1} - \hat{\mathbf{p}}_{l,1})_y \quad \cdots \quad (\mathbf{p}_{l,n} - \hat{\mathbf{p}}_{l,n})_x \quad (\mathbf{p}_{l,n} - \hat{\mathbf{p}}_{l,n})_y]\end{aligned}$$

Since the variable parameters have increased (vector $\mathbf{r}$ now include the mechanical parameters), an extended version of the geometrical Jacobian must be considered. Equation (3.31) now becomes (see Appendix C for the computation details of the different blocks A_i):

$$\dot{\mathbf{P}} = \mathbf{J_g}\dot{\mathbf{r}}_e = \left[\; A_{x_{ch}} \; \middle| \; A_{y_{ch}} \; \middle| \; A_\psi \; \middle| \; A_\mu \; \middle| \; A_\lambda \; \middle| \; A_\phi \; \middle| \; A_\theta \; \right] \dot{\mathbf{r}}_e. \tag{5.8}$$

According to (5.8), the *global* geometrical Jacobian will have at most 24 rows (it may be less according to the number of the extracted visual feature points) and 7 columns. Considering the same ordering as in the error vector, the first 4 rows relate the variations of the mechanical parameters $\mathbf{r}_m$ to the variations of the deviation from their nominal values $(\mathbf{r}^*_m - \mathbf{r_m})$. Since these quantities depend only on $\mathbf{r}_m$ itself, the top left corner of the global $\mathbf{J_g}$ (4×4 matrix) will be the opposite of the identity matrix. The rest of this matrix is made up of the composition of the $\mathbf{J_{g}}_i$ calculated for each upper ($\mathbf{J_{i,u}}$) and lower ($\mathbf{J_{i,l}}$) border point of the $i-$th section of the bending part of the instrument. The corresponding weighting matrix is:

$$\mathbf{W} = \left[\begin{array}{cccc|c} \zeta(e_1) & 0 & 0 & 0 & \\ 0 & \zeta(e_2) & 0 & 0 & \mathbf{O}_{4\times 20} \\ 0 & 0 & \zeta(e_3) & 0 & \\ 0 & 0 & 0 & \zeta(e_4) & \\ \hline \multicolumn{4}{c|}{\mathbf{O}_{20\times 4}} & \mathbf{W_{2d}} \end{array}\right].$$

5.2.2 Expressing the error wrt the Bézier curve

Observing more closely the images and the extracted features, it has been seen that the fitting of Bézier curve gives precise results whereas the exact placement of the corner point along the Bézier curve turns out to be more complicated. The strong frontal illumination seems to modify how the color is received by the endoscopic camera, consequently the inter-marker color transition in the image becomes wider (or even absent in case of saturation) and, therefore, a sharp limit is more complex to detect. Therefore, it can be stated that the misplacement error of the corner points location along the direction of the real border is greater than the one along the perpendicular direction. This aspect has been included to improve the optimization result. The idea is to express the 2D error vector $(\mathbf{p_i} - \hat{\mathbf{p}}_\mathbf{i})$ wrt a new frame whose axes are parallel ($s_\parallel$) and perpendicular ($s_\perp$) to the Bézier curve at the corresponding feature point ($\mathbf{p_i}$) as shown in Fig. 5.2.

This allows to weight differently the two new coordinates according to the experimental considerations explained earlier.

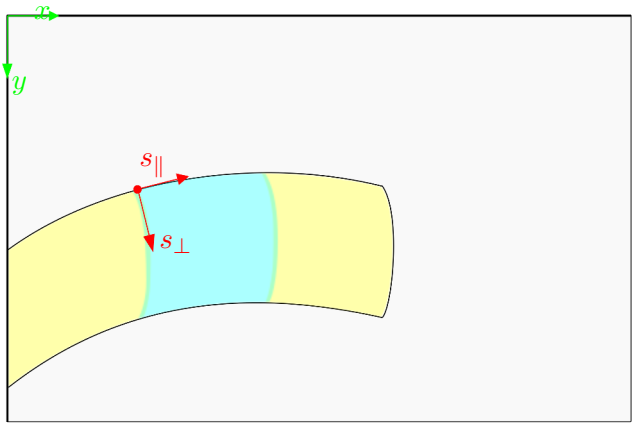


Figure 5.2: Scheme showing the reference frame used to express the visual error on the marker corners. $s_{\parallel}$ and $s_{\perp}$ are computed analytically based on the expression of the first derivative of the Bézier curve computed for each instrument border.

This improvement can be smoothly added to the optimization framework explained so far. Indeed, (3.25) is equivalent to:

$$\mathbf{T}_i \dot{\mathbf{p}}_\mathbf{i} = \mathbf{T}_i \mathbf{J}_{\mathbf{I_i}} \mathbf{J}_{\mathbf{g_i}} \dot{\mathbf{r}} \tag{5.9}$$

where $\mathbf{T}_i$ is the 2D homogeneous transformation between the image frame and the local frame associated to the i-th feature point. Thus, for each point $\mathbf{p_i}$ it can be written:

$$\delta\mathbf{r} = c\,(\hat{\mathbf{L}}_\mathbf{i}{}^T \mathbf{T_i}{}^T \mathbf{W_i}{}^T \mathbf{W_i} \mathbf{T_i} \hat{\mathbf{L}}_\mathbf{i} + \nu I)^{-1} \hat{\mathbf{L}}^T \mathbf{T}_i \mathbf{W_i}{}^T \mathbf{W_i} \mathbf{T}_i \mathbf{e_i} \tag{5.10}$$

where $\tilde{e} = \mathbf{T}_i \mathbf{e_i}$ is the error expressed in the Bézier -based coordinate frame and W_i will have different weights for each component. The generalization considering more than one corner point $\mathbf{p_i}$ is straightforward.

To compute $\mathbf{T}_i$, the tangent to the points of the Bézier curve must be known. This can be computed from the curve derivative

$$B'(t) = 2(t-1)(F_1 - F_0) + 2t(F_2 - F_1)$$

provided that the value of t corresponding to the considered p_i is known.

5.3 Simulation Study

Before applying the algorithm to real video sequences, some simulations tests have been executed. The aim of these test was simply to verify if the proposed mechanical parameters tolerances could cope with the uncertainties on the model. Therefore, no Bézier reference frame is employed. For these tests, two models have been considered: the first model (*correct* model) is considered to be the real one and it is fixed and used to create the virtual reference images (synthetic images substituting

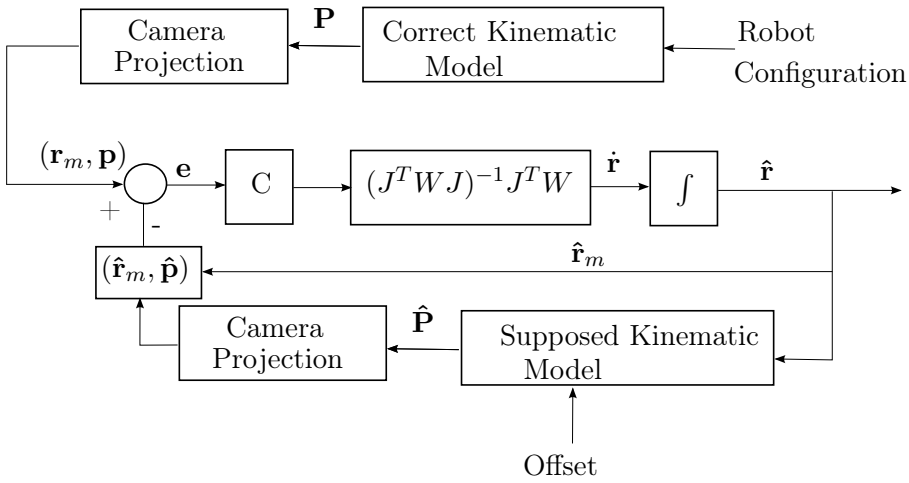


Figure 5.3: Scheme adopted for the simulation where the kinematic model used for the pose estimation has an offset on $\mathbf{r}_m$ wrt to the reference kinematic model.

the real endoscopic images) and the *supposed* model which differs from the reference model by some error (offset) on the mechanical parameters and which is used in the optimization process to estimate the 3D pose (scheme of Fig. 5.3). Obviously, the proposed weighting functions are centered around the supposed mechanical values.

To decouple the effects of the errors, the offset was added either on the channel position or on the channel orientation with respect to the camera (the cases are listed in table 5.2, the considered correct values are $(x_{ch}, y_{ch}) = (-13.3, 6.6)$ and $\psi = -10°$) and $\mu = 0$[3]. For each of the listed cases, the optimization algorithm was tested on 500 robot configurations and using two different settings for the weighting function (5.6):

Narrow DB	Large DB
$k_w = 15$, $a = 1$ for both ζ_{ch}	$k_w = 5$, $a = 3$ for both ζ_{ch}
$k_w = 100$, $a = \pi/180$ for ζ_ψ	$k_w = 10$, $a = \pi/60$ for ζ_ψ

As suspected, in case of uncertainty on the model (either on the ψ or channel position), considering a fixed mechanical model does not provide good results neither in terms of 3D position estimation (Fig. 5.4(a) and (b)) nor in terms of image convergence (Fig. 5.4(c) and (d)). On the other hand, the proposed method seems to be able to mitigate the counteracting effect deriving by model uncertainties. A fundamental role on this method, though, is played by the weighting policy. The dead-band (DB), indeed, must be chosen according to the expected error on the mechanical parameters: in the case where the DB is smaller than the error on the model, the optimal solution is far from being the correct one (red line in Fig. 5.4(a)). On the other hand, if the dead band is correctly centered, the convergence

[3]The mechanical parameters μ is considered known in this study. The chosen parameters are sufficient to discuss the point. Variation of μ would lead to totally similar results

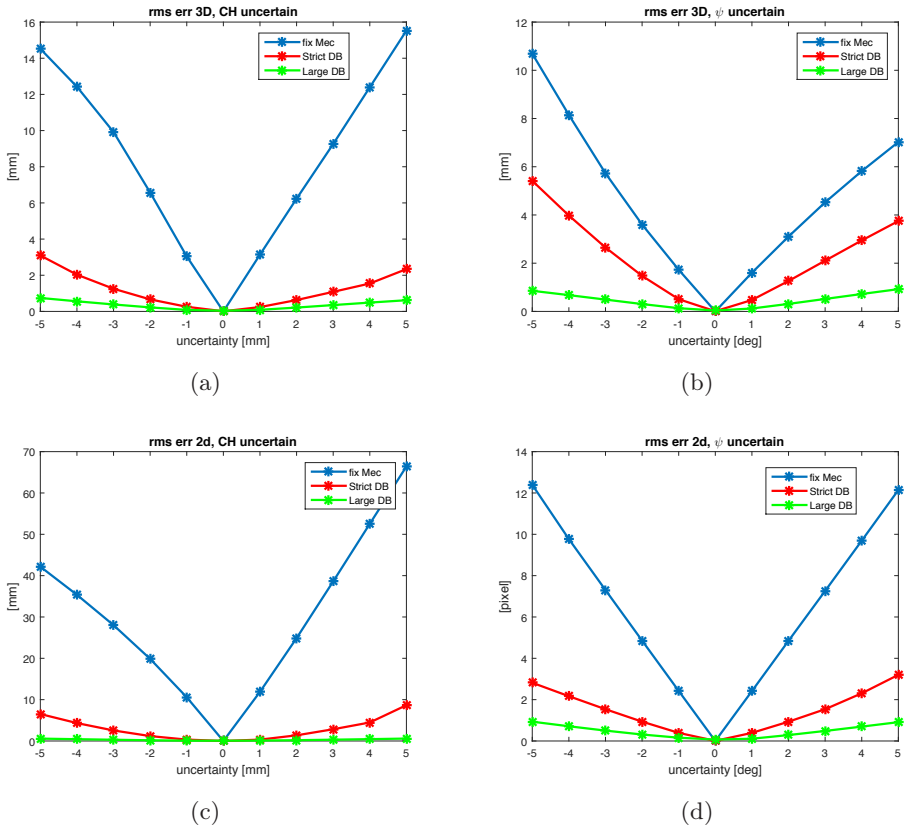


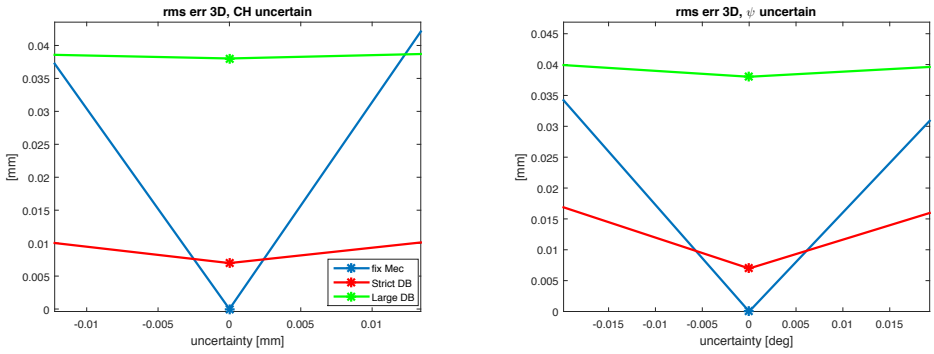
Figure 5.4: Results of simulation considering an error in the model on the instrument channel position. The graphic shows the outline of the RMS error on the tip 3D position estimation (top line) and of the 2D error over all the corners (bottom line) for different cases (cf 5.2). In (a) and (c) the error is introduced in the instrument channel position and in (b) and (d) the error is introduced only in the orientation of the channel wrt to the camera.

Test case Test case	Channel on supposed model [mm]	Test Case Test Case	ψ on supposed model [deg.]
1	$(x_{ch}-5, y_{ch}-5)$	12	$\psi-5$
2	$(x_{ch}-4, y_{ch}-4)$	13	$\psi-4$
3	$(x_{ch}-3, y_{ch}-3)$	14	$\psi-3$
4	$(x_{ch}-2, y_{ch}-2)$	15	$\psi-2$
5	$(x_{ch}-1, y_{ch}-1)$	16	$\psi-1$
6	(x_{ch}, y_{ch})	17	ψ
7	$(x_{ch}+1, y_{ch}+1)$	18	$\psi+1$
8	$(x_{ch}+2, y_{ch}+2)$	19	$\psi+2$
9	$(x_{ch}+3, y_{ch}+3)$	20	$\psi+3$
10	$(x_{ch}+4, y_{ch}+4)$	21	$\psi+4$
11	$(x_{ch}+5, y_{ch}+5)$	22	$\psi+5$

Table 5.2: Offsets used for the simulation tests. The supposed kinematic model used to estimate the 3D pose is equal to the correct model used to create the synthetic image except for the 3D position of the instrument channel (which is displaced) or the orientation of the channel wrt to the camera axis.

to the optimal solution is faster and slightly more precise (given an equal number of iterations) when the dead-band is narrow (Fig. 5.5(a) and (b)) .

Although, choosing an arbitrarily large DB would probably increase the number of local minima (increasing the risk of getting no optimal solution at all) due to the already mentioned redundancy between different DOFs. In fact, the variations of some components of $\mathbf{r}_e$ have similar effect on the image and, since the optimization criterion is based on image features only regarding the bendable part, a compensation between two or more parameters may occur leading to only slightly changes or no effect on the cost. This is the case, for example, of x_{ch} and ψ which do not have the same effect in 3D but several combinations of the two can lead to the same image. This is somehow confirmed by the results in Fig. 5.6(a): even though the model error is introduced only on ψ (cases 12-22 of table 5.2), the estimated channel exit position is farer from the actual value when a larger DB is used. In this case, the algorithm cannot identify the actual mechanical value and both ψ and the channel position seem to loose their physical meaning in favor of the global criteria. The risk of such behavior would be more evident in a larger scale. Let us consider, for example, an horizontal and bent instrument, very similar images can be obtained with the nominal values ($x_{ch} = -13.3$ and $\psi = -10°$ wrt the camera coordinate reference frame) but also with an x_{ch} far from the camera on the left and a positive ψ, as if the instrument were outside the channel and positioned almost parallel to the (x, y) camera plane. If the mechanical parameters are not limited around their nominal values, such solutions would be valid even though they would



(a) (b)

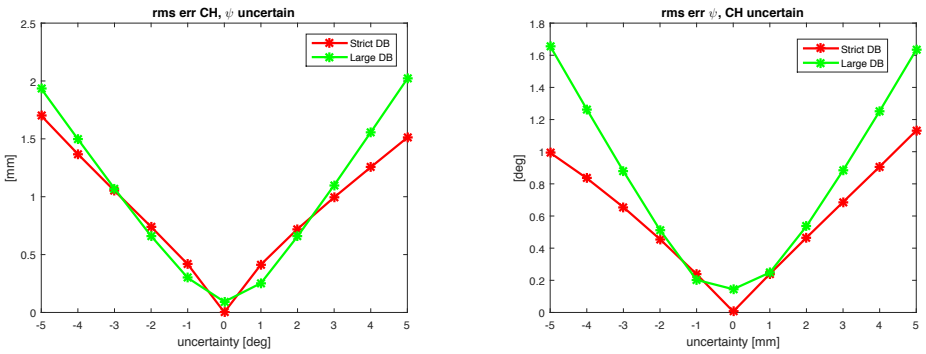
Figure 5.5: Zoom on the origin of Fig. 5.4(a) and (b). If the dead band is correctly centered (i.e. small uncertainty on the model), the convergence to the optimal solution is faster and slightly more precise (given an equal number of iterations) when the dead-band is narrow.

not have any physical meaning. This could be solved if the non-controllable part of the instrument were taken into account, but this is often not possible since, most of the time, only the bendable part is outside the channels during operations.

The beforehand mentioned behaviors are amplified when noise is added to the visual measurements. The results in Fig. 5.7(a) and 5.7(b) underline again the importance of the *a priori* knowledge for the centering of the dead-band. With the effect of noise, though, this aspect acquire more importance. In fact, in case of smal uncertainty, the adoption of a large DB policy provides lower accuracy results with respect to a narrow DB policy. However, large DB is applicable when the *a priori* is really untrustworthy or when large movement are expected to occur during the operation. Even though, in this case, would be better to update the center of the weight function (according to some data) and use a strict DB policy.

Also the forementioned parameters compensation issue is more evident in presence of image noise: when a large tolerance is admitted, the estimation error on the mechanical parameters consistently increases (channel position in Fig. 5.7(c) and ψ in Fig.5.7(d)) loosing their physical meaning.

Since, for biomedical applications, high accuracy is needed even with noisy data and considering that we can quite precisely centering the weighting function (i.e. assuming that the mechanical parameters nominal values furnished by the provider are sufficiently precise), the most adapt weighting policy is using a narrow DB.



(a) (b)

Figure 5.6: RMS errors committed in the estimation of the channel position when the supposed channel position correspond to the correct value (cases 12-22) (a) and those committed in the estimation of the ψ when the supposed ψ is equal to the correct value (cases 1-11) (b).

5.4 Motor Sensor Data Fusion

Earlier it was stated that no *a priori* was available for the real DOFs of the instrument, but this is not totally true. In fact, if on one hand no prior hypothesis or considerations can be done on the configuration, on the other hand robot sensor data can be exploited to have an initial guess of the instrument position. Actually, a good confidence can be given to the rotation and translation encoders values whilst encoders value of the deflection are difficult to relate to real instrument deflection value due to all the phenomena in the transmission chain (friction, cable extension or slack).

As pointed out previously, it is difficult to have a high precision on TCP position looking only at the encoders values, but it can be thought to exploit their information to discriminate those cases where the visual measurements are only partially available or in those positions of the visual features space where the uncertainty on the estimation is high (cf. sec. 3.5). In other words, the idea is to fusion the encoder data (or better said the instrument pose variation deriving by the encoder data variations) with the visual measurement. For this scope, a Kalman filter framework is used where the increment of encoder value is used to update the state prediction and the visual-based pose estimation is used as the measurement.

Thanks to this formalization, wrong estimations based on the image can be detected and mitigated. Furthermore, this allows to inherently consider a smoothness and continuity in the movement, although the precision on the encoders is not such to allow an accurate estimation in case of hidden instruments (and consequence lost of all the visual information).

For applying the Kalman framework, the state space model and the measure-

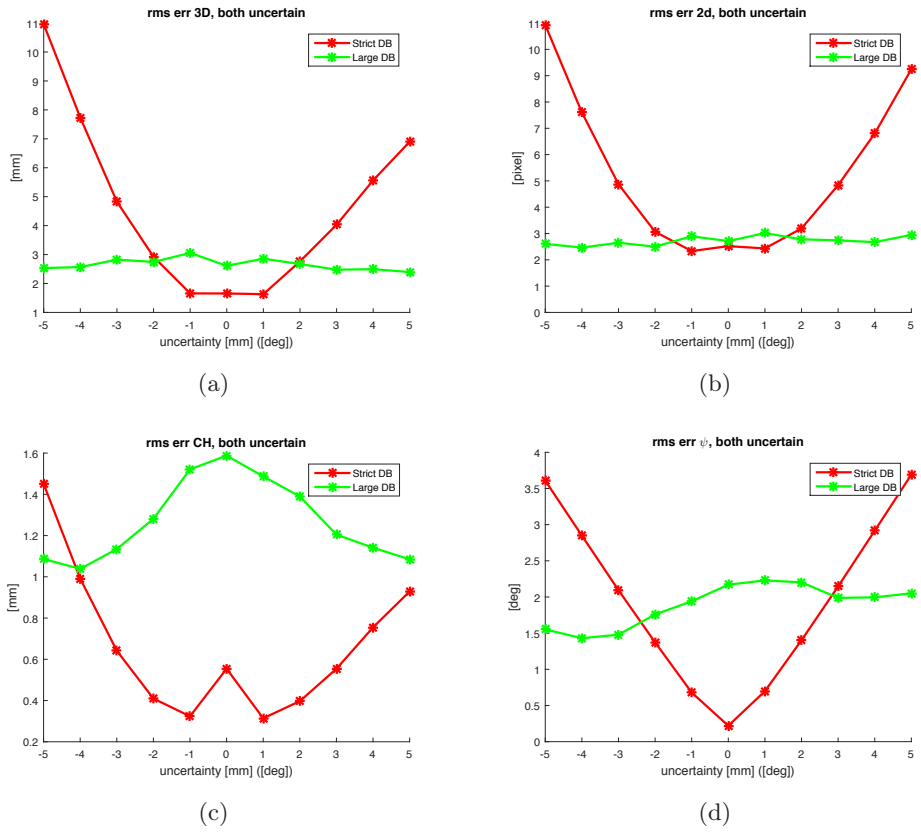


Figure 5.7: Effect of adding noise at the visual measurements (i.e. the position of the markers' corners) and considering uncertainty on both ψ and channel position (x_{ch}, y_{ch}). In (a) the RMS of the 3D position estimation errors, in (b) the RMS errors between actual and reprojected corners in the image (after convergence of the optimization algorithm). RMS error on channel position and orientation are shown respectively in (c) and (d).

ment model of the process are supposed to be known and both corrupted by noise which is usually considered to be additive, white, Gaussian and with zero mean. In mathematical terms, considering a linear, discrete-time dynamical system, the process equation is:

$$\mathbf{x}_{k+1} = \mathbf{F}_{k+1,k}\mathbf{x}_k + \mathbf{w}_k \tag{5.11}$$

where $\mathbf{F}_{k+1,k}$ is the transition matrix bringing the state $\mathbf{x}$ from time k to $k+1$ and $\mathbf{w}_k$ is the process noise such that

$$E[\mathbf{w}_n\mathbf{w}_k^T] = \begin{cases} \mathbf{Q}_k & \text{for} \quad n = k \\ 0 & \text{for} \quad n \neq k \end{cases} .$$

The measurement can be described by a similar equation:

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x}_k + \mathbf{v}_k \tag{5.12}$$

where $\mathbf{y}_k$ is measurable at time k and $\mathbf{H}_k$ is the measurement matrix. For measurement noise $\mathbf{v}_k$ the same hypothesis hold and

$$E[\mathbf{v}_n\mathbf{v}_k^T] = \begin{cases} \mathbf{R}_k & \text{for} \quad n = k \\ 0 & \text{for} \quad n \neq k \end{cases} .$$

In the case an external action has to be consider, the system model (5.11) can be extended as follows:

$$\mathbf{x}_{k+1} = \mathbf{F}_{k+1,k}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \mathbf{w}_k \tag{5.13}$$

where $\mathbf{u}_k$ is the aforementioned external control action and $\mathbf{B}$ describes how $\mathbf{u}_k$ acts on the state evolution.

The presented framework can be adapted to our purpose of data fusion in the following manner:

- The process can be considered like a time-discrete process whose sampling time coincide with the camera frame rate.

- It can be expressed as a linear process if the DOFs vector is chosen as state ($\mathbf{x}_k = [\lambda, \phi, \theta, x_{ch}, y_{ch}, \psi, \mu]^T$) and the measured variable is the state itself. Focusing on the "natural" DOFs (λ, ϕ, θ), it reasonable to assume that the state changes only if a motor action occurs. On the other hand, the mechanical parameters do not seem to follow any deterministic evolution law, therefore we preferred not to perform any update to this part of the state either. For these reasons, we decided to consider the transition matrix equal to the identity and take as measurement the state estimated by the visual information, this means: $\mathbf{F}_{k+1,k} = \mathbf{H}_k = \mathbf{I}$, $\forall k$.

- The motor velocity can be considered as the external action $\mathbf{u}_k$ which updates the state (i.e. the DOFs values). Consequently,

$$\mathbf{B} = \left[\begin{array}{ccc|c} 1 & 0 & 0 & \\ 0 & 1 & 0 & \mathbf{O}_{3\times4} \\ 0 & 0 & 1/r & \\ \hline \multicolumn{3}{c|}{\mathbf{O}_{3\times4}} & \mathbf{O}_{4\times4} \end{array}\right]$$

  The mechanical parameters, although estimable from visual information, are not real controllable DOFs and, thus, no motor velocity is available for them and neither their variation has a clear relationship with the encoder data variation. Thus, the last four component of $\mathbf{u}$ (encoder data) and the last four rows of $\mathbf{B}$ (relationship between motor variation and state) will be equal to zero. Anyway, considering true the supposition of the mechanical play and the movements in free space, the robot configuration (measured by the encoders) and the estimated values of the mechanical parameters should be related i.e. the position of the instrument inside the channel depends on the particular configuration.

  Since the bending is obtained by actuation on antagonist cables (cf Fig. 2.20 and 3.1(b)), the related encoder returns the difference between the two cables lengths. Dividing by radius of the instrument bendable section r, the deflection angle in radians is obtained.

- The measurement and prediction confidence trade-off can be made vary in function of the available quantity of the visual information. Logically, if the visual information is only partially available (only few corners are detected), the confidence of the model should be increased to the detriment of the confidence on the measurement.

To obtain a linear system, we have considered the estimated state as measurement. If on one hand, this choice allows to have an easy formalization of the Kalman filter on the other it threatens the hypothesis over the measurement noise. In fact, even though the image noise can be considered white Gaussian and zero mean, nothing can be told about the noise associated to a measurement (like the one proposed) that depends, in turn, of the image noise level. Therefore, it cannot be certainly stated that the measurement noise is still additive, white, Gaussian and with zero mean.

5.5 Experimental Results

To evaluate the accuracy and relevance of the proposed method for pose estimation, a sequence of instrument configurations was generated spanning the whole visible left instrument workspace. At each configuration (defined by three values of the 3

DOFs) an image was acquired simultaneously by the endoscopic camera and by the two cameras of a stereo system. The sequence of configurations was chosen such that the instrument makes a continuous movement in the picture too as if it were a real trajectory. Thanks to stereo vision the 3D position of a white ball in-built with the tip of the instrument is determined and used as Ground Truth (GT) (see appendix B for more details on the set-up for creating the 3D GT). The accuracy is, then, evaluated as the 3D distance between the estimated white ball position (based on instrument model and image features) and the GT.

To obtain some meaningful statistics on the accuracy of the proposed method, a sequence of 295 poses is considered. Over the corresponding images the segmentation process failed on 14 images which gives a useful set of 281 images. A strict Dead-Band policy (cf table in sec. 5.3) is used for the admitted tolerances of the mechanical DOFs using for μ the same weighting function as ψ. Following the Bézier -based reference frame, the weight associated to all the parallel components of the 2D errors is 8 and the one associated to the perpendicular components is 10.

Moreover, the analysis of several videos of surgical operations using the Anubis platform shows that the instruments displacements can be considered very small between successive frames. Thus, the solution configuration for the image frame j is used as initial guess for the optimization problem in frame $j+1$, with the result of speeding up the computation time for the pose estimation.

The good quality results over these images (Fig. 5.8) are supported by the round mean square (RMS) errors over each coordinate (in mm): 2.03 ± 1.65 on x, 2.07 ± 2.07 on y and 2.78 ± 2.57 on z axis.

To obtain these results, all the images have been considered including those where the features where only partially visible. In fact, when the instrument is very near to the camera the initial part of the instrument bendable section is hidden inside the channel or is outside the image. Moreover, in order to use projective geometry in pose estimation, the reference image must be rectified. It is a well known issue that for high-distorsion lenses the reliability of the camera model is weak near the border of the image. In the rectification process, then, the "reliable" field of view is limited and, usually, a part of the border is not taken into consideration causing a partial exclusion of the painted bendable section from the image.

When the validation 3D point is near a set of visual features (the corners associated to the tip in this case) used in the similarity criterion, a good convergence on this feature is likely to correspond to a good estimation of such point (3D position) no matter the estimated orientation is. In this case, a good 2D convergence on the corners associated to the tip of the instrument would suffice to have a good estimation on the x and y 3D coordinates of the instrument central axis terminal point (the tip of the instrument). Furthermore, since the instrument diameter is known, a quite good estimation on the scale and, indirectly, of the depth of the tip can be obtained. However, performing the validation on such point does not give any information on the orientation of the end-effector and, consequently, on the robot

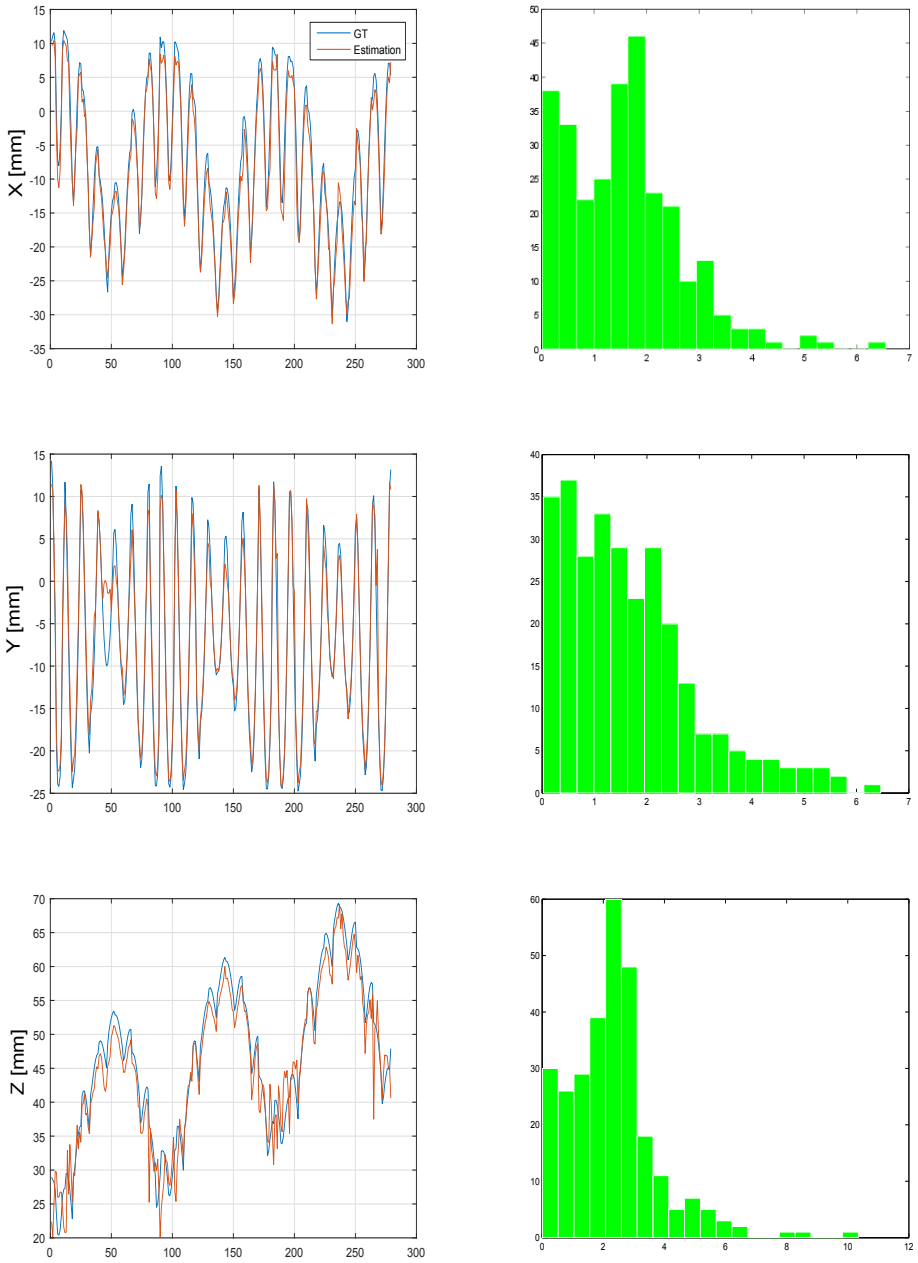


Figure 5.8: On the left: pose estimation result over a trajectory along x (top), y (middle) and z (bottom) coordinates: the blue line is the ground truth position retrieved by the stereo-vision system and the red line is the estimated position.
On the right column: distribution of the estimation absolute error along the different coordinates.

Method		Inspired by [Reilink 2012]	Ours
RMS 3D err [mm]	on x :	3.64	2.03
	on y :	3.45	2.07
	on z :	10.45	2.78
Max 3D Err [mm]		19.74	11.93
2D Error [pixel]		24.12	10.77

Table 5.3: Quantitative results comparison of the proposed model based method with the one inspired by the work of [Reilink 2012].

configuration, which can be wrong.

We have then chosen to employ a reference point (the white ball) distant from the ending of the bendable part so as to have a large 3D position error in those cases where the estimated orientation is wrong. This makes the RMS errors on the 3D position to represent, somehow, also the quality of the entire pose estimation.

Looking at the RMS it can be stated that our method provides a more precise orientation wrt the fix model approach. In fact, the RMS error over this same sequence using fix mechanical parameters increases abruptly when considering the white ball as the validation 3D point (instead of a EM sensor attached at the end of the bendable part) getting to $[3.35, 3.59, 10.54]$ mm respectively along x, y and z coordinates.

Even if the RMS error of the fix mechanical model approach are computed with a GT 3D reference point nearer to the instrument tip (as done in sec. 5.1 with EM sensors), the result is still worse than the one obtained by the proposed method.

Once demonstrated the effectiveness of the consideration of the mechanical plays in the optimization process, another interesting result is the one concerning the contributions of each one of the proposed implementation aspects: Bézier reference frame for the error and motor encoder data fusion. Table 5.4 and Fig. 5.9 summarize these aspects showing the sorted norms of the errors committed in the estimation of the TCP 3D position. Bézier based reference frame seems to improve the pose estimation only slightly and the major contribution is in term of maximum error. Maybe, further improvement could be achieved with a more exhaustive study over the trade-off between the parallel/perpendicular error components. What is evident from data, though, it is the capacity of the Kalman filter with dynamic confidence adjustment of mitigating the aberration in pose-estimation errors taking into account temporal data when visual information is incomplete.

The aforementioned problem deriving by the low reliability of the deflection encoder data is also visible in Fig. 5.10(b) where the 3D TCP positions are represented and a color code is used to represent the norm of the error (the more reddish the

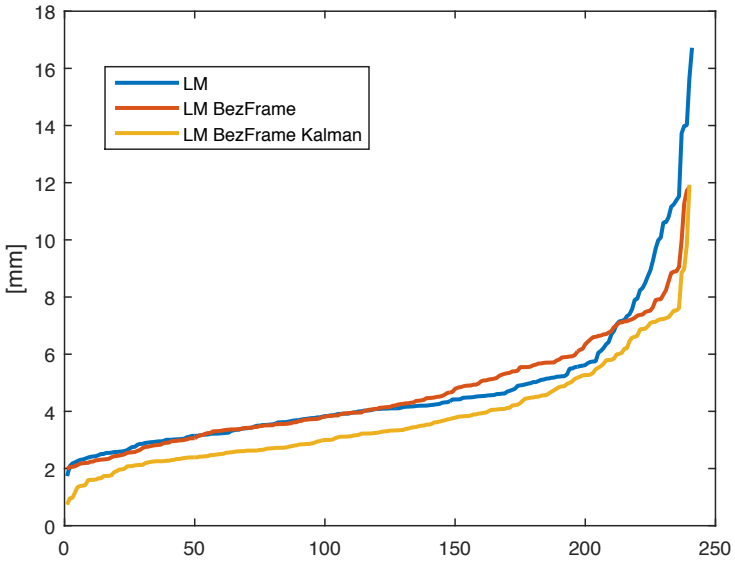
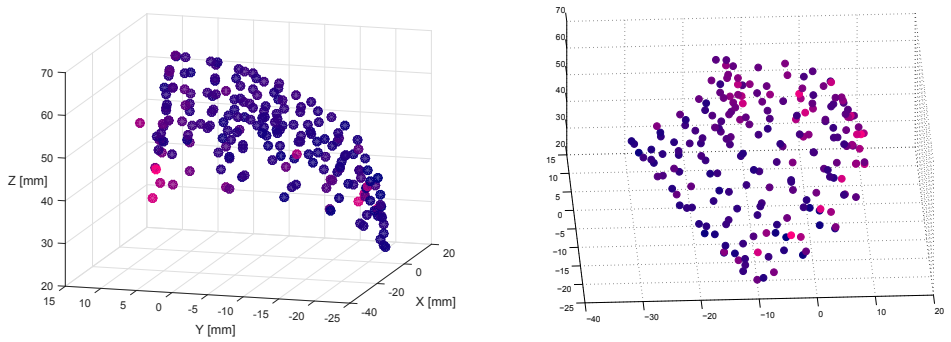


Figure 5.9: The three lines represent the sorted estimation errors norms made on a sequence of the different contribution of the proposed method: pose estimation using a simple LM optimization (blue line), LM optimization expressing the error wrt to a Bézier -based reference frame (red line) and adding motor encoder data to filter the results (yellow line).

Method		LM	LM Bézier Frame	Complete
RMS 3D err [mm]	on x :	3.22	2.98	2.03
	on y :	1.76	1.89	2.07
	on z :	3.76	3.43	2.77
Max 3D Err [mm]		16.72	11.84	11.93

Table 5.4: Numercial results of the different contributions of the proposed method. The three column (from left to rith) represent: pose estimation using a simple LM optimization, LM optimization expressing the error wrt to a Bézier -based reference frame and using motor encoder data to filter the results.



(a) (b)

Figure 5.10: Representation of the 3D TCP positions of the validation sequence and the associated position estimation error considering (in (a)) or not (in (b)) the kalman filtering. The error norm is color coded: the more reddish the higher is the norm of the estimation 3D error.

higher is the error). In this graphic, it can be noticed that some of the highest errors are located on a crown which correspond to those instant where an inversion of the deflection movement occurs. In this case, as already explain in precedence, the lack of the antagonist cable must be recovered before appreciating an actual movement of the instrument. This means that the motor is moving and, consequently, the state is update whereas the instrument stays still. Therefore, a high error is accumulated which, inevitably, affects the pose estimation in the successive frames.

Indeed, this "crown" of high errors is not visible in Fig. 5.10(a) where the Kalman filter is not employed and therefore the real behavior of the proposed model based method can be analyzed. Here, the highest errors seem to be concentrated in a specific region corresponding to those poses where the instrument is deflected, pointing upward/downward and near to the camera. The same behavior can be noticed even if the motor data are considered. This behavior can be explained by the fact that, in these configurations, the instrument appears in the left border of the image where the distortion is higher. This may make the visual measurements less reliable indirectly compromising the estimation pose.

5.5.1 *In-vivo* Qualitative Results

The same process has been applied on *in-vivo* video sequences, acquired in the abdomen of a pig during single port surgery realized with the manual short Anubis platform. No ground truth was available in this case and the proposed analysis is only qualitative. In Fig. 5.11 some examples are shown, where the instrument is first detected (as shown in the precedent chapter) and subsequently its 3D pose is computed based on the extracted corners (green crosses). The corresponding virtual image of the instrument is then reprojected in the image with the estimated

2D position of the center of the grasper (green dot).

These results seems to confirm the capability of the proposed method to estimate the pose of flexible instrument basing on the corners of colored markers.

Furthermore, the consideration of mechanical plays in the model allows the pose estimation process to converge even when the instrument is pushing on the organs and its position inside the channel is consequently modified (Fig. 5.11(a) and (b)). In this case the estimation of the position of the tip is biased because of the deformation of the instrument.

The adaptability of the proposed pose estimation scheme is also demonstrated by the fact that the flexible system used for *in vivo* experiments is different of the one used in the laboratory. Nonetheless the pose estimation works without any modification of the algorithm.

Finally, as also shown in the precedent chapter, the feature extraction algorithm fails when most of the instrument is hidden (Fig. 5.11(e)), since it does not take into consideration any template of the object nor multi-frame analysis. However, even with partial information the computed 3D pose is acceptable in terms of reprojection, indirectly demonstrating the interest of using a model in terms of segmentation.

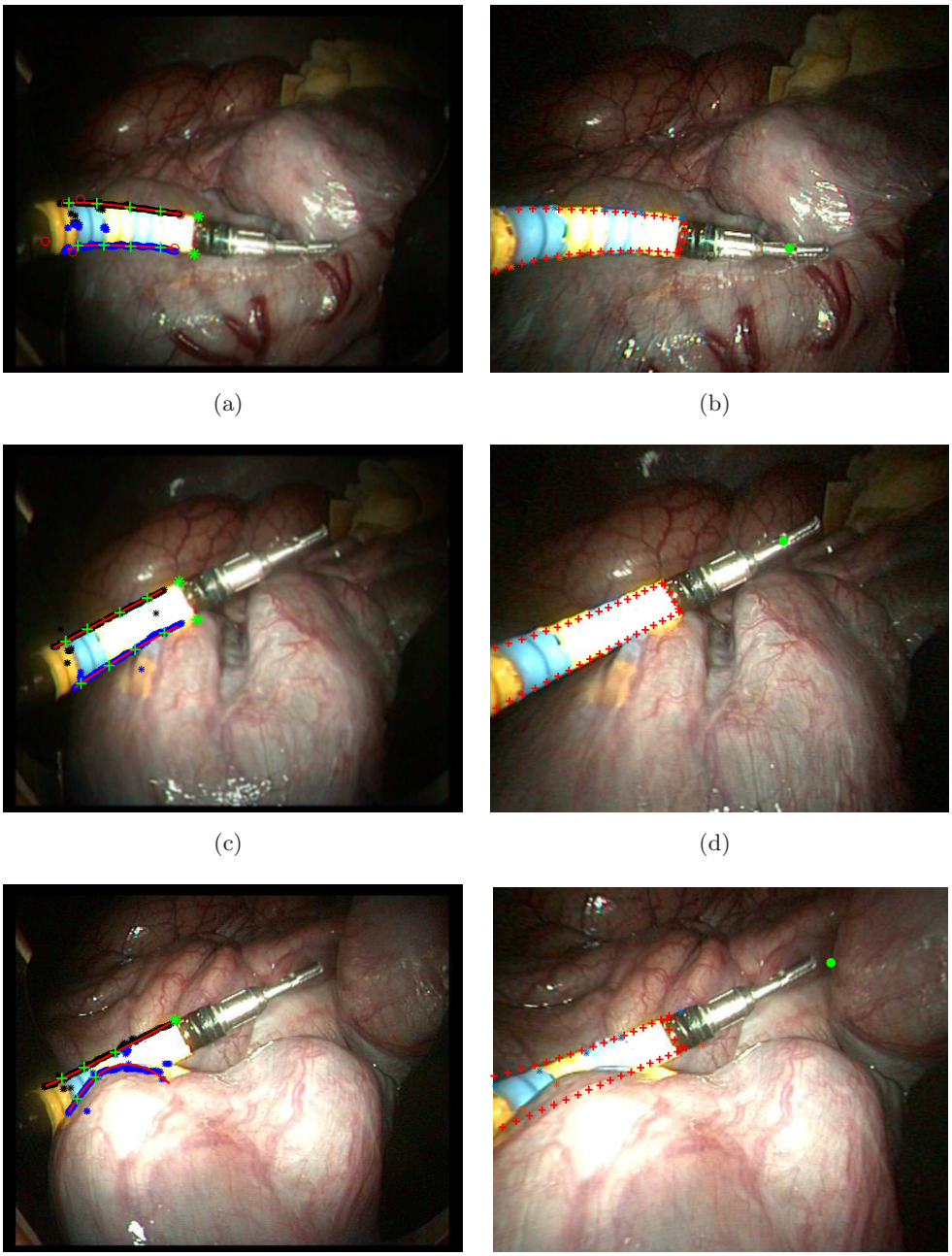
5.6 Conclusion

This chapter has presented the comparison between two model based methods: the first considers the geometrical model known, whereas the second takes into account the possible mechanical plays between the instruments and the housing channel.

Given the evident difference between the two approaches' results, it can be stated that the introduced tolerances on the mechanical parameters allow to account for the mentioned mechanical plays and help compensate uncertainties in the kinematic model (including those regarding the real position of the camera axis wrt to the axis of the channel) or, going a step further, the instrument deviation due to contact with organs (even though, in this case, the weighting function parameters should be adapted or modified along the operation such that the dead band could admit larger displacement either in term of translation or orientation).

Discrete qualitative results have been obtained also in *in-vivo* environment using a different system wrt the one used in laboratory. These results shows the adaptability of the proposed method to different (although similar) systems and, even, when the geometric structure is modified e.g. when the instrument is slightly pushing on the organ. Moreover, from a visual point of view, the same framework could be adapted to recover the instrument outline in case of partial occlusion.

Nevertheless, the observed errors are still large for a surgical application and they are probably the sum of intrinsic issues this approach presents. In addition to segmentation errors which obviously affect the pose accuracy, a part of the error may be explained by the fact that Kalman hypothesis are not totally satisfied. Other



(a) (b)

(c) (d)

(e) (f)

Figure 5.11: in-vivo results of the presented method: (Left) the segmentation process and (right) the reprojection of the instrument from the estimated pose. The green dot is the reprojection of the center of the grasper. In (a), the considered mechanical play allows to fit the instrument even when it is pushing on the organ. In (c), the real instrument is correctly discriminated from its reflection on the organ. Even though in (e) the segmentation process partly fails, an acceptable result is obtained in the instrument fitting and reprojection (f).

errors, then, can arise from considering an incorrect model (e.g. uniform curvature for the bending part, bending on a plane, ...) or can be partly explainable from imperfections in the registration between the camera and the Stereo system Ground Truth.

A learning approach can help avoiding many of these mentioned problems, since it learns the model from the input and output data without making any hypothesis on the system model. The next chapter presents a study of the potentialities of two learning approaches in the context of position estimation from a single 2D image.

Chapter 6

Learning Based Pose Estimation

Contents

It can be said that a machine *learns* when it can improve its performances through *its experiences*. Nowadays, there are many aspects that can be learned conferring to learning a wider and wider variety of meanings according to its applications: model identification, image segmentation, interpretation of data (scenes, situations,..), etc. However, two main categories of learning problems can be distinguished based on the kind of the output variables: in *regression* the objective is to learn a function and return a numerical data output, whereas *classification* aims at inferring a quality or category based on some input features/characteristics/quality.

Learning algorithm are usually divided into two phases: the machine is first trained on a specific set of known samples to learn the rules which it will apply to new samples to determine its output numerical value or its class.

For example, let us imagine that a function f has to be learned and denote h the hypothesis that can be made on the shape of f and $\mathbf{x} = (x_1, \cdots, x_n)$ the vector-valued input. Thus, h can be thought as being implemented by a device that has $\mathbf{x}$ as input and $h(\mathbf{x})$ as output (which, again, can be either a numerical value or a class according to the problem that has to be solved). The only needed *a-priori* is

on the class of function H to whom h pertains. The best h is selected on a training set Ξ of M input vector samples which is usually built according to experiments.

In this framework, there are two major settings in which the function can be learned. In *supervised learning*, both input ($\mathbf{x}$) and corresponding output (y) values of each sample of the training set are known. The optimization criteria, then, can be defined by comparing $h(\mathbf{x})$ and the associated output y.

In *unsupervised learning* the training set is composed only with input vectors without function values. The problem in this case is usually to find the best partitioning of the training set, obtaining some meaningful categories of the input data. The output values in this case will be the category name or the pertaining cluster of the input data.

Once the best guess among the hypothesis H is learned (based on the training set), this can be used with new samples to estimate the output value.

An important role in learning is played by the training set and how much it is representative of the general problem. It can be asserted that the higher is the cardinality (M) of the training set and the diversity of its samples, the better this algorithm performs in terms of estimation accuracy. This fact is quite intuitive: in fact, the more *representative* cases are considered in the training set the higher is the probability that the new *unseen* sample is near (or similar) to those already "seen" during machine training. On the other hand, learning method would loose its sense if the training set contains all the possible cases that can occurs: in this case, in fact, the approach would be more similar to a look-up table and no "estimation" is needed.

In the context of pose estimation, the basic idea is to *learn* the function relating an image feature with the 3D position of the TCP. As enunciated at the end of the previous chapter, the principal aim of adopting a learning approach is to try avoiding all those sources of error deriving by an incorrect system model. Achieving high accuracy in the geometrical model can become hard in the case of flexible instrument which can present several behaviors difficult to detect and, consequently, describe mathematically (e.g.: non constant curvature, no planar curvature, different behavior according to the shape of the rest of the body, ...).

In this case, though, the class (H) of the regression function is hard to be known *a priori* for this kind of problem. Moreover, since the input vector derive from image features, the training set is affected by noise.

Taking into account these characteristics, the searched solution should avoid noise over-fitting and should be able to easily adapt to any shape, allowing local variations fitting without affecting the whole approximation. For this reason, polynomials or linear fitting are unsuitable. A good choice in these terms is the use of a network of Radial Basis Function (RBF) with Gaussian Kernels [Broomhead 1988], [Moody 1989] which provides the desired local approximation (according to an initial cluster computation) and, thanks to the chosen kernels, can easily approximate any smooth function.

Although, the not satisfying quality of the RBF results and the particular shape of the considered function brought us to investigate better approaches both for the clustering (Sec. 6.2) and for position estimation (Sec. 6.4). The proposed methods are first tested with synthetic data with a particular attention to input noise effect on the learning process (Sec. 6.5). Finally, the experimental results are presented and analyzed.

6.1 Radial Basis Functions Network

6.1.1 Principle

The idea underneath this method is very simple: it proposes to address, in a *local* way, the *global* problem of approximating an unknown function. Basically, the desired function is expressed as a combination of local functions (named Kernels) which are responsible of the shape of only a portion of the original function domain. These kernels are defined *radial* functions since their influence diminishes with the increasing distance from their centers.

Therefore, the regression result is a Network of Radial Basis Function (RBF) which, in turn, can be thought as a 2-layers artificial neural network (ANN) that employs RBF as activation functions.

Thus, given $\mathbf{x} = [x_1, \cdots, x_{2n}]^T$ the $2n$ dimensional input vector, the estimation of the output ($\hat{y}$) set up by RBF network is a weighted sum of K kernel functions φ_k:

$$\hat{y} = h(\mathbf{x}) = \sum_{k=1}^{K} \gamma_k \varphi_k(\mathbf{x}) + \gamma_{K+1} \tag{6.1}$$

where γ_k are the learned coefficients which describe the contribution of the k-th kernel to the estimation of the output. In the case of Gaussian kernels, φ_k takes the form [Bors 2001, Lendasse 2003]:

$$\varphi_k(\mathbf{x}) = \mathcal{Y}_k \, e^{-\frac{1}{2}(\mathbf{x}-\mathbf{C}_k)^T \Sigma_k^{-1} (\mathbf{x}-\mathbf{C}_k)} \tag{6.2}$$

where $\mathbf{C}_k$ is the center of the Gaussian kernel, Σ_k its variance and $\mathcal{Y}_k$ its amplitude.

As any other learning approach, a training phase is needed for both layers: firstly the location and shape of Gaussian kernels are computed through unsupervised learning (first layer) and, secondly, the best linear combination of these kernels is computed by supervised learning to approximate the function. Before giving the implementation details for the case concerning image-to-3D-position, an one-dimensional example is presented to clarify and visualize what the different layers mean and how they are built.

6.1.1.1 Example

Let say that we want to approximate with a RBF Network the function in Fig. 6.1(a) which describes the behavior of the system that has to be learned. Obviously, we do

not know the continuous outline of that function, what we do know, instead, is a set of measurements each of whom is composed by the input value and the associated system response to such input. This collection of data represents the training set which is visualized in Fig. 6.1(b).

As previously described, the *first* stage of training is to compute the location and shape of the Gaussian kernels. A technique of *unsupervised learning* (e.g. k-means) is used to cluster the input data of the training set into subsets containing "similar" data (indicated by the different colors in fig 6.1(c)). Several similarity criteria can be used also depending on which unsupervised learning method is going to be used. Once obtained the clusters, the barycenters of the data composing each cluster are taken as the centers of the Gaussian kernels ($\mathbf{C}_k$) whereas their covariance matrix is taken as Σ_k. The amplitude $\mathcal{Y}_k$, instead, is computed as the average of the corresponding output values pertaining to each cluster k.

Thus, to each cluster a Gaussian kernel is associated which represents a *local* model of the corresponding data (Fig. 6.1(d)). From the same picture it can be observed the local influence of each Gaussian which decreases in radial direction with the distance from the center.

The *second* stage is to compute the optimal combination (i.e. compute γ_k) of these local models according to a global criteria, such that it minimizes the residual error wrt the collected output values (*supervised learning*) obtaining the final result in Fig. 6.1(e) (green curve). Now, an approximation of the searched function is known and it can be used to compute the value of any new sample even if it did not pertain to the training set. What is evident, though, it is the fact that the approximation is poor in that portion of the input space where few or no training samples are available.

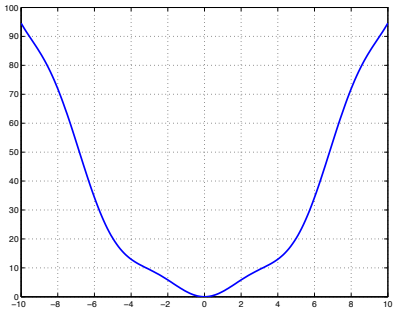
6.1.2 Learning the 3D position from image

In the case of 3D position estimation, the input vector $\mathbf{x}$ is composed by some image descriptors of the instrument and the output values will be one of the 3D coordinates of the tip.

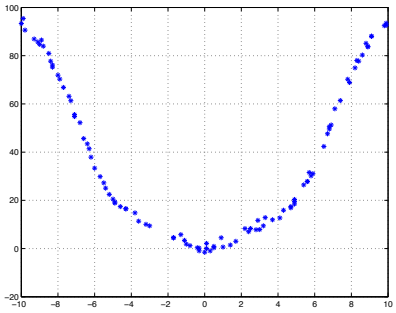
The input space is composed by salient visual features (as state vector components) coming from the visual measurements.

To simplify the learning process, the entire problem has been subdivided into three sub-problems aiming to learn three single-output functions describing the relationship between the image features and each of the coordinates of the 3D tip position. More precisely, the objective is to compute the function relating the $2n$ dimensional column vector $\mathbf{x}$ composed by the $2n$ (x, y) coordinates of the n chosen salient image features to, alternatively, one of the three coordinates of the 3D TCP position i.e. X, Y or Z (depending on the considered case).

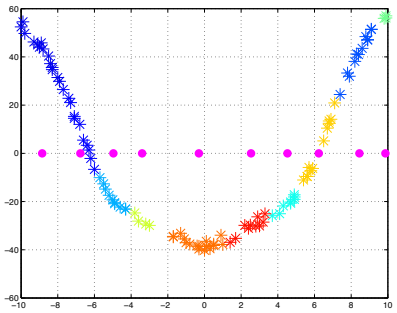
More formally, defining $\mathbf{y} = [X, Y, Z]$, the function h_i of each sub-problem is responsible of computing the estimation of the i-th output variable $\hat{y}_i$ (alternatively



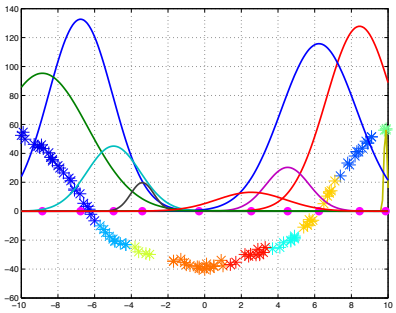
(a)



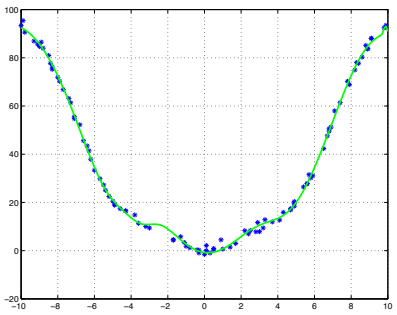
(b)



(c)



(d)



(e)

Figure 6.1: Different steps of the learning process for a single input single output function.

K	Number of clusters
M	Number of elements in the training set
N_k	Number of elements in each cluster: $\sum_k N_k = M$
$\mathbf{C}_{k,i}$	Center of gravity of each cluster k in the estimation problem of the i-th 3D coordinate.
$\gamma_{k,i}$	Coefficient weighting the influence of the k-th kernel for the estimation of the i-th 3D coordinate.
$\mathbf{x_m}$	Elements of the training set $m = 1 \cdots M$
$\mathbf{x_{m_k}}$	Elements of the training set pertaining to cluster k
$\hat{y}_i$	Estimation of the i-th coordinate $y_1 = X$, $y_2 = Y$ and $y_3 = Z$
$\hat{y}_{m,i}$	i-th component of the output sample corresponding to $\mathbf{x}_m$

Table 6.1: Symbols used in the RBF formalization and their meanings.

X, Y or Z):

$$\hat{y}_i = h_i(\mathbf{x}).$$

Consequently, equation (6.1) can be rewritten as follows for each 3D coordinate $\hat{y}_i$:

$$\hat{y}_i = h_i(\mathbf{x}) = \sum_{k=1}^{K} \gamma_{k,i}\varphi_{k,i}(\mathbf{x}) + \gamma_{K+1,i} \tag{6.3}$$

where, this time, $\gamma_{k,i}$ are the learned coefficients that describe the contribution of the k-th kernel to the estimation of the i-th output. The Gaussian kernel (6.2), $\varphi_{k,i}$ takes the form:

$$\varphi_{k,i}(\mathbf{x}) = \mathcal{Y}_{k,i}\, e^{-\frac{1}{2}(\mathbf{x}-\mathbf{C}_{k,i})^T \Sigma_{k,i}^{-1} (\mathbf{x}-\mathbf{C}_{k,i})} \tag{6.4}$$

The training process must, then, be repeated three times to solve the three sub-problems. Therefore, three training sets are needed: each of them contains the M samples composed by the $2n$ dimensional input vector $\mathbf{x}_m$ (with m going from 1 to M) and the corresponding output $y_{m,i}$ that must be learned.

The steps described in the one-dimensional example can be straightforwardly adapted and employed to this case. Firstly, the input data $\mathbf{x}_m$ are divided into clusters: one cluster for kernel. For this aim, K randomly selected $\mathbf{x}_m$ have been taken as initial centers for a k-mean-like algorithm. Once the algorithm converged, for each of the found K clusters the center of gravity, the variance and amplitude are

calculated from the points composing the specific clusters. These values are respectively used as centers ($C_{k,i}$), amplitude ($\mathcal{Y}_{k,i}$) and variance ($\Sigma_{k,i}$) of the Gaussian kernels (cf. equation (6.4)):

$$\begin{aligned} C_k &= \frac{1}{N_k}\sum_{m_k=1}^{N_k}\mathbf{x_{m_k}} \\ \mathcal{Y}_k &= \frac{1}{N_k}\sum_{m_k=1}^{N_k} y_{m_k} \\ \Sigma_k &= \mathbf{X}_k^T\mathbf{X}_k \end{aligned}$$

where $\mathbf{x_{m_k}}$ and y_{m_k} are the $\mathbf{x_m}$ and $y_{m,i}$ samples pertaining to the k-th cluster, N_k is the number of elements of each cluster ($\sum_k N_k = M$), whereas $\mathbf{X}_k$ is the matrix containing the different $\mathbf{x_{m_k}}$ in its columns:

$$\mathbf{X}_k = \left[\mathbf{x}_1|\cdots|\mathbf{x}_{N_k}\right]$$

Secondly, the K coefficients $\gamma_{k,i}$ (cf. (6.3)) are computed so as to minimize the difference between the estimation $\hat{y}_i$ and the corresponding training set values $y_{m,i}$:

$$\gamma_{k,i} = \arg\min \sum_{m=1}^{M}(y_{m,i} - \hat{y}_{m,i})$$

For each 3D coordinate, the problem can be translated in minimizing the following cost function:

$$\chi_i^2 = \frac{1}{2}\sum_{m=1}^{M}\left(y_{m,i} - \sum_{k=1}^{K}\gamma_{k,i}\varphi_{k,i}(\mathbf{x}_m) + \gamma_{K+1,i}\right)^2$$

which, in matrix form, can be written as:

$$\chi_i^2 = [\mathbf{Y_i} - \Phi_i^T\Gamma_i]^T[\mathbf{Y_i} - \Phi_i^T\Gamma_i] \tag{6.5}$$

where $\mathbf{Y_i}$ is the vector containing all the M 3D i-th coordinate output values of the training set, Γ_i is the vector of the $K+1$ regression coefficients

$$\Gamma_i = \begin{bmatrix} \gamma_{1,i} \\ \vdots \\ \gamma_{k,i} \\ \vdots \\ \gamma_{K,i} \\ \gamma_{K+1,i} \end{bmatrix} \tag{6.6}$$

and

$$\Phi_i = \begin{bmatrix} \varphi_{1,i}(\mathbf{x}_1) & \cdots & \varphi_{1,i}(\mathbf{x}_M) \\ \vdots & \vdots & \vdots \\ \varphi_{k,i}(\mathbf{x}_1) & \cdots & \varphi_{k,i}(\mathbf{x}_M) \\ \vdots & \vdots & \vdots \\ \varphi_{K,i}(\mathbf{x}_1) & \cdots & \varphi_{K,i}(\mathbf{x}_M) \\ 1 & \cdots & 1 \end{bmatrix} \tag{6.7}$$

contains the kernels (computed for the i-th 3D output coordinate) evaluated in the distinct input samples of the training set $\mathbf{x_m}$. Each row is corresponding to one kernel k.

Since the cost function (6.5) is linear wrt to the coefficients, the minimization has a closed form solution given by:

$$\mathbf{\Gamma_i} = (\mathbf{\Phi_i}^T \mathbf{\Phi_i})^{-1} \mathbf{\Phi_i}^T \mathbf{Y_i} \tag{6.8}$$

Such a function approximation is efficient as long as enough points are available so as to finely follow the function variations (consider Fig. 6.1(e) in the region where no training point are available). Indeed, the reconstructed function drops to zero as soon as there is no more points to support it [Benoudjit 2003]. Counterbalancing this behavior by increasing the number of kernels would result in less smooth functions with, possibly, large oscillations (with the risk of overfitting). Since RBF are efficient locally and since linear approximation better fits slow and global variations, a combination of both is proposed. Thus, the role that the RBF network assumes is to refine the linear estimation. The training process is totally similar to what described so far with the only difference that the new fitting data for the RBF network are expressed with respect to the linear estimation result. So the new training output data for the RBF network will be:

$$\tilde{y}_{m,i} = y_{m,i} - \boldsymbol{\alpha}^T \mathbf{x}_m$$

where $\boldsymbol{\alpha}$ is the vector of the global linear regression:

$$\boldsymbol{\alpha} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y_i}$$

where $\mathbf{X}$ is the matrix containing the training input samples in its rows.

Gaussian kernel has great versatility in function approximation but, given its shape, it seems to provide a better fit for continuous functions with smooth outline, whereas it seems to be less adapt for piecewise linear or discontinuous functions.

To obtain a good local representation of the data, it must be assured that the clusters used for computing the Gaussian kernels gather samples whose output outline (i.e., in this case, one of the three 3D coordinates of the TCP) is smooth enough to be adequately approximated by a Gaussian. In other words, to achieve a good approximation when performing the clustering on the input domain, we should ensure that proximity and smoothness in the output domain (3D space) is reflected in the input domain (visual feature space). However, this is not always the case when the input samples derive from the perspective projection and this is even amplified due to the peculiar deformation of the instrument.

Let us consider, for example, the case where the instrument is deflected and rotating around the channel axis: the depth wrt to the camera plane only slightly

changes, whereas the input vector deriving from visual features coordinates related to the instrument body (such as the markers centroids or corners) would be subjected to a substantial variation. This leads to the following

Consideration 1. Proximity in the output domain (3D space) is not always reflected in the image feature space.

With the aim of managing this kind of situation and group together all those data pertaining to a smooth 3D manifold, the output should be taken into account in the clustering step. A new *clustering* domain is then defined and it will be described in next section.

6.2 Improved Clustering Domain

In the previous section the conventional procedure to built a network of RBF has been presented pointing out the risks of a clustering performed only on input domain and the consequent necessity of considering the output value in forming the clusters.

Indeed, if we retake the example of a bent instrument rotating around its principal axis and we consider the x coordinate of the last marker as the input of our function and the distance of the instrument tip from the camera as the output, we would obtain the outline described in Fig. 6.2(b). This behaviour would be correctly described by a very wide Gaussian as shown in Fig. 6.2(d). Unfortunately, if the clustering algorithm take into account solely the input domain, there can be the risk of obtaining a result more similar to the one of Fig. 6.2(c). In this case, even the combination of the radial basis functions cannot properly describe the tendency of the function in this portion of the domain.

We then propose to perform the clustering algorithm in a joint input/output domain so that the distance among the samples is influenced also by the output values. The objective is to promote results similar to those in Fig. 6.2(d) where the data which are near in the output domain are gathered together in the same sub-set. In this case, the gaussian describing the cluster will take into account the variance of those new data assuming a wider shape.

Before going on with the discussion, we should distinguish, at this point, between *clustering* and *training* domains. With *clustering domain* is intended the set where the k-means-like algorithm is applied and its dimensionality will be indicated with q_{cl}, whereas the *training domain* will be the one containing the input samples used to retrieve the 3D coordinate estimation and it will be of dimension q_{tr}.

To have also in the clustering domain the same dimensionality as the input, PCA analysis can be carried out so as to take into consideration only those principal components necessary to complete (or not) the output information.

The transformation of the data to this new clustering domain is easily achieved by taking the first q'_{cl} principal components (according to the system DOFs or a more exhaustive cross validation on the training set) and stacking them to the three

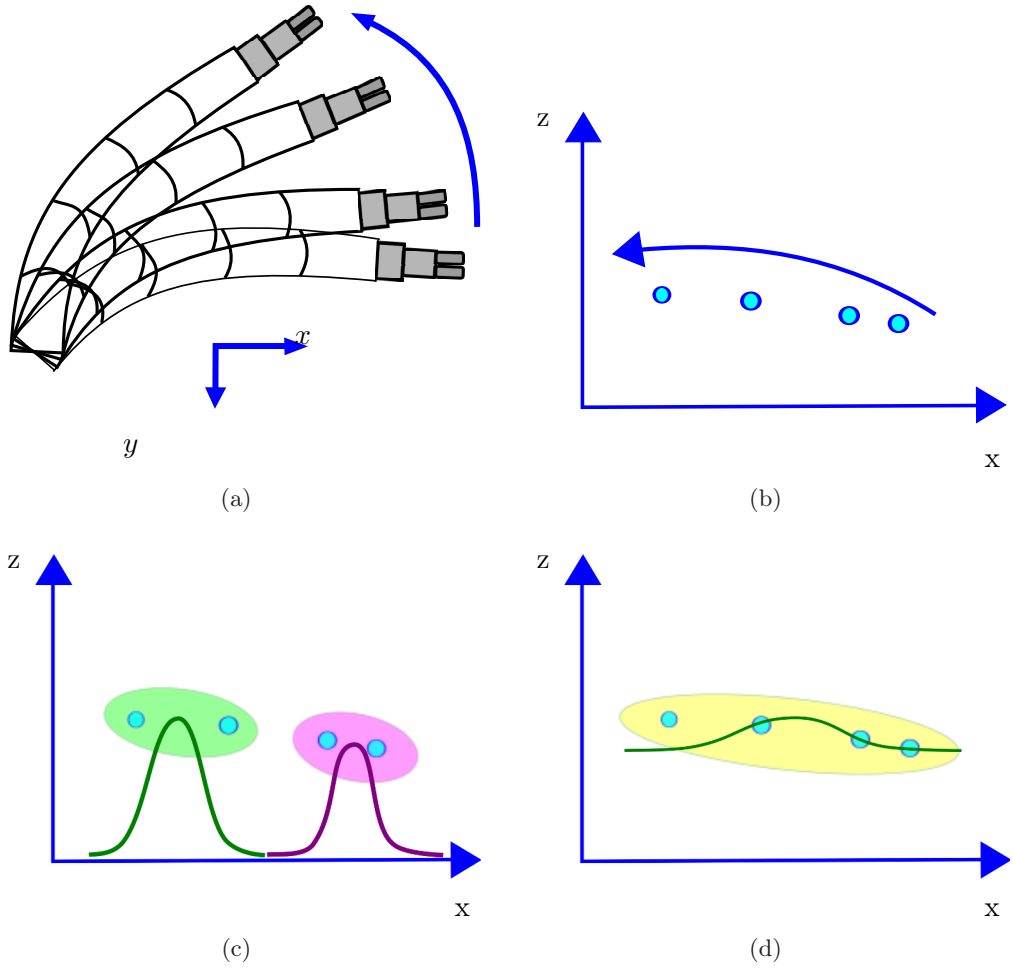


Figure 6.2: Example illustrating the problem related with the initial clustering of the RBF approach. Let us suppose that the instrument is bent and turn around its main axis (a) and that we want to approximate the relationship between the x coordinate of the tip of the instrument and its distance from the camera z (b). If the clustering algorithm (k-means like) is applied considering only the input data (x coordinate), it would lead to a bad estimation (c). If, on the other hand, we take into account also the output data (the z coordinate describing the depth), it will be more probable to obtain a final clustering result similar to the one shown in (d) which leads to a Gaussian whose shape better describes the tendency of the x-to-z relationship.

coordinates ($q_{out} = 3$) of the TCP position. Thus, $q_{cl} = q'_{cl} + 3$. In the next table and section this case will be referred to as $q'_{cl}(+3)$: for example, if the clustering set samples are composed by the first four PCs and the 3 output variables, then it will be denoted as 4(+3).

An alternative could be to take into consideration solely the concerned i-th coordinate of the TCP position ($q_{out} = 1$). In that case, $q_{cl} = q'_{cl} + 1$ and, in the next sections, it will be referred to as $q'_{cl}(+1)$ as described in the precedent paragraph.

To proceed with the clustering in this joint-space, some precautions should be taken. Since the measurement units between the input and output spaces are different, recentering the data wrt their means and normalizing the coordinates by their standard deviation (computed over all the samples) helps to even out the clustering space in the sense that each sample will contribute with the number of standard deviation from its mean.

The new samples, then, can be defined as

$$\begin{aligned} \mathbf{x'}_m &= \Sigma_{x_{PCA}} \mathbf{x}_{PCA_m} \\ \mathbf{y'} &= \Sigma_y (\mathbf{y} - \boldsymbol{\mu}_y) \end{aligned}$$

where $\Sigma_{\cdot}$ is a diagonal matrix whose elements are the reciprocal of the standard deviation of each component (PC or output variable) computed over all the samples and $\boldsymbol{\mu}_y$ is the column vector of the means of each output coordinate.

The new clustering set is composed by the samples formed as follows:

$$\mathbf{x}_{cl_m} = [\underbrace{x'_1, \cdots, x'_{q'_{cl}}}_{\text{input first } q'_{cl} \text{ principal component}}, \underbrace{y'_1, \cdots, y'_{q_{out}}}_{\text{output variables}}]^T_m \tag{6.9}$$

therefore $q_{cl} = q'_{cl} + q_{out}$. On the other hand, the input training set is composed by samples composed by solely input data parameters:

$$\mathbf{x_{tr}} = [x'_1, \cdots, x'_{q_{tr}}]^T_m. \tag{6.10}$$

where q_{tr} can be chosen no matter the value assigned to q_{cl}.

The k-means-like algorithm is applied to the clustering set of samples $\mathbf{x_{cl}}$ which corresponds to a clustering domain of dimension q_{cl}. The indexes of the samples composing each cluster (m) are retained and used to recreate the same clusters in the training space (made by samples $\mathbf{x_{tr}}_m$)) where the centers of gravity are finally computed.

For sake of clearness, it may be useful to reiterate that we want to learn the relationship between the visual measurement (input set) and one of the coordinate of the 3D TCP position (output set) whose value is not directly available. The input training set must be formed, then, by the same kind of samples of the wanted input set (i.e. visual measurement) and $\mathbf{x_{tr}}_m$ cannot contain output information which will be unavailable when testing. For the same region, the cluster computed in the

joint space, must be reformed in the training input space taking into account only the visual measurements.

6.3 RBF: Simulation Result

With the aim to reduce the input domain dimensionality so as to ease the regression process, we consider, as input, the vector formed by the coordinates of the apparent 2D centroids of the colored markers.

Given the difficulty of imagining the shape of such learned function, a simulation study is done to interpret the role of each parameter, the influence of the input dimension and to retrieve more details on the centroids-to-3D-TCP-coordinates relationship.

For this simulation, a synthetic training set has been built. The space of the flexible instrument DOFs ($[\lambda, \phi, \theta]$) is uniformly sampled and, using kinematic and camera model, the coordinates of the markers centroids with the corresponding 3D TCP position is recorded.
A test set has been built in the same way but for different instrument configurations so as to test the learned function over a set of "unseen" positions. In other words, the intersection between the set of the tip 3D position of the training set with that of the 3D positions of the test set is empty.

However, a special attention must be conferred in choosing the training test according to the expected test set so as to avoid extrapolation. Indeed, as already mentioned, RBF network tends to zero in absence of training samples and, therefore, good estimation results cannot be attained for those test samples that fall in a region of the input domain where no training samples are present.
Let us define the *visited workspace* as that portion of the workspace visited by the TCP 3D positions given a set of robot configuration $[\lambda_i, \phi_i, \theta_i]$. To avoid extrapolation, then, the test visited workspace must be included in the training visited workspace which must be dense enough to prevent from the presence of wide empty regions[1].

More specifically, both for training and testing set used in simulation, the same visited workspace is considered and defined by

$$\begin{aligned} \lambda &\subseteq [25, 45] \\ \phi &\subseteq \left[-\tfrac{\pi}{2} + \tfrac{\pi}{10}, \tfrac{\pi}{2} - \tfrac{\pi}{10}\right] \\ \theta &\subseteq \left[\tfrac{\pi}{10}, \tfrac{\pi}{2} - \tfrac{\pi}{10}\right]. \end{aligned}$$

The training set is composed of 280 samples whose corresponding configurations are recovered by taking respectively 5, 8 and 7 equally spread values inside the described intervals. For the test set, the considered robot configuration were defined by taking

[1] Actually, the described problem could be solved by iterative learning and reinforcement learning, but this is not in the scope of this thesis.

set #	# DOFs	considered feature (n)	input dim. ($2n$)	# samples
1	3	3 centroids	6	280
2	3	3 centroids	6	180
3	3	5 centroids	10	280
4	3	5 centroids	10	180
5	6	5 centroids	10	280
6	6	5 centroids	10	180

Table 6.2: Sets used for the simulations considering only 3 or 6 DOFs. Sets 2, 4 and 6 are used to asses the accuracy of the function learned respectively on sets 1, 3 and 5.

5 equally expatiated values for λ, 5 for ϕ and 6 for θ leading to a set of 180 samples. Building the sets in this way allows to obtain a test set with no common elements with the training set assuring, though, that the test visited workspace is included in the training visited workspace (which means that no extrapolation problems will be considered).

A first batch of simulation tests were carried out considering only three DOFs for the instruments (so no variability in the mechanical parameters) and, as visual features, either the three centroids of the yellow markers (set 1 and 2 table 6.2) or all the five centroids (set 3 and 4 of the same table).
Thus, two input spaces were considered: one of six and the other of ten dimensions. Set 1 and 3 in table 6.2 are both derived by exactly the same instrument configurations, the only thing that changes is the input dimension. For set 1 the position of the 3 yellow centroids is considered, whilst for set 3 the input data is the vector of the 5 centroids coordinates. The same discussion is valid for set 2 and 4.

With the scope of presenting results independent by the random initialization, the RMS errors committed over each 3D coordinate are computed considering the estimation results over 20 different random initialization as a unique sequence. This means that the whole training and test process is repeated 20 times with 20 different initialization and the RMS error (given in mm) is computed over 20×180 estimations (in the case where the validation set is either set 2, 4 or 6).
For each simulation test, in addition to this RMS error, the worst 3D error norm for the best initialization (in terms of RMS error) will be given (*Worst 3D err* column in the tables of the next sections).

6.3.1 Cluster Initialization

Due to the particular manner of building the training set, the input samples composing it do not uniformly cover the input domain. Indeed, the uniform discretization used over the robot DOFs does not correspond to a uniform allocation of the marker centroids in the image. It can occur, then, that portions of the input space do not

train: 280 elements; test: 180

K	Avg number of elements per cluster	3D RMS error
5	56	2.3659
10	28	1.9970
15	18.67	1.9
20	15.0556	1.9868

train: 140 elements; test: 90

K	Avg number of elements per cluster	3D RMS error
3	46.6667	2.7574
5	28	2.4408
8	17.5000	2.4512
10	15.1	2.8070

Table 6.3: Effect of the number of cluster over the 3D position estimation accuracy.

contain any sample. That is the reason why the initialization of the cluster centers for the k-means is done by randomly choosing K training set samples rather than opting for a uniform sampling in the input domain with the risk of having empty or not representative clusters.

6.3.2 Choice of K

A first clear result that can be extrapolated from simulation concerns the proper choice of K value which was not discussed so far. The best value for K (in terms of estimation error on the test set) seems to strongly depend on the density and the quantity of samples composing the training set. The choice of K value can be described by the number of elements per cluster, which resulted to be between 15 to 30 per cluster. Analyzing the results, fewer samples per cluster are not enough to have a good generalization and representation for building the Gaussians, whereas too many samples results in an exaggerate smoothing effect.

The study of the best K value has been carried out using set no. 1 for training and the no. 2 to test and the results are shown in table 6.3).

As an alternative solution, a k-fold cross-validation step can be used to determine the best number of clusters according to the available training set. Before starting the training process, the same training set can be divided into, for example, 4 subsets. One of these subsets is retained as the validation data for testing the model trained over the other 3 sub-sets. The cross-validation process is then repeated 3 times (the folds), with each of the 3 subset samples used exactly once as the validation data. The 4 results from the folds can be combined together and the

3D estimation error can be computed. This can be repeated as many time as it is desired according to how precise the value of K is wanted to be know, without the need of creating a validation (test) set.

6.3.3 Input Domain

The interest of study the effect of input domain dimensionality arises from the following

Consideration 2. *Taking into account only three DOFs in building the training and test set, inherently, implies that the shape of the instrument depends only by three factors or three variables. This should be, somehow, reflected also in the image-to-pose function provided that the chosen feature can fully capture the 3D shape. In this case, the domain of the function relating the* $2n$*-dimensional input space with one 3D coordinate (either* X*,* Y *or* Z*) of the TCP should lie on a manifold of dimension 3.*

However, this is not always true, since it depends on the chosen feature and information representation. Most of the time, this consideration would lead to the conclusion that the input domain dimensionality should be *at least* equal to the number of DOFs. Moreover, in certain conditions, spatial localization can become meaningless in high dimensional domains because, typically, with increasing dimensionality all data points tend to have the same distance to each other [Beyer 1999] and this could have side effects on the good convergence of the k-means-like algorithm.

In this case, though, domain dimensionality is not critical but we want to study which is the needed quantity of visual information to have an accurate 3D position estimation. This could also be useful with noisy visual features to isolate only the information useful for 3D position retrieval.

A dimension reduction can be carried out performing a Principal Component Analysis (PCA) over the input samples so as to observe if the first 3 principal components (PCs) are enough to estimate the 3D position.

Consideration 2 seems to be confirmed analyzing the results when using set 1 and 2 (respectively for the training and for the test). In fact, there is no substantial difference between the RMS error using 3 PCs or the whole information given by the three centroids (case 10 and 11)[2].

However, this assertion is contradicted when 5 markers (set 3 and 4) are used instead of 3. In fact, even if case 10 and 20 are comparable, the results improve considerably when q_{cl} and q_{tr} are equal to 6 (cf case 21 and 11). Nevertheless no further improvement is appreciable when the whole 5 centroids information is used (case 22) except for the estimation of Y coordinates which seems to get some

[2]After taking the three PCs of each sample of set 1 and 2, the clustering and input domain are equals between each other but their dimension has decreased from 6 to 3 ($q_{cl} = q_{tr} = 3$).

Case	trained on	tested on	q_{cl}	q_{tr}	3D RMS	Worst 3D err
10	1	2	3	3	0.5914 0.7039 1.5546	6.2484
11	1	2	6	6	0.6312 0.4797 1.7528	5.2821
20	3	4	3	3	0.4384 0.5231 1.3416	6.5967
21	3	4	6	6	0.5971 0.4689 0.8496	3.1330
22	3	4	10	10	0.5446 0.2705 0.7006	3.1936

Table 6.4: Quantitative results of pose estimation accuracy when considering 3DOFs, different clustering and training domains and different visual features. Cases 10 and 11 exploit the 3 yellow markers centroids whereas cases 20, 21 and 22 the all 5 centroids. The worst 3D error value is computed (as for all the other tables) selecting the estimation result deriving by the best initialization in terms of the RMS of the norms of estimation errors.

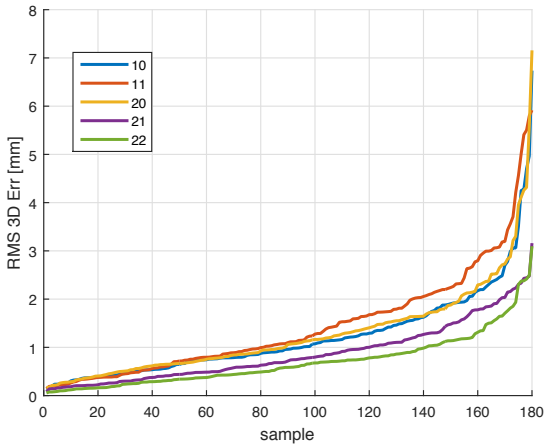


Figure 6.3: Sorted 3D error norms for the cases listed in table 6.4.

advantage from the additional information. The observed behavior suggests a first conclusion:

Conclusion 1. *Other things being equal, the use of 5 centroids brings about some piece of information useful for position and, mainly, for depth estimation.*

Furthermore, comparing cases 20 and 21 one can observe that 3 components are not sufficient to get the best position estimation result. This is may due to the fact that PCA is not the right choice to reduce dimensionality in this case. Indeed, directions with high variance become a principal component even though such direction has no influence or relation with the output.

6.3.3.1 Mechanical Parameters Variations

As already analyzed in the precedent chapter, considering only 3 DOFs does not allow to represent all the kinematics of the instrument. Mechanical play between instrument and its housing channels must be taken into account since they influence the appearance of the instrument itself in the image.

If the effect of the mechanical plays want to be considered, the training and test sets must be modified taking into account 7 DOFs.

There are many possibilities to extended the proposed 3 DOFs sets. The first can be to consider that the mechanical parameters variations are totally random. In such a case, to create the synthetic data set, it is sufficient to perturb the mechanical nominal values before computing the 3D instrument pose with the kinematic model.

If. instead, the mechanical DOFs are considered as real "controllable" DOFs, a new workspace should be defined based on these 7 DOFs. The training and test sets, then, would be composed as set 3 and 4 by uniformly sampling the 7 DOFs.

A third possibility, the one employed here, is to consider that the position and orientation of the instrument in the channel depends on the particular instrument configuration. In this case, similarly to the first option, the values of the 4 mechanical parameters are computed before retrieving the entire pose with the kinematic model (and, finally, the synthetic image by the camera model). Thus, the new training and test set (indicated respectively as set 5 and 6) are computed in the exact same manner as for sets 3 and 4, the only different aspect is that the mechanical parameters are assuming variable values depending on the specific configuration $[\lambda, \phi, \theta]$:

- $x_{ch} = x^*_{ch} + 3 * sin(\phi)$
- $y_{ch} = y^*_{ch} + 2 * sin(\phi)$
- $\psi = \psi^* + \pi/90 * sin(\theta - \pi/4)$
- $\mu = 0$: as for the model-based simulation tests, is considered fix and equal to 0.

Case	trained on	tested on	q_{cl}	q_{tr}	3D RMS	Worst 3D err
30	5	6	3	3	0.5369 0.8704 2.9192	9.0981
31	5	6	6	6	0.6423 0.8574 2.0556	5.1269
32	5	6	7	7	0.5658 0.8904 1.3337	4.6466
33	5	6	10	10	0.3545 0.3829 1.3150	3.8549

Table 6.5: Tendency of RMS and worst 3D estimation errors using RBF Network according to the clustering (q_{cl}) and training (q_{tr}) domain dimensionality. The highest accuracy is achieved when exploiting the entire visual information $q_{tr} = 10$ (which corresponds to the 2×5 2D coordinates of the markers' centroids).

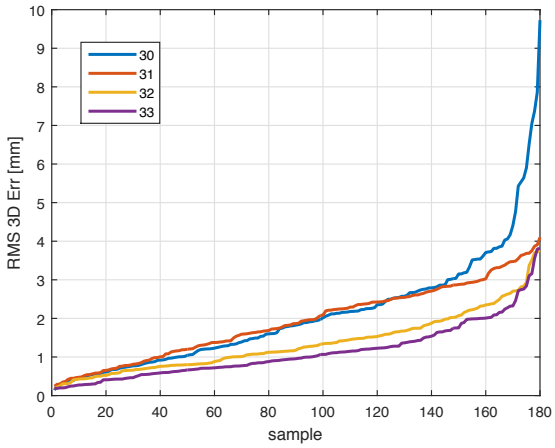


Figure 6.4: Sorted 3D error norms for the cases listed in table 6.5.

Obviously, we have no guarantee that these functions describe the real relationship between the instrument configuration and the mechanical parameters, but they allows to implement a more realistic set that takes into account the mechanical plays.

Consideration 2 may be referred also to this case with 6 DOF[3]. Therefore, 3 PCs does not seem to be sufficient to have a good estimation. This is confirmed also by simulation results (cf. table 6.5), where six or even better seven PCs (cases 31 and 32) already give an acceptable result even though augmenting the clustering and training domain dimensionality (i.e. using more visual information) seems to increase also the estimation accuracy.

However, the achieved precision is not satisfying considering that no noise has been added yet to the data. The possible reasons could be that the shape of the kernel is not correctly computed (maybe due to an imperfect clustering) or that the chosen kernel are not suitable to approximate the function. These two problems are now addressed and possible solutions are proposed.

[3]Let us remind that the fifth mechanical DOF is considered fix and equal to its nominal value.

Case no.	trained on	tested on	q_{cl}	q_{tr}	3D RMS	Worst 3D err
31	5	6	6	6	0.6423 0.8574 2.0556	5.1269
40	5	6	3(+3)	6	0.5281 0.7827 1.7625	5.1788
41	5	6	5(+1)	6	0.5910 0.8719 1.7874	5.5241

Table 6.6: Average and maximum errors (cases 40-41) for two different clustering domains: considering the first 3 input PCs and the whole output ($q_{cl} = 3(+3)$) or the first 5 input PCs and the output variable that is wanted to be estimated ($q_{cl} = 5(+1)$). These strategies seem to slightly benefit the average accuracy above all in Z (wrt to case 31) supporting the hypothesis that such clustering domains help gathering together those data whose output are better represented by a Gaussian.

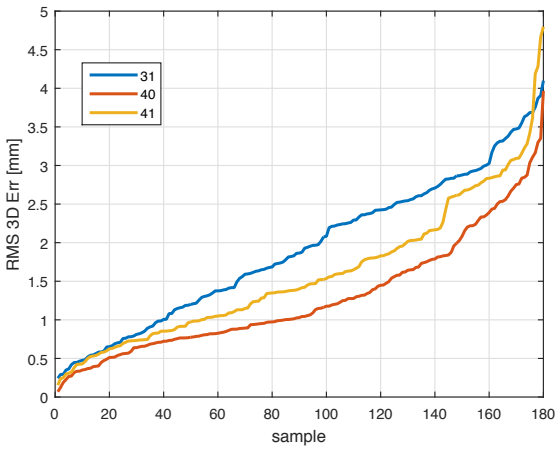


Figure 6.5: Sorted 3D error norms for the cases listed in table 6.6.

6.3.3.2 Clustering the data

To improve the computation of the Gaussian shapes, consideration 1 has been taken into account and, consequently, decided to carry out the clustering algorithm in the input-output joint space obtained as described in Sec. 6.2.

The adoption of this alternative clustering space seems to bring only small improvements in terms of 3D estimation accuracy (cf. tab. 6.6). In fact, the clustering and training domain dimension being equal (case 40, 41 and 42 have all $q_{cl} = q_{tr} = 6$), the 3D RMS error and even the worst 3D err are comparable. Nevertheless, the result over Z seems to improve when considering the output (either the 3 coordinates or the specific coordinate), confirming somehow the adequacy of such clustering domain.

This leads to the following consideration

Consideration 3. *It seems that the output carries some information that can substitute (or even improve) the input principal components.*

This fact can be exploited in those cases where the confidence on the output

of the training set is higher than the confidence on the input samples (the image features in this case). One can think to substitute the input information with the output without any substantial loss of accuracy in the estimation.

Consideration 1 does not only suggest a clustering strategy but, also, that the RBF Gaussian kernel can be inadequate to describe the 2D-to-3D Z-coordinate relationship. In fact, considering again the case of the instrument rotating with a fix deflection, the function describing the relationship between the centroids and the Z 3D coordinate would appear as an hyperplan in the $2n+1$ dimensional domain.

That is why another approach has been investigated where another kernel such as hyperplanes is used instead of Gaussians. Directly substituting the Gaussian kernels by planes would result in a piecewise linear function which presents strong discontinuities in correspondence of the inter-clusters frontiers.

Another approach, then, could be investigated. We propose to use Locally Weighted Regression which are described in the next section.

6.4 Locally Weighted Regression (LWR) Method

To understand how LWR works, a step backward is necessary to introduce linear regression. Assuming, without loss of generality, that the input $\mathbf{x}$ and output y_i are zero-mean, the model function in linear regression is:

$$y_i = \gamma^T \mathbf{x} + \varepsilon_y \tag{6.11}$$

where $\mathbf{x}$ is the q_{tr}-dimensional input vector, y_i is the output value, γ are the regression coefficients and ε_y is a noise variable independent of $\mathbf{x}$.

Furthermore, let M be the number of training data points such that $\mathbf{X}_{tr} = \{(\mathbf{x_m}, y_{m,i})\}_{m=1}^{M}$. The coefficients γ can be obtained by minimizing the squared error χ^2:

$$\chi^2 = \frac{1}{2}\sum_{m=1}^{M} ||y_{m,i} - \gamma^T \mathbf{x_m}||^2 \tag{6.12}$$

which results in the conventional least square solution:

$$\gamma = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}_i$$

where the rows of $\mathbf{X}$ are the input vector of the training set (image centroids coordinates) and $\mathbf{y_i}$ the corresponding output values (one of the three space coordinate).

Locally Weighted Regression is funded on the principle that each data of the training set should contribute to the computation of the linear regression proportionally to its distance from the point where the regression has to be computed (also called *query point*) [Atkeson 1997].

Even though all the data of the training set are used to compute such regression, the term "locally" stays for describing the fact that each term in (6.12) is weighted

according to the distance from the query point $\mathbf{x}_q$. Therefore, the new cost function has the form:

$$\chi^2 = \frac{1}{2}\sum_{m=1}^{M} w_m(\mathbf{x}_q)\,||y_{m,i} - \gamma^T \mathbf{x_m}||^2 \tag{6.13}$$

and the weights can be expressed as

$$w_m(\mathbf{x}_q) = \exp\left(-\frac{1}{2}(\mathbf{x_m} - \mathbf{x}_q)^T D(\mathbf{x_m} - \mathbf{x}_q)\right) \tag{6.14}$$

where D is a positive semi-definite matrix computing a distance metric that determines the region of influence of the so-defined local model (but that still depend on all the $\mathbf{x_m}$).

In this case, the solution for the regression coefficients is:

$$\gamma = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y} \tag{6.15}$$

where $\mathbf{W}$ is a diagonal matrix containing the w_m on its diagonal.

Recalling the one-dimenional example used for RBF, in the case of LWR, when the estimation for a new sample has to be computed (red line in Fig.6.6(a)), a weight is assigned to all the data of the training set according to (6.14) (resulting for example in the distance metrics of Fig. 6.6(b)) so as to find the "local" version of the linear regression coefficients (cf. eq. (6.15)) obtaining the local approximation (green line of Fig. 6.6(a)). This kind of learning method is also named Memory-Based learning since it needs the entire training set to estimate a new sample.

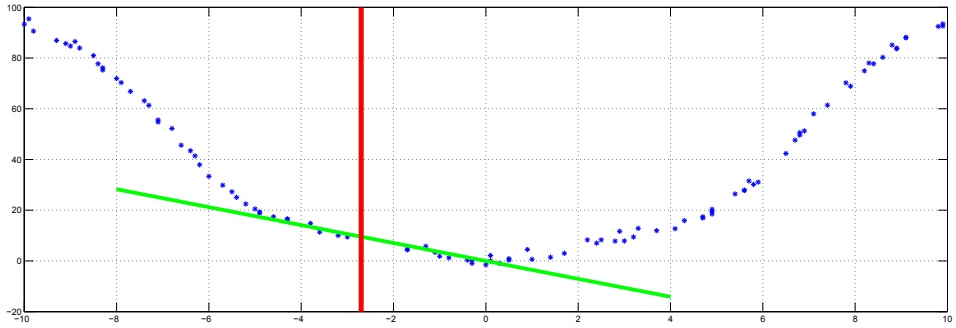
The two main problems of such approaches lie

- in the inversion of the covariance matrix eq. (6.15), which cannot handle redundant information,
- and the high computational cost due to the weights computation, which has to be done for all the data of the training set.

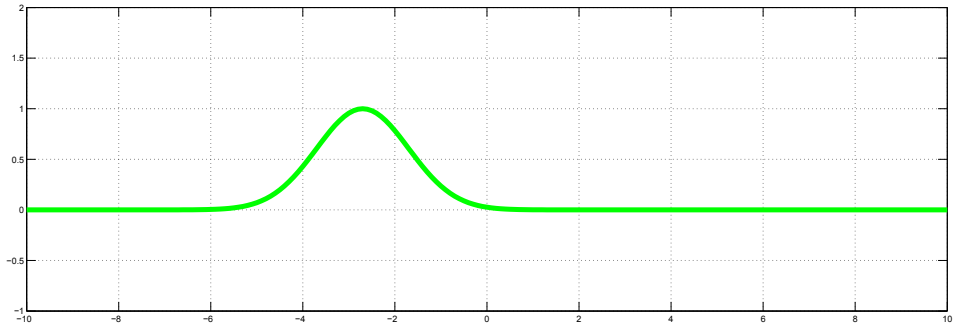
To solve the described problems, the dimensionality of the training domain can be reduced with PCA (to avoid the use of redundant information which may bring to singularities) and the weights computation may be limited to the first j nearest neighbors of the query points to lighten the whole process.

Taking into account considerations 1 and 3, our proposal is to take into account the output in computing the nearest neighbors. Since no information on the output is furnished for the query point, no input-output joint distance can be computed for deriving the weights. A way to consider the output is to pre-cluster the training data in the input-output joint space as described in sec. 6.2.

In the training space (defined as (6.10)), a distance metric can be defined and, consequently, the distance from the query point (expressed wrt to its principal components) to the clusters centers can be computed. The 2 nearest clusters are selected and the weights are computed only for the sample composing such clusters.



(a)



(b)

Figure 6.6: Principle of the LWR approach on a one-dimensional example. For a new query point, the weights associated to the all training set samples are computed according to a distance metric (such as the one in (b)) and with respect to the query point in question. After that, equation (6.15) to obtain the linear approximation valid only in a neighborhood of the query point.

Case no.	trained on	tested on	q_{cl}	q_{tr}	3D RMS	Worst 3D err
31	2	6	6	6	0.2320 0.3156 0.7800	2.5978
32	2	6	7	7	0.2003 0.2301 0.6230	1.4606
33	2	6	10	10	0.1496 0.1334 0.5665	1.4581
40	2	3	3(+3)	6	0.2964 0.3597 1.0054	2.9219
41	2	3	5(+1)	6	0.3453 0.4231 1.1592	3.8243

Table 6.7: LWR Approach: simulation results using 6 DOFs and 5 centroids and various clustering and training domain dimensionality.

The objective of this pre-clustering is to group the data according also to the output values and transfer, somehow, this output influence in defining the query point neighborhood where no output information is available.

However, consideration 1 was more related with the Gaussian shape and the necessity of clustering together smooth data which should be better represented by the local Gaussian models. The choice of using planes for the regression has its origin in trying to deal with cases similar to consideration 1 and more specifically the case of centroids-to-Z 3D coordinate relationship. If this is the case, the effect of considering the output would not be as much important as for the RBF Network. Output consideration, though, can be useful in the perspective of consideration 3 to "substitute" noisy data with the possibly more trustful output data.

6.4.1 Simulation Results

As before, the LWR-based method is tested in simulation and actually the achieved accuracy is better than that achieved with RBF Network. The same tendency as before reappears here showing that the higher the dimensionality the better the performance is (see Tab. 6.7 and Fig. 6.7).

The results here seems to be in contradiction with the previous assertion about the need of dimensionality reduction. However, in case of higher redundancy (using 3 DOFs and a training and clustering set of dimension 10) the described singularity problems arise more often.

We can observe, though, that considering output in the clustering domain does not improve the estimation precision as in the other case.

6.5 Counteracting the noise

The effect of noise deserves a different section since it changes the results observed so far above all for RBF and conclusions are less intuitive and difficult to generalize.

A first result is over accuracy. Indeed, as one may expect, when considering input noise (which is unavoidable in image), the position estimation accuracy drops.

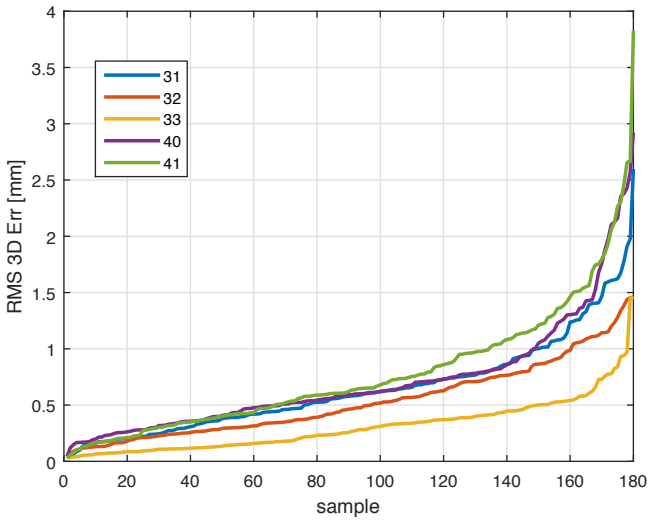


Figure 6.7: Sorted 3D error norms for the cases listed in table 6.7.

Case	trained on	tested on	q_{cl}	q_{tr}	3D RMS	Worst 3D err
50	5	6	6	6	RBF 1.1411 1.5936 3.5830	11.0444
					LWR 0.9030 1.3017 2.3512	7.6843
51	5	6	10	10	RBF 1.2365 1.7560 4.0120	12.2903
					LWR 0.9993 1.4225 2.6475	8.1172
52	5	6	3	3	RBF 0.8489 1.4579 3.2583	9.2711
					LWR 0.8276 1.2487 2.2021	7.1378

Table 6.8: Simulation test results considering image noise (and consequent error on visual feature extraction) both in training and test set. In this case, higher training and clustering domain dimensionality does not imply a higher accuracy.

Moreover, with the introduction of noise, augmenting the dimension does not coincide any more with increasing the estimation precision, on the contrary it decreases when passing from 3 to 6 PCs to 10 (cf. cases 50, 51 and 52). This fact is appreciable above all looking at the worst error in Tab. 6.8.

In this simulation tests (where both image features and TCP 3D position are computed using camera and kinematic models), a uniformly distributed random noise between 0 and 5 pixels was added to each feature point (both to train and test set input) whereas no noise were added to the TCP position.

To counteract the effect of noise we propose two strategies. The first one is to increase the trustworthiness of the clustering data by relying on more trustful data and the second one is to "learn" the effect of the noise.

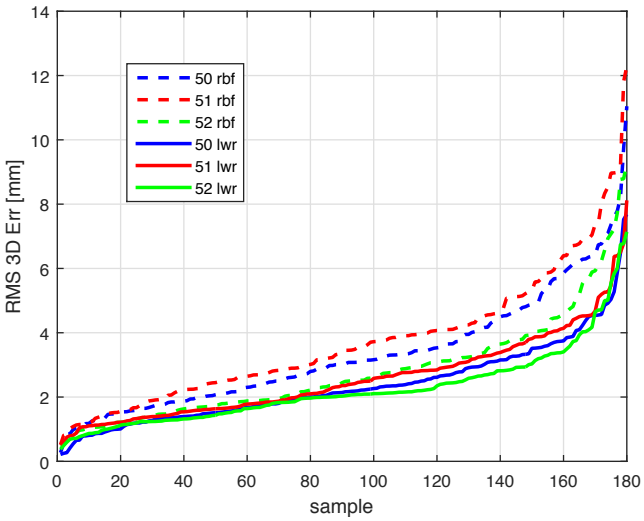


Figure 6.8: Sorted 3D error norms for the best initialization of the cases listed in table 6.8.

Improved Clustering

For the *first solution*, consideration 3 provides already a possible trustful information to use i.e. the output value. In fact, the output value is a measurement issued from a stereoscopic system which usually furnish very trustful results (provided that the cameras intrinsic parameters are known). Therefore, using the same framework described beforehand, the clustering space would be composed by the first principal component of the input (image feature) vector and the output (3D coordinate) variables. Moreover, the fact of using PCA relegate the noise contribution to the last component allowing to eliminate a part of the noise by taking only the first components (cf. cases 50 and 52).

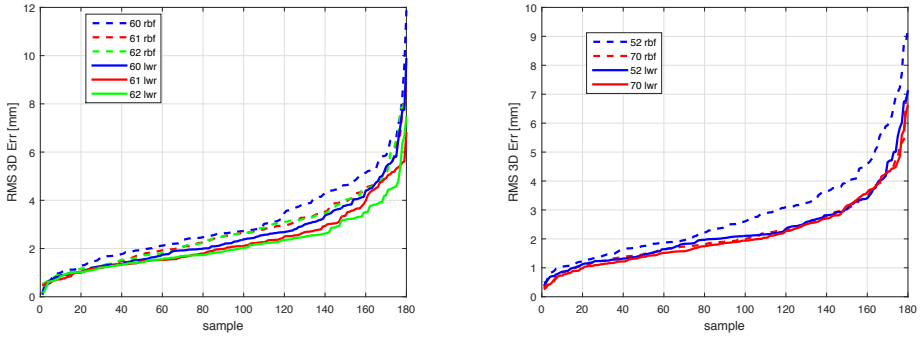
This correction improves the results for RBF when considering all the output values. On the other hand, the same case ($q_{tr} = 3(+3)$) it seems to have almost no effect for LWR, even though a slight improvement in the maximum error is noticeable (cf. cases 60, 61, Tab. 6.9 and Fig. 6.9(a)).

This capability of noise reduction when using the output is confirmed also when comparing the results obtained using the output space as clustering with those obtained using the 3 input PCs (Fig. 6.9(b) and Tab. 6.10). Also in this case the effect is more appreciable for the RBF approach leading to the conclusion that planes may better represent the image-to-3D function being capable of filtering outliers.

The results deriving from the consideration of the output alone imply suggest that the benefits go beyond the noise counteraction especially for RBF. It seems that the output information is better for clustering purpose than the information carried by input principal component. The simulation results show that the accuracy that can be achieved with a lower clustering and training dimensionality (case 70,

Case	trained on	tested on	q_{cl}	q_{tr}	3D RMS	Worst 3D err
60	5	6	5(+1)	6	RBF 1.0931 1.6832 3.3868 LWR 0.9189 1.3978 2.7112	11.8784 8.1118
61	5	6	3(+3)	6	RBF 0.9928 1.4646 2.7167 LWR 0.8304 1.3211 2.3229	8.5503 6.8322
62	5	6	4(+3)	7	RBF 0.9571 1.6823 2.8944 LWR 0.8460 1.3820 2.1930	8.6669 5.4759

Table 6.9: Effect of the image noise on the accuracy of the 3D position estimation for various clustering domains. The output, which is supposed to be less affected by noise, can be used to build the clustering set to improve the Gaussian shape and, consequently, the position estimation. The same effect is less evident when using LWR whose shape seems to be more adequate for this regression problem.



(a) (b)

Figure 6.9: Sorted 3D error norms for the cases listed in table 6.9 in (a) and table 6.10 in (b).

Case	trained on	tested on	q_{cl}	q_{tr}	3D RMS	Worst 3D err
70	5	6	0(+3)	3	RBF 0.8247 1.435 2.4534 LWR 0.8358 1.3471 2.2032	7.0378 6.6369
52	5	6	3	3	RBF 0.8489 1.4579 3.2583 LWR 0.8276 1.2487 2.2021	9.2711 7.1378

Table 6.10: Comparison of accuracy results considering 3-dimensional clustering and training domains. When using the 3D output samples as clustering set, the global accuracy seems to improve above all for RBF network suggesting a higher reliability of the output GT information wrt to the visual information which are more corrupted by noise.

90	5	6	6	6	RBF 0.8530 1.4258 2.6253 LWR 0.8275 1.3685 1.9785	7.1746 6.6369
91	5	6	3(+3)	6	RBF 0.7290 1.3650 2.0916 LWR 0.8244 1.3964 2.1033	6.7925 6.0101
92	5	6	5(+1)	6	RBF 0.8207 1.4104 2.3777 LWR 0.8685 1.4210 2.4416	6.7114 7.1330

Table 6.11: Accuracy results with domain enrichment to "learn" the noise effect on the image centroids. Comparing them with cases from 50 to 52 (Tab. 6.8), such enrichment seems to help the position estimation precision above all for RBF Network and also in terms of maximum 3D error. As for the other cases, the maximum 3D error is computed over the best initialization in terms of RMS error.

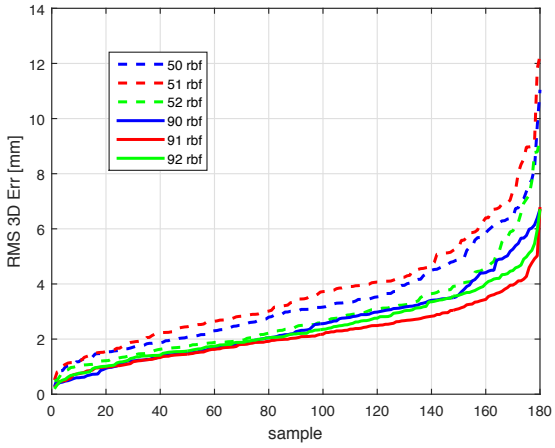
Tab. 6.10) can be as good as (or even better) the one obtained considering higher dimensionality (e.g. case 62, Tab. 6.9).

Training Set Enrichment

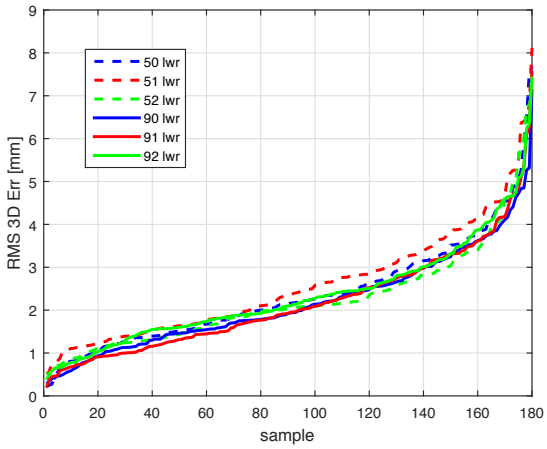
The *second proposal* is a kind of *training set enrichment*. The idea is to "teach" to the regression machine that the same output can be obtained with different input configurations due to noise effect. For this aim, the cardinality of the training set can be increased by adding for each original training sample a new sample with the same output value (as the original one) but with the input value corrupted by the expected noise e.g. the error in extracting the visual feature. This average error can be determined empirically analyzing a sequence of images. The benefit of such enrichment is evident comparing result in Tab. 6.11 with cases 50, 51 and 52. The positive effect is more evident for the RBF Network (cf. Fig. 6.10(a)) even though a (more slight) improvement can be noticed also for the LWR (Fig. 6.10(b) and (c)).

The other objective of the training set enrichment is to adequately add samples in the case where the training set is not dense enough. At the beginning of the chapter and also in Sec. 6.1 has been underlined how local learning approaches need a representative and "dense" training set to avoid to drop to zero if no data are present. Cases 100 and 101 (Tab. 6.12) confirm this aspect. In case 100, an uniformly sampled subset of set 5 is used for training (subset of cardinality 140) whereas the test is always done on test 6 (180 samples). In case 101, domain enrichment is used to "complete" the training set finally obtaining a slightly better result (cf. table 6.12).

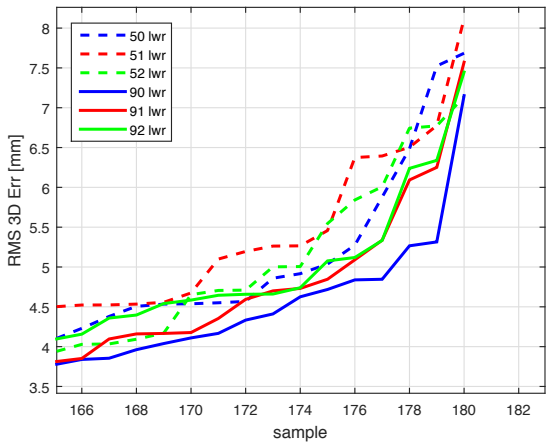
The simulation results show that the two proposed improvements increase the average accuracy of both RBF and LWR methods even when training set is affected by noise. However, the two methods do not response in the same way to the same



(a)



(b)



(c)

Figure 6.10: Sorted 3D error norms for the cases listed in table 6.11 and comparison with cases from 50 to 52 where no domain enrichment was performed. In (c), a zoom on the last samples of (b) is shown.

100	5	6	6	6	RBF 1.1329 1.6025 3.8530 LWR 0.9342 1.3813 2.8512	10.6994 7.8852
101	5	6	3(+3)	6	RBF 0.9778 1.6017 2.7073 LWR 0.8923 1.4183 2.1218	9.3821 6.5195

Table 6.12: Effect of domain enrichment used to increase the cardinality of the training test. In case 100, only 140 samples are used to train the regression machine which is tested on a sequence of 180 samples. In case 101, the training domain density is increased as in the case of the domain enrichment and output is used to form the clustering set.

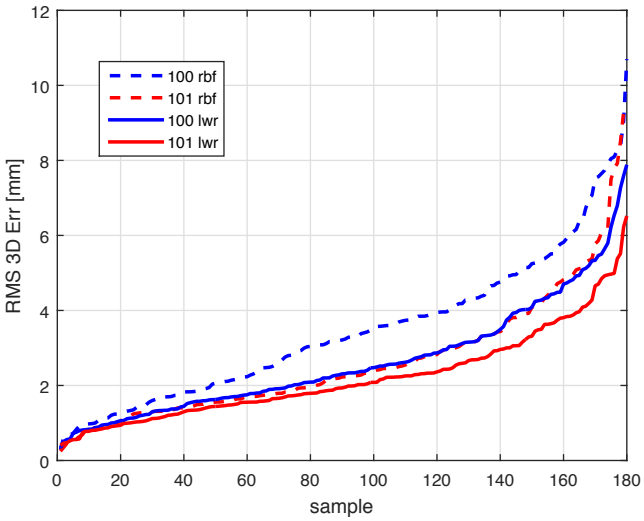


Figure 6.11: Sorted 3D error norms considering the best initialization for the cases listed in table 6.12.

action.

According to simulations where noise is taken into account, joint cluster domain brings a higher benefit on RBF than LWR-based method (cf cases 52 and 70). This may be explainable reminding consideration 1. In fact, LWR is already adequate to fit (hyper-)planes to the data, whilst, adding output shape information obliges, somehow, the Gaussian to take a flatter shape in the case of depth estimation.

These simulation experiments suggest three interesting cases to test with the data retrieved from the robotic experimental cell:

- $q_{cl} = 3(+3)$, $q_{tr} = 6$
- $q_{cl} = 6$, $q_{tr} = 6$
- $q_{cl} = 0(+3)$, $q_{tr} = 3$

always considering the training enrichment.

6.6 Experimental Results

As emerged from the preceding discussion, the needed data set for the RBF network training is composed of the image coordinates of the five markers centroids and the corresponding measurements of the 3D position of the tip of the instrument.

The method described in Ch. 4 can be directly applied for extracting the 5 centroids of the markers by using the nodes of the resulting directed path (without the need of border and corner detection).
The GT of the 3D position is obtained using the same experimental setup as detailed in appendix B.

With the aim to compare model-based and learning-based methods the same set is used here to perform the experimental validation.

In this case, a 4-fold cross validation process is adopted for obtaining the quantitative results of the estimation error. The available data are partitioned into 4 subsets whose components are randomly chosen. The train/test process is repeated 4 times: each time 75% of the data (three packages) are used for the training and the remaining 25% for performing the test. The RMS errors are then computed considering the errors over all the 4 testing packages (i.e. over the whole available data set).

In the precedent chapter the mechanical parameters were modeled as additional DOFs. However, provided that no contacts occur between instruments and environment, these are "passive" DOFs and, consequently, one can think that their values are associated to the particular configuration of the robot. The simulations were carried out along the line of this reasoning and, based on that, the best result parameters have been chosen for the experimental tests i.e. $q_{cl} = 3(+3)$ and $q_{tr} = 6$.

With such choices, the result is poor in terms of RMS estimation error (computed over 237 values):

q_{cl}	q_{tr}	RMS err	Worst 3D err
0(+3)	3	RBF 0.8295 1.1266 2.6353	9.5452
		LWR 0.8539 0.802 2.0959	7.0924
3(+3)	6	RBF 0.8274 1.2193 2.8493	9.089
		LWR 0.747 0.7019 1.803	6.1562
6	6	RBF 0.9285 1.5184 3.0284	8.6384
		LWR 0.7749 0.7634 1.6569	5.3356

Table 6.13: RMS and maximal errors computed for the two different strategies (RBF and LWR-like approaches) for an in-laboratory sequence of 237 configurations with input domain enrichment. The best results for RBF are obtained considering the output in the clustering domain ($q_{cl} = 0(+3)$), whereas LWR is capable of approximating the function only with input information ($q_{cl} = 6$).

- for (Pre-Clustered) LWR: $[0.815, 1.03, 2.4] \pm [0.54, 0.73, 1.57]$
- for RBF: $[1.02, 1.98, 4.15] \pm [0.65, 1.28, 2.57]$

but an important improvement is noticed when the training set enrichment is performed by adding 3 (other values can be used as well) samples for each original sample. Doing so, the obtained RMS errors are:

- (Pre-Clustered) LWR: $[0.7, 0.74, 1.74] \pm [0.43, 0.49, 1.08]$
- RBF: $[0.92, 1.2, 2.49] \pm [0.61, 0.72, 1.49]$

As also explained before, the choice of the value of parameter K changes according to the cardinality of the training set. The results shown for the original training set are obtained with $K = 10$ and for the case of the enriched training $K = 18$.

As announced in the simulation section, other experiments has been carried out which are summarized in table 6.13. These last results confirm the expectation of the simulation tests. Indeed, LWR is definitively better than RBF in approximating the 2D-to-3D function above all in terms of depth retrieving. Even if output information is not considered in the clustering domain ($q_{cl} = 6$), a good approximation of the Z coordinate is achieved thanks to LWR endorsing, somehow, the consideration 1 concerning the particular almost planar shape of the 2D-to-Z-coordinate function.

Comparing the experimental results with those obtained in simulation, the tendency of the RMS error in the different cases (clustering domain, training set enrichment, ...) is almost maintained (Fig. 6.12) whereas the RMS value are lower for the experimental RMS. This difference in value can be explained by the fact that, maybe, the hypothesis made over noise magnitude (both for input or output measurement) or over the modeling of the mechanical DOFs are only partially true and maybe exaggerate wrt to reality.

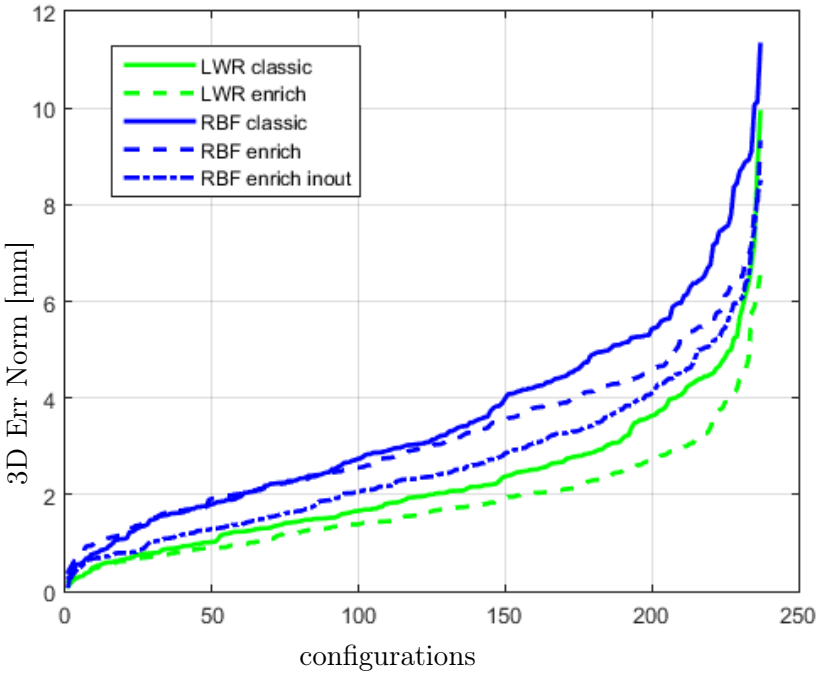


Figure 6.12: Effect of the two improvement proposed for the learning based method. On the graphic, we can observe that the training set enrichment improves the quality of the position estimation obtained by the RBF (cf. curves *RBF classic* vs *RBF enrich*) and by LWR (cf. *LWR classic* vs *LWR enrich*). Finally, in the case of RBF, realizing the clustering in the input/output joint domain (curve *RBF enrich inout*) provides better results than those obtained with only the training set enrichment.

Comme montré en figure 6.12, ces deux modifications apportent des bénéfices aux deux approches en améliorant la précision de l'estimation de la position 3D du bout de l'instrument.

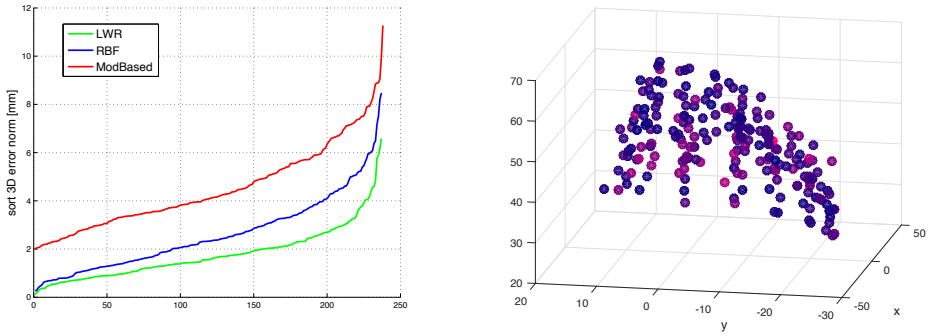
6.7 Comparison with Model-Based Method

Comparing the root mean square errors of the model-based and learning-based methods suggests that the latter approach seems to be more capable of capturing the intrinsic nature of the image-to-TCP relationship. The same tendency is confirmed by the maximum error which is almost 3 mm worse in the case of model based (Fig. 6.13(a)). In terms of maximum error, though, both approaches are still far from the possibility of being used in the envisaged biomedical applications.

Differently of what observed with the model based approach, here, there is not a clear pattern in the distribution of the highest 3D errors in the 3D space. The largest error are on isolated points and are usually derived from the fact that no dense information is available in that specific portion of the space.

The substantial advantage of learning approach is that it does not require to model all the different relationships of the considered system. This allows to reduce the errors deriving from incomplete models, which often result to be incorrect due to uncertainty or identification issues. This can also avoid camera calibration step and camera distortion estimation since these procedures are inherently considered in the learning process. In fact, the coordinates of the features extracted in the original images and used during the training do not need any rectification. The 2D-to-3D function, then, is retrieved directly with the distorted information. This does not imply any problem as far as the image distortion is the same for all the sequences where the trained regression is wanted to be applied. Furthermore, the needed features (markers centroids) are easier (and faster) to extract wrt to markers corners.

On the other hand, learning-based approaches are strictly dependent on the ground-truth they have been trained on: the GT must be trustful and complete to achieve satisfactory precision and its obtaining is often laborious and time-demanding. If, during test, a "not seen" sample shows up the trained machine cannot give any trustful estimation. This is also confirmed by the fact that the largest errors are on the frontier of the training set i.e. where the information is less dense (cf. Fig. 6.13(b)). In this aspect, a model-based approach beats the learning-based because it does not need to *see* any specimen and, nevertheless, thanks to model hypotheses, can infer pose information even if the information is only partially available. Furthermore, it shows a more adaptability with different systems (in Ch. 5 the same model worked for experimental cell and manual flexible system used *in-vivo*), whereas in the learning based method another training set should be created.



(a) (b)

Figure 6.13: The norm of the 3D error committed with the different approaches is sorted and shown in (a) (red: Model Based, blue: RBF and green: LWR). In (b) a color code is used to highlight the norm of the error in the workspace. The more red they are the higher is the error committed in its estimation. The highest errors seem to be located at the borders of the training set.

Considering a scenario where the instruments interact with the environment (tissues, organs, ...) a model-based approach would allow to still interpret the scene, with the only necessity of upgrading the model accordingly to the deformation that have to be taken into account (e.g. considering more DOFs for being able to estimate non-constant curvature configurations or additional bending sections). From the learning point of view, a complex training action should be performed if all the possible deformations have to be taken into account.

Since the model-based approach estimates the best joint configuration, it allows to estimate the entire pose of the instrument which could be useful for motion planning purposes or if the TCP orientation has to be controlled. This is an important aspect in surgery, since cutting or sewing action generally requires a particular tool orientation to perform the most adequate gesture and, then, attain the optimal result. The approach adopted here in the learning-based solution only allows to have the TCP position estimation. It could be thought to extend the proposed solution considering orientation data. However, a precise orientation GT seems difficult to obtain. With respect to this aspect several attempts were made with electromagnetic sensors but their reliability is limited and even drops when interact with metallic parts (which are difficult to avoid in endoscopic surgical contexts).

Chapter 7

Conclusion and Future Work

Contents

7.1 Conclusion

The control of continuum robots is a challenging task and it becomes even more challenging in the context of surgical operations carried out with flexible miniaturized instruments. Without accounting for the difficulties linked to *in-vivo* operations, the complex models needed to describe their dynamics lead to the necessity of external measurements of the instrument pose to close the feedback position control loop. With this final aim, we have proposed and compared in this manuscript, two *vision based* methods for measuring the 3D position of flexible instruments.

The problem has been tackled in all its stages: from camera and flexible endoscopic system modeling, through the selection and segmentation of suitable visual features and to the inferring of 3D information from such 2D features.

7.1.1 Chapter 3

The whole kinematic chain from the camera till the TCP of the instrument has been identified and expressed in the camera reference frame. Based on this formalization, a theoretical study has been carried out over different descriptors to find the most suitable (among them) to infer 3D information from monocular vision. In spite of

the fact that the described study has only a local validity, it allowed to quantify the sensitivity of certain features and, consequently, make the decision of using the corners of some colored markers located on the body of the bendable section of the surgical instrument.

7.1.2 Chapter 4

Although these features are always visible whatever is the orientation of the instruments, they resulted to be difficult to automatically segment in *in-vivo* environments. We proposed, then, a novel segmentation algorithm that is easily adaptable to any marked bendable instrument with few a-priori information on the object. It showed its robustness to strong specularities present in *in-vivo* environment and false-positive reflection of the instrument on glossy surfaces contiguous to the instrument. The core of this algorithm is composed by two stages. The first is a graph-based method which allows defining the relationship between regions of interest for selecting solely the true positive and combining them according to some visual criteria. In our case, topological and shape criteria have been used, but the same formalization can be used for other criteria. Thanks to this method, the partial markers regions can be recomposed till obtaining the entire instrument body in the image.
The second stage takes advantage of the skeleton of the segmented instrument (defined as the piecewise linear curve connecting the marker centroids) to search its upper and lower borders points in the proper directions. Our proposal of using M-estimator to fit Bézier curves to such points allowed a two-fold action: to filter border outliers out and to get a continuous representation of the border to achieve sub-pixel precision in corner extraction. Moreover, it also help to recover partially missing information. The method, though, is not robust to external or self occlusions, yet.

One model-based and one model-free approaches have been advanced in this report to tackle the 3D pose estimation starting from 2D information and some a-priori information either represented by model knowledge or training set.

7.1.3 Chapter 5 and 6

The model-based approach improves the state-of-the-art solution, by considering an extended kinematic model with additional "spurious" DOF. This factor allows to model the mechanical plays that the considered system may have and, consequently, enhance the adaptability of such method, providing good results even when the geometrical model (camera-to-instrument-channel chain) changes during the surgical operation.

It has been also shown how, in a robotic context, motor encoder information can be exploited (within a Kalman filter framework) to strengthen or complete visual

information throughout an image sequence.

On one hand this method provides the complete pose information, but, on the other hand, it showed accuracy limits in terms of RMS TCP position estimation errors, whose main cause seemed to be due to model uncertainties (for example a non constant curvature).

To overcome model uncertainties, a learning based algorithm has been developed to approximate the image-to-TCP-position relationship. Given the difficulty of obtaining an a-priori information on the nature of such relationship, local regression approaches have been preferred over regression methods based on global costs, since they can easily fit local variability without affecting the outline of the function in the whole domain. Two methods were conceived inspired by Radial Basis Function (RBF) and Locally Weighted Regression (LWR). The first one approximates the function with a linear combination of Gaussians fitted on pre-clustered data. The second one fits a plane on the training data weighted according to the distance from the query point. Differently from conventional LWR the neighborhood is composed by those samples pertaining to the two nearest clusters of the query points.

To mitigate the known issues of local approaches when dealing with high-dimensional domains, the choice of performing the clustering (needed by both methods) in the space defined by the first k principal components of the input-output joint space showed to be effective. The proposed training set enrichment showed an important effect on counteracting the noise.

Comparing the RMS error results, the learning approaches are more accurate (in average). Moreover, since the regression is based solely on input/output data, easier-to-extract (but still coherent) visual features, such as the centroids, can be chosen wrt the model-based approach. However, the difficulty here lies in creating a representative and dense training set, which can be laborious.

The choice of one method over the other, though, still depends on the meant application. In endoscopic surgery, for example, the orientation of the surgical tool (TCP) is as important as the position to ensure the right gesture and perform the operation faster. In other cases where the trajectories of the tool need to be controlled or its workspace to be limited, only the position is sufficient and the learning method proposed here can be utilized as it is.

Both model-based (thanks to the modular way in which the Jacobian is written) and learning based (for its intrinsic independence from the model) can be adapted to other camera-to-instrument configurations, e.g. conventional endoscopy where a single working channel is provided usually located underneath the camera or to flexible unarticulated instruments by fixing the parameter θ.

7.2 Future Work

A list of complementary works can be drawn up, starting from the current identified limits of the presented solution.

7.2.1 Feature Segmentation

The algorithm for extracting visual information, for example, can be made more robust to occlusion by introducing a tracking algorithm. The more direct solution would be to consider the model of the instrument itself and use a Kalman or particle filter [Isard 1998] to predict the position of the features in the image. In this case, of 3D tracking, a good 2D measurement or, alternatively, a good drift detection should be individuated. An example for such measurement can be the distance of the estimated outline to the actual apparent contour of the instrument both labeled according to which markers they pertain. This allows to have some 3D instrument related features which can be employed as harder constraints to possibly help limiting tracking drifting. Using a model, though, always presents the same limits related to the quality of the model and, consequently, can jeopardize the tracking accuracy. An alternative would be to consider the Bézier curves that can adapt at any shape even though loose the 3D meaning. Another approach could be to use 2D feature tracking (Bézier or others) for a preliminary raw estimation that can be refined with dense information (such as Mutual Information). However, in this case, a precise model of instrument appearance should be known. In any case, as a first approximation, a constant velocity model can be considered or, in alternative, take the motor encoder data increment to compute the prediction.

Regarding the segmentation, it would be worth keeping on working on the formalization of the graph-based method described in Sec. 4. Such scheme could be inscribed in a *pictorial structure* framework (or the more general *deformable part model*) which is presently used, for example, to detect person skeleton [Felzenszwalb 2005]. Firstly, the image is processed with a specific detector for each body part and, subsequently, a global cost function is used to combine the "candidate" body parts in one person skeleton. The difference with our problem is the difficulty of having characteristic features for each marker (for example, the first and last markers are not distinguishable between them, except for their relative position wrt the instrument) and the fact that the structure of the skeleton may changes (in fact, saturation does not always affect the instrument and, therefore, the instrument body can be composed by 5 or more elements). Therefore, two aspects should be enriched: a suitable characteristic to train the markers detectors (or put a particular color for each marker) and a global cost function allowing the right ordering and (re-)composition of the candidate regions.

7.2.2 Model-Based Approach

The considerable error resulting from the model-based approach can be mainly imputed to the imperfect modelling of the instrument. As future work, the possible improvement deriving from the employment of a more complete model should be verified. The weakest hypothesis seems to be the one regarding the constant curvature of the instrument in a plane. One solution would be considering all the vertebrae as independent joints, but this would dramatically increase the number of parameters over which perform the optimization. An alternative is to study, thanks to the stereoscopic system used for Ground Truth, the specific curvature of the instrument so as to describe it with a limited number of parameters.

Another improvement would be considering, in addition to corners points, the entire apparent contour of the instrument or the computed Bézier curve to reinforce the visual information and help convergence to the right solution following the works in [Colombo 2005].

Also a more clever image-motor data-fusion could be foresee. An example could be, at least, detecting by vision the histeresis characterizing the deflection: avoiding state update when no motion on the image is perceived even though a motor increment is detected.

7.2.3 Learning-Based Approach

On the other hand, the learning approach lacks of the orientation of the TCP. Given its acceptable precision on 3D position estimation, the result of the learning-based method could be used, in turn, as a new constraint for the optimization process in the model-based method. This should improve the local-minima avoidance in addition to provide an idea of the orientation of the tool.

One can also imagine to learn solely the shape of the instrument, separating it from the translational (and scale) component of the base of the bendable part. The idea would be to learn the relationship between the image features and the two parameters (ϕ and θ) responsible for modifying the aspect and orientation (in the image) of the bendable part. In this way, the shape can be learned even on synthetic data and, consequently, it could be easier to insert in the training set some non-conventional poses that can be assumed in case of interaction with the environment. Bézier curves parameters can be a good input vector for such new learning approach, since the control points inherently contain the apparent shape and orientation information.

In this learning based method there is the space of another improvement related to the way the domain dimensionality is reduced. Indeed, PCA is not the best way to choose efficient projections: components with high dispersion will be considered as principal component even though they have no effect on the output. One alternative could be PLS (Partial Least Square) which recursively performs single variable

regression along input projection and residual errors of the previous step. The key ingredient of PLS is on the choice of such projection direction, which coincides with the direction of maximal correlation between the residual errors and the input data. In the same perspective used in our work of considering output data for clustering, PLS assures that the main input projections are those that have the greatest influence on the output.

Other clustering initializations can also be studied maybe based on the output outline (or variation), for example, locating the initial centers according to the gradient of the training samples output. Alternatively, an incremental clustering can be adopted to avoid agglomerating initial guesses in one region of the space with the risk, though, to create clusters with almost no data.

7.2.4 Simulation Experiments

To strengthen the conclusions derived from results and simulations, further tests and theoretical studies can be envisaged. For example, other features or descriptors can be considered in the sensitivity preliminary study or additional tests with other state-of-the-art learning algorithm (e.g. Random Forest) could be performed to see the actual potentiality of the proposed solution. From the segmentation point of view, the strengths and weaknesses of graph-based segmentation and Bézier fitting of the borders could be tested on additional video sequences to verify its robustness in cluttered scenes or in different *in-vivo* background. It will be interesting, as well, applying the framework developed here to other robotic or manual systems used in general endoscopy or, even, concentric tubes (single bending section concentric tube at a first instance) to test its inter-class applicability.

7.2.5 Instrument Control

Although, the real flowering of this work (already evoked in the title) would be the exploitation of pose estimation for controlling the flexible instrument with visual servoing. The most direct implementation would be an Image Based Visual Servoing (IBVS), where the reference can be a picture of the instrument in a given configuration. At each time step a control action can be computed in a similar manner to the optimization process used in pose estimation and the resulting increment for the DOF values can be used as velocity reference for those motors actuating the 3 DOF of the flexible instrument.

7.2.6 Possible Further Application

Another really interesting application of such work would be contact detection with the environment or force sensing. If the force applied on the tip causes movements of the bendable section which can be attributed to the conventional movements (i.e. that can be explained with the kinematic model), the contact can be detected looking

at the difference between the encoder incremental values and the increment of the DOF revealed from the image. More precisely, if the estimated DOFs experience a variation which is not reflected by encoders data, one can attribute such movements to external forces acting on the instrument.

As the reliability of deflection encoders is poor for retrieving the real instrument deflection, the data referring to this DOF would not be trustful enough to associate a force estimation.
Nevertheless, in all the cases, the resolution and accuracy reachable with such technique is not clear *a priori* and probably would not be sufficient for high-precision applications, but would only give an idea on the force magnitude or increment.

Appendix A

Apparent Border Points

A.1 Computation of the 3D Corresponding Points

In chapter 3 we pointed out the dependency of the extended image Jacobian from the depth of the point for which it is computed. Therefore, we have to be capable of computing such point for different configurations of the instrument.

The instrument can be considered as a torus of internal radius r which changes its 3D shape according to the variations of the 3 natura DOFs (λ, ϕ, θ). Knowing the instrument kinematic model and r these points can be computed according to perspective projective relationship for the torus. Fig. A.1 shows the profile of a traversal section of the instruments and the projection rays of its boarders towards the camera. Looking at the scheme of this figure, we can define some properties for those point whose projection will form the apparent border of the instrument:

- The projection rays are tangent to the circumference in the points $P_{i,u}$ and $P_{i,l}$ and pass through the principal point of the camera C.
- Unlike C, $P_{i,u}$ and $P_{i,l}$ will lay on the same plane defined by the circumference (which is the same defined by the x and y axis of $\mathcal{F}_{sec}$).
- The distance of $P_{i,u}$ and $P_{i,l}$ from S will be equal to r.

To make all the computation easier, these geometrical relationships can be expressed taking as reference frame the one associated to the center of the circumference (i.e. $\mathcal{F}_{sec}$). Therefore, defining

$$P_i^{sec} = \begin{bmatrix} X_i \\ Y_i \\ 0 \end{bmatrix}, C^{sec} = \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix}$$

the following equations can be written:

$$X_i^2 + Y_i^2 = r^2 \tag{A.1}$$

since P_1 and P_2 pertain to the circumference and

$$[X_i\, Y_i\, 0] \begin{bmatrix} X_i - X_c \\ Y_i - Y_c \\ -Z_c \end{bmatrix} = 0 \tag{A.2}$$

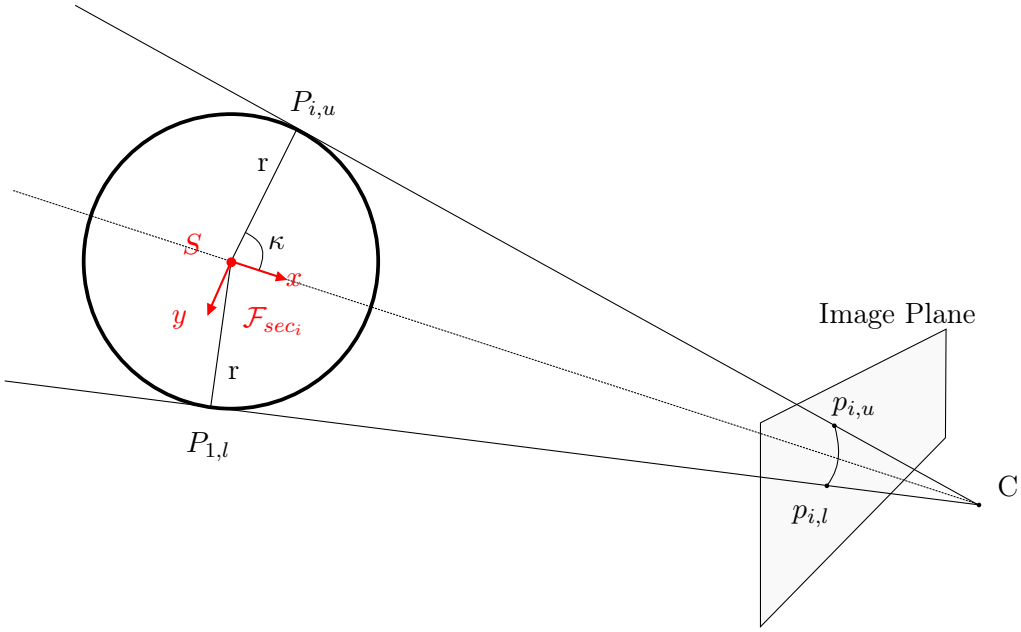


Figure A.1: Perspective projection of a torus shaped instrument.

for the perpendicularity condition.

Executing the multiplication in equation (A.2) and considering equation (A.1), X_i can be expressed as

$$X_i = \frac{r^2 - Y_i Y_c}{X_c}.$$

Then, equation (A.1) can be written as follows:

$$Y_c^2\, Y_i^2 + (X_c^2 - 2R^2 Y_c)\, Y_i + r^4 - r^2 - r^2 X_c^2 = 0$$

which is a second order equation with two solutions: one is the upper boarder point $\mathbf{P}_u^{sec} = (X_u, Y_u, 0)$ and the other is the lower $\mathbf{P}_l^{sec} = (X_l, Y_l, 0)$.

The position with respect to the camera frame can be easily found from the solutions just obtained as:

$$\mathbf{P}_u^c = t_{c,sec}^c + R_{sec}^c \begin{bmatrix} X_u \\ Y_u \\ 0 \end{bmatrix}$$

and the same thing for P_l.

A.2 Computation of the Jacobian for Apparent Points

Corners of the markers are not completely defined physically. Their actual position on a section varies depending on the pose of the instrument with respect to the camera, and as a consequence the velocity of the apparent point on the section

In Appendix A	
${}^{d}V^{c}_{a/b}$	velocity between $\mathcal{F}_a$ and $\mathcal{F}_b$ computed in c and expressed wrt $\mathcal{F}_d$
${}^{a}C$	3D position of C wrt $\mathcal{F}_a$ reference frame
${}^{b}R_a$	Matrix expressing the orientation of $\mathcal{F}_a$ wrt $\mathcal{F}_b$

Table A.1: New notation used in section A.2

is also affected. This effect has to be taken into account in the geometrical 3D Jacobian.

In this section, a change in the notation is performed to precisely express all the relationships needed for the computation of the Jacobian for apparent points (cf. Tab. A.1).

A.2.1 Velocity of an apparent point onto its section

Let consider an apparent point A (either $P_{i,u}$ or $P_{i,l}$), which can slide onto a given section of the instrument of center S. Then

$$ {}^{cam}V^{A}_{A/cam} = {}^{cam}V^{A}_{A/sec} + {}^{cam}V^{A}_{sec/cam}, $$

where the first term of the right side of the equation is the velocity of the apparent point with respect to physical points of the considered section, and the second term is the velocity of the physical point with respect to the camera.

The second term can be expressed conventionally in function of the joint velocities as : ${}^{cam}V^{A}_{sec/cam} = J^{A}_{3D}\dot{r}$.

To express the velocity of the apparent point A onto the section it pertains (the first term), one first expresses the constraint of a point on the section to be an apparent point: orthogonality between SA and CA.

In the frame of the section, one has :

$$ r(cos(\kappa), \pm sin(\kappa), 0)({}^{sec}C - r \begin{pmatrix} cos(\kappa) \\ \pm sin(\kappa) \\ 0 \end{pmatrix}) = 0 $$

which can be rewritten as:

$$ Q = r(cos(\kappa), \pm sin(\kappa), 0)\ {}^{sec}C - r^2 = 0 $$

By noting ${}^{sec}C = (x_c, y_c, z_c)^T$ and stating that the previous relation should be constant with respect to time, one has $\frac{dQ}{dt} = 0$ and one obtains:

$$ \frac{dQ}{dt} = \frac{\partial Q}{\partial\ {}^{sec}C} \begin{pmatrix} \dot{x_c} \\ \dot{y_c} \\ \dot{z_c} \end{pmatrix} + \frac{\partial Q}{\partial \kappa}\dot{\kappa} = 0 $$

which gives (considering, without loss of generality, only the lower apparent point):

$$\dot{\kappa} = \frac{-cos(\kappa)\dot{x_c} - sin(\kappa)\dot{y_c}}{y_c cos(\kappa) - x_c sin(\kappa)}.$$

Noting that ${}^{sec}\omega_{A/sec} = (0, 0, \dot{\kappa})^T$, and that

$$\begin{aligned} {}^{cam}V^A_{A/sec} &= {}^{cam}R_{sec}\ {}^{sec}V^A_{A/sec} = {}^{cam}R_{sec}(\ {}^{sec}V^S_{A/sec} + \ {}^{sec}\omega_{A/sec} \wedge \ {}^{sec}SA) \\ &= {}^{cam}R_{sec} \begin{pmatrix} -\dot{\kappa} r sin(\kappa) \\ \dot{\kappa} r cos(\kappa) \\ 0 \end{pmatrix} \end{aligned}$$

one gets

$${}^{cam}V^A_{A/sec} = \ {}^{cam}R_{sec} M \begin{pmatrix} \dot{x_c} \\ \dot{y_c} \\ \dot{z_c} \end{pmatrix}$$

with

$$M = \frac{1}{y_c cos(\kappa) - x_c sin(\kappa)} \begin{pmatrix} r sin(\kappa) cos(\kappa) & r sin(\kappa)^2 & 0 \\ -r cos(\kappa)^2 & -r sin(\kappa) cos(\kappa) & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

One also has $(\dot{x_c}, \dot{y_c}, \dot{z_c})^T = \ {}^{sec}V^c_{cam/sec} = -{}^{sec}R_{cam}\ {}^{cam}V^C_{sec/cam}$.
We note ${}^{cam}V^C_{sec/cam} = J_{g\_gen}\dot{r}$ where J_{g_gen} is the Jacobian relating the joint velocities to the velocity of point C, considered as a point attached to $\mathcal{F}_{sec}$. This Jacobian is close to the one of eq. 3.24, although it is different because C does not generally belong to the plane of the considered section.

Actually, the effect of $\dot{\lambda}$, $\dot{\phi}$, $\dot{\psi}$, $\dot{x_{ch}}$, $\dot{y_{ch}}$ and $\dot{\mu}$ are not modified by this fact. Only the effect of $\dot{\theta}$ (A_θ) is modified, because $\dot{\theta}$ produces different effects on different sections of the instrument.

For deriving the component A_θ for J_{g_gen}, one can follow the same process as for J_g, but taking into account a non null third component for $t^b_{S,\mathbf{C}}$:

$$t^b_{S,\mathbf{C}} = \begin{bmatrix} x_c cos\theta_i + sin(\theta_i) z_c \\ y_c \\ -x_c sin\theta_i + cos(\theta_i) z_c \end{bmatrix} \tag{A.3}$$

$$t^b_{b,\mathbf{C}} = \begin{bmatrix} R_{curv}(1 - cos\theta_i) \\ 0 \\ R_{curv} sin\theta_i \end{bmatrix} + \begin{bmatrix} x_c cos\theta_i + sin(\theta_i) z_c \\ y_c \\ -x_c sin\theta_i + cos(\theta_i) z_c \end{bmatrix}. \tag{A.4}$$

$$A_\theta = R^c_b \frac{\partial \mathbf{C}^b}{\partial \theta}$$

with

$$\frac{\partial \mathbf{C}^b}{\partial \theta} = \begin{pmatrix} k_\theta L/\theta \sin\theta_i - C_x^{sec} k_\theta \sin\theta_i + C_z^{sec} k_\theta \cos\theta_i - L/\theta^2(1-\cos\theta_i) \\ 0 \\ k_\theta L/\theta \cos\theta_i - C_x^{sec} k_\theta \cos\theta_i - C_z^{sec} k_\theta \sin\theta_i - L/\theta^2 \sin\theta_i \end{pmatrix}$$

Finally

$$^{cam}V^A_{A/cam} = J^A_{g_{app}} \dot{r}$$

with

$$J^A_{g_{app}} = -\,^{cam}R_{sec} M J^C_{g\_gen} + J^A_g.$$

The 3D geometric Jacobian of the apparent point A can hence be obtained from the usual 3D Jacobian of the same point but considered as a physical point attached to the section, the generalized Jacobian expressed at the origin of the camera C (considering C as part of the instrument), from the rotation between the camera and the section the point A belongs to, and matrix M. M depends on the position of the apparent point on the section and the position of the camera with respect to the section.

Appendix B

Validation Process

For the validation process to be meaningful, the estimated quantity (3D TCP pose) needs to be compare with the value obtained with an alternative measurement method. A possibility for retrieving such measurement is attaching an external sensor, such as electromagnetic sensors, to the tip of the instrument. However, in previous experiments, poor quality results have been provided probably due to the effect of metallic parts of the Anubis and its motorization on the electromagnetic field.

To avoid these problems, an alternative method is used here based on stereo vision. A system composed by two cameras has been placed looking towards the left instrument workspace so that the white ball was visible in both images.

To build the ground truth (GT), a sequence of several instrument configurations is created. For each of them the endoscopic image is stored together with 3D position of the white ball of the electric tool (as shown in Fig. 3.8) estimated using stereo vision.

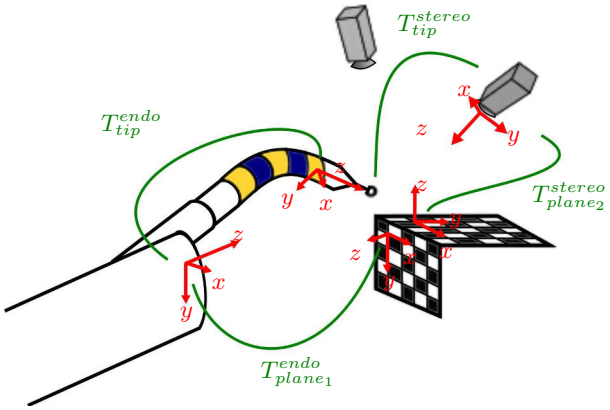
The pose estimation based on endoscopic images provides the 3D pose wrt to the endoscopic camera whereas the ground truth as defined so far is expressed wrt the stereo-vision system. A registration step between the stereo system and the endoscopic camera is needed to be able to compute the root mean square (RMS) estimation errors over the 3D position of the electric tool tip (white ball).

To carry out this registration, the particular object shown in Fig. B.1(b) is placed so that one grid is captured by the endoscopic camera and the other is captured by the stereo system as shown in (Fig. B.1. Knowing the intrinsic parameters of the cameras (thanks to a previous calibration process for both stereoscopic system and endoscopic cameras) and detecting the corners of the corresponding grids, the extrinsic parameters can be computed for both endoscopic and stereoscopic system defining, respectively, $T^{left}_{grid_2}$ and $T^{endo}_{grid_1}$.

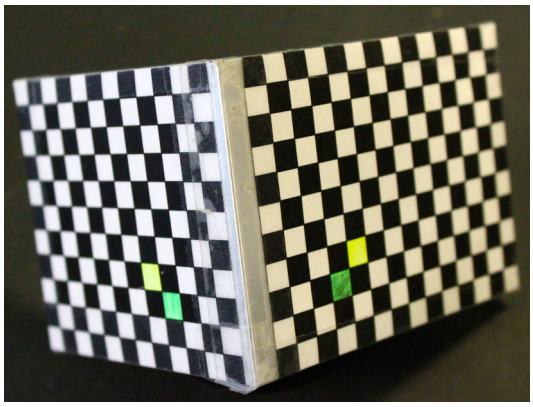
To finalize the registration, the transformation ($T^{grid_1}_{grid_2}$) between the reference frame of the two grids (whose origins is the common corner of the green yellow colored squares) must be known. This depends on the geometry of the chosen object which has been retrieved using the stereo-vision system.

In this manner, the global transformation between one camera of the stereo system and the endoscopic camera can be computed as follows:

$$T^{endo}_{left} = T^{endo}_{grid_1} T^{grid_1}_{grid_2} [T^{left}_{grid_2}]^{-1}.$$



(a)



(b)

Figure B.1: (a) Scheme of the validation test-bed. The stereo-vision system returns the 3D position of the white ball. The calibration grid object in (b) allows to know the transformation from the endoscopic camera and the stereo system.

Appendix C

Computation of Jacobian Block for the Mechanical Parameters

When assuming tolerances on the mechanical parameters $(x_{ch}, y_{ch}, \psi, \mu)$, their variability effect on the image has to be taken into account in the optimization process of the model-based method. For that reason, the geometrical Jacobian (that is used to compute the extended image Jacobian) must include the relationship describing the 3D point velocity due to the mechanical parameters variations.

We just list here the blocks A_i related to the mechanical parameters which are employed in equation (5.8).

The variation of the channel position directly modify the position of any 3D point pertaining to the instrument obtaining:

$$A_{x_{ch}} = \frac{\partial t^c_{c,\mathbf{P_i}}}{\partial x_{ch}} = \frac{\partial t^c_{c,b}}{\partial x_{ch}} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \tag{C.1}$$

and

$$A_{y_{ch}} = \frac{\partial t^c_{c,\mathbf{P_i}}}{\partial y_{ch}} = \frac{\partial t^c_{c,b}}{\partial y_{ch}} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}. \tag{C.2}$$

For ψ and μ the same principle used for ϕ is employed here, obtaining:

$$A_{\psi} = \frac{\partial t^c_{ch,\mathbf{P_i}}}{\partial \phi} = -[(P_i - t^c_{c,ch}) \times y^c_{ch})]. \tag{C.3}$$

and

$$A_{\mu} = \frac{\partial t^c_{ch,\mathbf{P_i}}}{\partial \mu} = -[(P_i - t^c_{c,ch}) \times [\cos(\psi)\, 0 \; -\sin(\psi)]^T)]. \tag{C.4}$$

Bibliography

[Agrawal 2010a] V. Agrawal, W.J. Peine and Bin Yao. *Modeling of Transmission Characteristics Across a Cable-Conduit System.* IEEE Transactions on Robotics, vol. 26, no. 5, pages 914–924, October 2010. (Cited on page 29.)

[Agrawal 2010b] Varun Agrawal, William J. Peine, Bin Yao and SeungWook Choi. *Control of cable actuated devices using smooth backlash inverse.* In Robotics and Automation (ICRA), 2010 IEEE International Conference on, pages 1074–1079, 2010. (Cited on page 30.)

[Agrawal 2012] Vishal Agrawal, William J. Peine and Bin Yao. *Dual loop control of cable-conduit actuated devices.* In American Control Conference (ACC), 2012, pages 2621–2626. IEEE, 2012. (Cited on page 31.)

[Atkeson 1997] Christopher G. Atkeson, Andrew W. Moore and Stefan Schaal. *Locally Weighted Learning.* In David W. Aha, editeur, Lazy Learning, pages 11–73. Springer Netherlands, 1997. DOI: 10.1007/978-94-017-2053-3_2. (Cited on page 138.)

[Azizian 2014] Mahdi Azizian, Mahta Khoshnam, Nima Najmaei and Rajni V. Patel. *Visual servoing in medical robotics: a survey. Part I: endoscopic and direct vision imaging – techniques and applications.* The International Journal of Medical Robotics and Computer Assisted Surgery, vol. 10, 2014. (Cited on page 37.)

[Bardou 2010] Berengere Bardou, Philippe Zanne, Florent Nageotte and Michel de Mathelin. *Control of a multiple sections flexible endoscopic system.* In Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on, pages 2345–2350, 2010. (Cited on pages 28 and 31.)

[Bardou 2012] Berengere Bardou, Florent Nageotte, Philippe Zanne and Michel de Mathelin. *Improvements in the control of a flexible endoscopic system.* In Robotics and Automation (ICRA), 2012 IEEE International Conference on, pages 3725–3732, 2012. (Cited on page 31.)

[Barron 2015] Jonathan Barron and Jitendra Malik. *Shape, Illumination, and Reflectance from Shading.* IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–1, 2015. (Cited on page 43.)

[Bartoli 2015] Adrien Bartoli, Yan Gerard, Francois Chadebecq, Toby Collins and Daniel Pizarro. *Shape-from-Template.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 10, pages 2099–2118, October 2015. (Cited on page 43.)

[Beaton 1974] Albert E. Beaton and John W. Tukey. *The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data*. Technometrics, vol. 16, no. 2, pages 147–185, May 1974. (Cited on page 82.)

[Benoudjit 2003] Nabil Benoudjit and Michel Verleysen. *On the Kernel Widths in Radial-Basis Function Networks*. Neural Process. Lett., vol. 18, no. 2, pages 139–154, October 2003. (Cited on page 126.)

[Beyer 1999] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan and Uri Shaft. *When is "nearest neighbor" meaningful?* In Database Theory—ICDT'99, pages 217–235. Springer, 1999. (Cited on page 133.)

[Blanz 1999] Volker Blanz and Thomas Vetter. *A morphable model for the synthesis of 3D faces*. In Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. (Cited on page 43.)

[Bors 2001] Adrian Gheorghe Bors. *Introduction of the radial basis function (rbf) networks*. February 2001. (Cited on page 121.)

[Bouguet 2013] Jean-Yves Bouguet. *Camera Calibration Toolbox for Matlab*, 2013. (Cited on page 56.)

[Broomhead 1988] D.S. Broomhead and D. Lowe. *Multivariable Functional Interpolation and Adaptive Networks*. Complex Systems, vol. 2, pages 321–355, 1988. (Cited on page 120.)

[Caglioti 2006] Vincenzo Caglioti and Alessandro Giusti. *Reconstruction of canal surfaces from single images under exact perspective*. In Computer Vision–ECCV 2006, pages 289–300. Springer, 2006. (Cited on pages 33 and 44.)

[Camarillo 2008a] David B. Camarillo, Kevin E. Loewke, Christopher R. Carlson and J. Kenneth Salisbury. *Vision based 3-D shape sensing of flexible manipulators*. In Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on, pages 2940–2947, 2008. (Cited on page 32.)

[Camarillo 2008b] D.B. Camarillo, C.F. Milne, C.R. Carlson, M.R. Zinn and J.K. Salisbury. *Mechanics Modeling of Tendon-Driven Continuum Manipulators*. IEEE Transactions on Robotics, vol. 24, no. 6, pages 1262–1273, December 2008. (Cited on page 30.)

[Camarillo 2009] D.B. Camarillo, C.R. Carlson and J.K. Salisbury. *Configuration Tracking for Continuum Manipulators With Coupled Tendon Drive*. IEEE Transactions on Robotics, vol. 25, no. 4, pages 798–808, August 2009. (Cited on page 30.)

[Chen 2007] Datong Chen. *Modeling vs. Segmenting Images Using A Probabilistic Approach.* In Image Processing, 2007. ICIP 2007. IEEE International Conference on, volume 2, pages II–277. IEEE, 2007. (Cited on page 70.)

[Chhatkuli 2014] Ajad Chhatkuli, Daniel Pizarro and Adrien Bartoli. *Non-Rigid Shape-from-Motion for Isometric Surfaces using Infinitesimal Planarity.* In Proceedings of the British Machine Vision Conference. BMVA Press, September 2014. (Cited on page 43.)

[Chirikjian 1993] G.S. Chirikjian. *A continuum approach to hyper-redundant manipulator dynamics.* In Proceedings of the 1993 IEEE/RSJ International Conference on Intelligent Robots and Systems '93, IROS '93, volume 2, pages 1059–1066 vol.2, July 1993. (Cited on page 30.)

[Chirikjian 1994] G.S. Chirikjian and J.W. Burdick. *A modal approach to hyper-redundant manipulator kinematics.* IEEE Transactions on Robotics and Automation, vol. 10, no. 3, pages 343–354, June 1994. (Cited on page 30.)

[Cipolla 1992] Roberto Cipolla and Andrew Blake. *Surface shape from the deformation of apparent contours.* Int J Comput Vision, vol. 9, no. 2, pages 83–112, November 1992. (Cited on page 42.)

[Colombo 2005] Carlo Colombo, Alberto Del Bimbo and Federico Pernici. *Metric 3D reconstruction and texture acquisition of surfaces of revolution from a single uncalibrated view.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 27, no. 1, pages 99–114, 2005. (Cited on page 155.)

[Criminisi 1999] Antonio Criminisi, Ian Reid and Andrew Zisserman. *Single view metrology.* In Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, volume 1, pages 434–441. IEEE, 1999. (Cited on page 42.)

[Croom 2010] Jordan M. Croom, D. Caleb Rucker, Joseph M. Romano and Robert J. Webster III. *Visual sensing of continuum robot shape using self-organizing maps.* In Robotics and Automation (ICRA), 2010 IEEE International Conference on, pages 4591–4596. IEEE, 2010. (Cited on pages 32 and 93.)

[Curcillo 2010] Paul G. Curcillo, Andrew S. Wu, Erica R. Podolsky, Casey Graybeal, Namir Katkhouda, Alex Saenz, Robert Dunham, Steven Fendley, Marc Neff, Chad Copper, Marc Bessler, Andrew A. Gumbs, Michael Norton, Antonio Iannelli, Rodney Mason, Ashkan Moazzez, Larry Cohen, Angela Mouhlas and Alex Poor. *Single-port-access (SPATM) cholecystectomy: a multi-institutional report of the first 297 cases.* Surgical Endoscopy, vol. 24, no. 8, pages 1854–1860, February 2010. (Cited on page 1.)

[Cuschieri 1991] A. Cuschieri, F. Dubois, J. Mouiel, P. Mouret, H. Becker, G. Buess, M. Trede and H. Troidl. *The European experience with laparoscopic cholecystectomy.* Am. J. Surg., vol. 161, no. 3, pages 385–387, March 1991. (Cited on page 1.)

[De Donno 2013] Antonio De Donno, Lucile Zorn, Philippe Zanne, Florent Nageotte and Michel de Mathelin. *Introducing STRAS: A new flexible robotic system for minimally invasive surgery.* In 2013 IEEE International Conference on Robotics and Automation (ICRA), pages 1213–1220, 2013. (Cited on page 21.)

[Degani 2006] Amir Degani, Howie Choset, Alon Wolf, Takeyoshi Ota, Marco Zenati and others. *Percutaneous intrapericardial interventions using a highly articulated robotic probe.* In Biomedical Robotics and Biomechatronics, 2006. BioRob 2006. The First IEEE/RAS-EMBS International Conference on, pages 7–12. IEEE, 2006. (Cited on page 22.)

[Dhumane 2011] Parag W Dhumane, Michele Diana, Joel Leroy and Jacques Marescaux. *Minimally invasive single-site surgery for the digestive system: A technological review.* J Minim Access Surg, vol. 7, no. 1, pages 40–51, 2011. (Cited on page 8.)

[Di Biase 2009] Luigi Di Biase, Andrea Natale, Conor Barrett, Carmela Tan, Claude S. Elayi, Chi Keong Ching, Paul Wang, Amin Al-Ahmad, Mauricio Arruda, J. David Burkhardt, Brian J. Wisnoskey, Punam Chowdhury, Shari De Marco, Luciana Armaganijan, Kenneth N. Litwak, Robert A. Schweikert and Jennifer E. Cummings. *Relationship between catheter forces, lesion characteristics, "popping," and char formation: experience with robotic navigation system.* J. Cardiovasc. Electrophysiol., vol. 20, no. 4, pages 436–440, April 2009. (Cited on page 24.)

[Dogangil 2010] G. Dogangil, B. L. Davies and F. Rodriguez y Baena. *A review of medical robotics for minimally invasive soft tissue surgery.* Proc Inst Mech Eng H, vol. 224, no. 5, pages 653–679, 2010. (Cited on page 2.)

[Doignon 1999] Christophe Doignon and Gabriel Abba. *A practical multi-plane method for a low-cost calibration technique.* In Proceedings of the European Control Conference, Karlsruhe, September 1999. (Cited on page 56.)

[Doignon 2004] C. Doignon, F. Nageotte and M. de Mathelin. *Detection of grey regions in color images : application to the segmentation of a surgical instrument in robotized laparoscopy.* In 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings, volume 4, pages 3394–3399 vol.4, September 2004. (Cited on page 68.)

[Doignon 2007] Christophe Doignon. *An Introduction to Model-Based Pose Estimation and 3-D Tracking Techniques.* In Pseudo Stereovision System (PSVS): A Monocular Mirror-based Stereovision System. INTECH Open Access Publisher, 2007. (Cited on pages 44 and 93.)

[Dupont 2010] P.E. Dupont, J. Lock, B. Itkowitz and E. Butler. *Design and Control of Concentric-Tube Robots.* IEEE Transactions on Robotics, vol. 26, no. 2, pages 209–225, April 2010. (Cited on page 30.)

[Fanfani 2015] Francesco Fanfani, Giorgia Monterossi, Anna Fagotti, Cristiano Rossitto, Salvatore Gueli Alletti, Barbara Costantini, Valerio Gallotta, Luigi Selvaggi, Stefano Restaino and Giovanni Scambia. *The new robotic TELELAP ALF-X in gynecological surgery: single-center experience.* Surg Endosc, April 2015. (Cited on page 17.)

[Felzenszwalb 2005] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. *Pictorial structures for object recognition.* International Journal of Computer Vision, vol. 61, no. 1, pages 55–79, 2005. (Cited on pages 72 and 154.)

[festo 2015] festo festo. *Bionic Handling Assistant | Festo Corporate*, 2015. (Cited on page 28.)

[Gidaro 2012] Stefano Gidaro, Maurizio Buscarini, Emilio Ruiz, Michael Stark and Anna Labruzzo. *Telelap Alf-X: a novel telesurgical system for the 21st century.* Surg Technol Int, vol. 22, pages 20–25, December 2012. (Cited on page 16.)

[Gidaro 2014] Stefano Gidaro, Emanuela Altobelli, Cristina Falavolti, Alfredo Maria Bove, Emilio Morales Ruiz, Michael Stark, Giuliano Ravasio, Sara Simona Lazzaretti and Buscarini Maurizio. *Vesicourethral anastomosis using a novel telesurgical system with haptic sensation, the Telelap Alf-X: a pilot study.* Surg Technol Int, vol. 24, pages 35–40, March 2014. (Cited on page 17.)

[Gravagne 2000] I.A. Gravagne and I.D. Walker. *Kinematic transformations for remotely-actuated planar continuum robots.* In IEEE International Conference on Robotics and Automation, 2000. Proceedings. ICRA '00, volume 1, pages 19–26 vol.1, 2000. (Cited on page 30.)

[Hannan 2003] M. Hannan and I. Walker. *Vision based shape estimation for continuum robots.* In IEEE International Conference on Robotics and Automation, 2003. Proceedings. ICRA '03, volume 3, pages 3449–3454 vol.3, 2003. (Cited on page 32.)

[Hartley 2004] R.I. Hartley and Andrew Zisserman. Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521540518, second édition, 2004. (Cited on page 42.)

[Heikkila 1997] Janne Heikkila and Olli Silvén. *A four-step camera calibration procedure with implicit image correction.* In Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, pages 1106–1112. IEEE, 1997. (Cited on page 56.)

[Hel-Or 1994] Yaacov Hel-Or and Michael Werman. *Model based pose estimation of articulated and constrained objects.* In Jan-Olof Eklundh, editeur, Computer Vision — ECCV '94, numéro 800 de Lecture Notes in Computer Science, pages 262–273. Springer Berlin Heidelberg, May 1994. DOI: 10.1007/3-540-57956-7_31. (Cited on page 93.)

[Horn 1970] Berthold K.P. Horn. *Shape From Shading: A method for Obtaining the Shape of a Smooth Opaque Object From One View.* Rapport technique 232, Massachusetts Institute of Technology, Cambridge, Mass., 1970. (Cited on page 42.)

[Isard 1998] Michael Isard and Andrew Blake. *Conditional Density Propagation for Visual Tracking.* International Journal of Computer Vision, vol. 29, no. 1, pages 5–28, 1998. (Cited on page 154.)

[Ivanescu 1995] Mircea Ivanescu and Viorel Stoian. *A variable structure controller for a tentacle manipulator.* In Robotics and Automation, 1995. Proceedings., 1995 IEEE International Conference on, volume 3, pages 3155–3160. IEEE, 1995. (Cited on page 30.)

[Ivanescu 2004] Mircea Ivanescu. *On the Dynamical Control of Hyper Redundant Manipulators.* In Advances in Automatic Control, pages 141–158. Springer, 2004. (Cited on page 30.)

[Johnson 2013] Paul J. Johnson, Carlos M. Rivera Serrano, Michael Castro, Richard Kuenzler, Howie Choset, Stephen Tully and Umamaheswar Duvvuri. *Demonstration of transoral surgery in cadaveric specimens with the medrobotics flex system.* Laryngoscope, vol. 123, no. 5, pages 1168–1172, May 2013. (Cited on page 22.)

[Jones 2006] B.A. Jones and I.D. Walker. *Kinematics for multisection continuum robots.* IEEE Transactions on Robotics, vol. 22, no. 1, pages 43–55, February 2006. (Cited on page 30.)

[Kalloo 2004] Anthony N. Kalloo, Vikesh K. Singh, Sanjay B. Jagannath, Hideaki Niiyama, Susan L. Hill, Cheryl A. Vaughn, Carolyn A. Magee and Sergey V. Kantsevoy. *Flexible transgastric peritoneoscopy: a novel approach to diagnostic and therapeutic interventions in the peritoneal cavity.* Gastrointest. Endosc., vol. 60, no. 1, pages 114–117, July 2004. (Cited on page 8.)

[Kapadia 2011] Apoorva Kapadia and Ian D. Walker. *Task-space control of extensible continuum manipulators.* In Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on, pages 1087–1092. IEEE, 2011. (Cited on page 30.)

[Kehoe 2014] B. Kehoe, G. Kahn, J. Mahler, J. Kim, A. Lee, A. Lee, K. Nakagawa, S. Patil, W.D. Boyd, P. Abbeel and K. Goldberg. *Autonomous multilateral debridement with the Raven surgical robot.* In 2014 IEEE International Conference on Robotics and Automation (ICRA), pages 1432–1439, May 2014. (Cited on page 16.)

[Kesner 2010] Samuel B. Kesner and Robert D. Howe. *Design and control of motion compensation cardiac catheters.* In Robotics and Automation (ICRA), 2010 IEEE International Conference on, pages 1059–1065. IEEE, 2010. (Cited on page 30.)

[Kwoh 1988] Y. S. Kwoh, J. Hou, E. A. Jonckheere and S. Hayati. *A robot with improved absolute positioning accuracy for CT guided stereotactic brain surgery.* IEEE Trans Biomed Eng, vol. 35, no. 2, pages 153–160, February 1988. (Cited on page 15.)

[Lendasse 2003] A. Lendasse, J. Lee, E. de Bodt, V. Wertz and M. Verleysen. *Approximation by Radial Basis Function Networks - Application to Option Pricing.* In in Connectionist Approaches in Economics and Management Sciences, pages 201–212. Kluwer academic publishers, 2003. (Cited on page 121.)

[Lepetit 2005] Vincent Lepetit and Pascal Fua. Monocular model-based 3d tracking of rigid objects. Now Publishers Inc, 2005. (Cited on page 44.)

[Litynski 1998] Grzegorz S. Litynski. *Erich Mühe and the Rejection of Laparoscopic Cholecystectomy (1985): A Surgeon Ahead of His Time.* JSLS, vol. 2, no. 4, pages 341–346, 1998. (Cited on pages 1 and 7.)

[Loeve 2010] A. Loeve, P. Breedveld and J. Dankelman. *Scopes Too Flexible...and Too Stiff.* IEEE Pulse, vol. 1, no. 3, pages 26–41, November 2010. (Cited on page 25.)

[Longuet-Higgins 1986] H. C. Longuet-Higgins. *Visual motion ambiguity.* Vision Research, vol. 26, no. 1, pages 181–183, 1986. (Cited on page 42.)

[Lowe 1991] David G. Lowe. *Fitting parameterized three-dimensional models to images.* IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 5, pages 441–450, 1991. (Cited on pages 98 and 99.)

[Maghooa 2015] Farahnaz Maghooa, Agostino Stilli, Yohan Noh, Kaspar Althoefer and Helge A Wurdemann. *Tendon and pressure actuation for a bio-inspired*

manipulator based on an antagonistic principle. In 2015 IEEE International Conference on Robotics and Automation (ICRA), pages 2556–2561, May 2015. (Cited on page 25.)

[Marchand 2002] E. Marchand and F. Chaumette. *Virtual Visual Servoing: a framework for real-time augmented reality*. Eurographics, vol. 21, no. 3, 2002. (Cited on page 93.)

[Marescaux 2007] Jacques Marescaux, Bernard Dallemagne, Silvana Perretta, Arnaud Wattiez, Didier Mutter and Dimitri Coumaros. *Surgery without scars: report of transluminal cholecystectomy in a human being*. Arch Surg, vol. 142, no. 9, pages 823–826; discussion 826–827, September 2007. (Cited on page 8.)

[McMahan 2006] William McMahan, V. Chitrakaran, M. Csencsits, D. Dawson, Ian D. Walker, Bryan A. Jones, M. Pritts, D. Dienno, M. Grissom and Christopher D. Rahn. *Field trials and testing of the OctArm continuum manipulator*. In Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on, pages 2336–2341. IEEE, 2006. (Cited on page 28.)

[Mehling 2006] Joshua S. Mehling, Myron Diftler, Mars Chu, Michael Valvo and others. *A minimally invasive tendril robot for in-space inspection*. In Biomedical Robotics and Biomechatronics, 2006. BioRob 2006. The First IEEE/RAS-EMBS International Conference on, pages 690–695. IEEE, 2006. (Cited on page 28.)

[Mochiyama 2003] Hiromi Mochiyama and Takahiro Suzuki. *Kinematics and dynamics of a cable-like hyper-flexible manipulator*. In Robotics and Automation, 2003. Proceedings. ICRA'03. IEEE International Conference on, volume 3, pages 3672–3677. IEEE, 2003. (Cited on page 30.)

[Moody 1989] John Moody and Christian J. Darken. *Fast Learning in Networks of Locally-tuned Processing Units*. Neural Comput., vol. 1, no. 2, pages 281–294, 1989. (Cited on page 120.)

[Moreno-noguer 2003] Francesc Moreno-noguer, Juan Andrade-cetto and Alberto Sanfeliu. *Fusion of Color and Shape for Object Tracking Under Varying Illumination*. In Proc.IBPRIA, LNCS 2652, pages 580–588. Springer, 2003. (Cited on page 71.)

[oc 2015] oc oc. *OC Robotics*, 2015. (Cited on page 27.)

[Padoy 2012] Nicolas Padoy and Gregory D. Hager. *Deformable Tracking of Textured Curvilinear Objects*. In BMVC, pages 1–11, 2012. (Cited on pages 33 and 93.)

[Penning 2011] Ryan S. Penning, Jinwoo Jung, Justin Borgstadt, Nicola J. Ferrier, Michael R. Zinn and others. *Towards closed loop control of a continuum robotic manipulator for medical applications*. In Robotics and Automation (ICRA), 2011 IEEE International Conference on, pages 4822–4827. IEEE, 2011. (Cited on page 31.)

[Phee 2009] S. J. Phee, S. C. Low, V. A. Huynh, A. P. Kencana, Z. L. Sun and K. Yang. *Master and slave transluminal endoscopic robot (MASTER) for natural Orifice Transluminal Endoscopic Surgery (NOTES)*. In 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pages 1192–1195, Sept 2009. (Cited on page 21.)

[Piccigallo 2010] M. Piccigallo, U. Scarfogliero, C. Quaglia, G. Petroni, P. Valdastri, A. Menciassi and P. Dario. *Design of a Novel Bimanual Robotic System for Single-Port Laparoscopy*. IEEE/ASME Transactions on Mechatronics, vol. 15, no. 6, pages 871–878, December 2010. (Cited on page 17.)

[Podolsky 2010] Erica R. Podolsky and Paul G. Curcillo. *Single port access (SPA) surgery–a 24-month experience*. J. Gastrointest. Surg., vol. 14, no. 5, pages 759–767, May 2010. (Cited on page 8.)

[Raja 1998] Y. Raja, S.J. McKenna and Shaogang Gong. *Tracking and segmenting people in varying lighting conditions using colour*. In Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998. Proceedings, pages 228–233, April 1998. (Cited on page 71.)

[Ramesh 2013] B. Ramesh, Madhuri Vidyashankar and Pooja Sharma Dimri. Single Port Laparoscopic Surgery in Gynecology. JP Medical Ltd, September 2013. (Cited on page 12.)

[Rattner 2006] D. Rattner, A. Kalloo and ASGE/SAGES Working Group. *ASGE/SAGES Working Group on Natural Orifice Translumenal Endoscopic Surgery. October 2005*. Surg Endosc, vol. 20, no. 2, pages 329–333, February 2006. (Cited on page 8.)

[Reddy 2004] N. Reddy and P. Rao. *Per oral transgastric endoscopic appendectomy in human*. In Proceedings of the 45th Annual Conference of the Society of Gastrointestinal Endoscopy of India, pages 28–29, 2004. (Cited on page 8.)

[Reichl 2013] Tobias Reichl, José Gardiazabal and Nassir Navab. *Electromagnetic servoing-a new tracking paradigm*. 2013. (Cited on page 31.)

[Reilink 2011] Rob Reilink, Stefano Stramigioli and Sarthak Misra. *Three-dimensional pose reconstruction of flexible instruments from endoscopic images*. In Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on, pages 2076–2082, 2011. (Cited on page 33.)

[Reilink 2012] R. Reilink, S. Stramigioli and S. Misra. *Pose reconstruction of flexible instruments from endoscopic images using markers.* In 2012 IEEE International Conference on Robotics and Automation (ICRA), pages 2938–2943, 2012. (Cited on pages 33, 62, 94, 95 and 112.)

[Reiter 2011] Austin Reiter, Roger E. Goldman, A. Bajo, K. Iliopoulos, N. Simaan and P.K. Allen. *A learning algorithm for visual pose estimation of continuum robots.* In 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 2390–2396, 2011. (Cited on page 33.)

[Reiter 2012] A. Reiter, A. Bajo, K. Iliopoulos, N. Simaan and P.K. Allen. *Learning-based configuration estimation of a multi-segment continuum robot.* In 2012 4th IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob), pages 829–834, 2012. (Cited on page 33.)

[Robinson 1999] G. Robinson and J. Bruce C. Davies. *Continuum robots-a state of the art.* In Robotics and Automation, 1999. Proceedings. 1999 IEEE International Conference on, volume 4, pages 2849–2854. IEEE, 1999. (Cited on page 27.)

[Roesthuis 2013] R.J. Roesthuis, S. Janssen and S. Misra. *On using an array of fiber Bragg grating sensors for closed-loop control of flexible minimally invasive surgical instruments.* In 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 2545–2551, November 2013. (Cited on page 31.)

[Rosa 2011] B. Rosa, B. Herman, J. Szewczyk, B. Gayet and G. Morel. *Laparoscopic optical biopsies: In vivo robotized mosaicing with probe-based confocal endomicroscopy.* In 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 1339–1345, September 2011. (Cited on page 28.)

[Schapire 1990] Robert E. Schapire. *The strength of weak learnability.* Machine learning, vol. 5, no. 2, pages 197–227, 1990. (Cited on page 70.)

[Seneci 2014] C.A. Seneci, Jianzhong Shang, K. Leibrandt, V. Vitiello, N. Patel, A. Darzi, J. Teare and Guang-Zhong Yang. *Design and evaluation of a novel flexible robot for transluminal and endoluminal surgery.* In 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2014), pages 1314–1321, September 2014. (Cited on page 18.)

[Shang 2012] Jianzhong Shang, Christopher J. Payne, James Clark, David P. Noonan, Ka-Wai Kwok, Ara Darzi and Guang-Zhong Yang. *Design of a multi-tasking robotic platform with flexible arms and articulated head for minimally invasive surgery.* In Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ

International Conference on, pages 1988–1993. IEEE, 2012. (Cited on page 28.)

[Simaan 2004] N. Simaan, Russell Taylor and P. Flint. *A dexterous system for laryngeal surgery.* In 2004 IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04, volume 1, pages 351–357 Vol.1, April 2004. (Cited on page 18.)

[Slama 1980] Chester C. Slama, editeur. Manual of Photogrammetry. American Society for Photogrammetry and Remote Sen, Falls Church, Va, 4th revised edition edition édition, December 1980. (Cited on page 56.)

[Stark 2008] Michael Stark and Tahar Benhidjeb. *Natural Orifice Surgery: Transdouglas Surgery—a New Concept.* JSLS, vol. 12, no. 3, pages 295–298, 2008. (Cited on page 14.)

[Stewart 1999] Charles V. Stewart. *Robust parameter estimation in computer vision.* SIAM review, vol. 41, no. 3, pages 513–537, 1999. (Cited on page 82.)

[Stilli 2014] A. Stilli, H.A. Wurdemann and K. Althoefer. *Shrinkable, stiffness-controllable soft manipulator based on a bio-inspired antagonistic actuation principle.* In 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2014), pages 2476–2481, September 2014. (Cited on page 25.)

[Sturm 1999] P.F. Sturm and S.J. Maybank. *A Method for Interactive 3D Reconstruction of Piecewise PlanarObjects from Single Images.* In Proceedings or the Tenth British Machine Vision Conference, pages 27.1–27.10. British Machine Vision Association, 1999. (Cited on page 42.)

[Su 2012] Hao Su, D.C. Cardona, Weijian Shang, A. Camilo, G.A. Cole, D.C. Rucker, R.J. Webster and G.S. Fischer. *A MRI-guided concentric tube continuum robot with piezoelectric actuation: A feasibility study.* In 2012 IEEE International Conference on Robotics and Automation (ICRA), pages 1939–1945, May 2012. (Cited on page 29.)

[Swanström 2008] Lee L Swanström, Yashodan Khajanchee and Maher A Abbas. *Natural Orifice Transluminal Endoscopic Surgery: The Future of Gastrointestinal Surgery.* Perm J, vol. 12, no. 2, pages 42–47, 2008. (Cited on page 8.)

[Szewczyk 2001] Jèrôme Szewczyk, V. De Sars, Ph Bidaud and G. Dumont. *An active tubular polyarticulated micro-system for flexible endoscope.* In Experimental Robotics VII, pages 179–188. Springer, 2001. (Cited on page 28.)

[Tatlicioglu 2007] Enver Tatlicioglu, Ian D. Walker and Darren M. Dawson. *Dynamic modelling for planar extensible continuum robot manipulators.* In

Robotics and Automation, 2007 IEEE International Conference on, pages 1357–1362. IEEE, 2007. (Cited on page 30.)

[Tomasi 1992] Carlo Tomasi and Takeo Kanade. *Shape and motion from image streams under orthography: a factorization method.* International Journal of Computer Vision, vol. 9, no. 2, pages 137–154, 1992. (Cited on page 42.)

[Torres 2012] L.G. Torres, R.J. Webster and R. Alterovitz. *Task-oriented design of concentric tube robots using mechanics-based models.* In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 4449–4455, October 2012. (Cited on page 29.)

[Traeger 2014] Mattias F. Traeger, Daniel B. Roppenecker, Matthias R. Leininger, Florian Schnoes and Tim C. Lueth. *Design of a spine-inspired kinematic for the guidance of flexible instruments in minimally invasive surgery.* In Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on, pages 1322–1327. IEEE, 2014. (Cited on page 18.)

[Utcke 2003] Sven Utcke and Andrew Zisserman. *Projective reconstruction of surfaces of revolution.* In Pattern Recognition, pages 265–272. Springer, 2003. (Cited on page 42.)

[Weber 2012] Bernhard Weber, Paul Zeller and K. Kuhnlenz. *Multi-camera based real-time configuration estimation of continuum robots.* In Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on, pages 3350–3355. IEEE, 2012. (Cited on page 32.)

[Webster 2009] R.J. Webster, J.M. Romano and N.J. Cowan. *Mechanics of Precurved-Tube Continuum Robots.* IEEE Transactions on Robotics, vol. 25, no. 1, pages 67–78, February 2009. (Cited on page 30.)

[Wong 2004] Kwan-Yee K. Wong, Paulo RS Mendonça and Roberto Cipolla. *Reconstruction of surfaces of revolution from single uncalibrated views.* Image and Vision Computing, vol. 22, no. 10, pages 829–836, 2004. (Cited on page 42.)

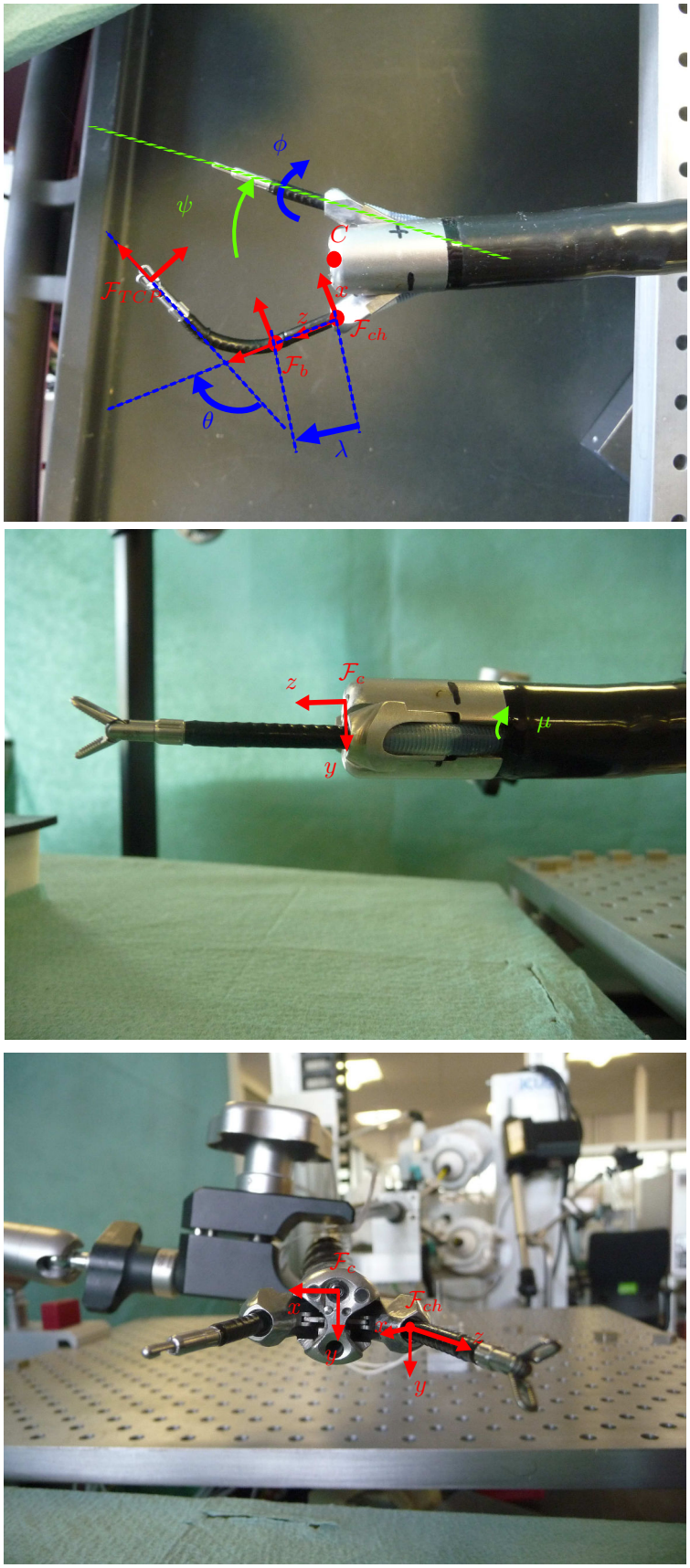
[Xu 2006] Kai Xu and N. Simaan. *Actuation compensation for flexible surgical snake-like robots with redundant remote actuation.* In Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006, pages 4148–4154, May 2006. (Cited on page 30.)

[Yip 2014] M.C. Yip and D.B. Camarillo. *Model-Less Feedback Control of Continuum Manipulators in Constrained Environments.* IEEE Transactions on Robotics, vol. 30, no. 4, pages 880–889, August 2014. (Cited on pages 31 and 32.)

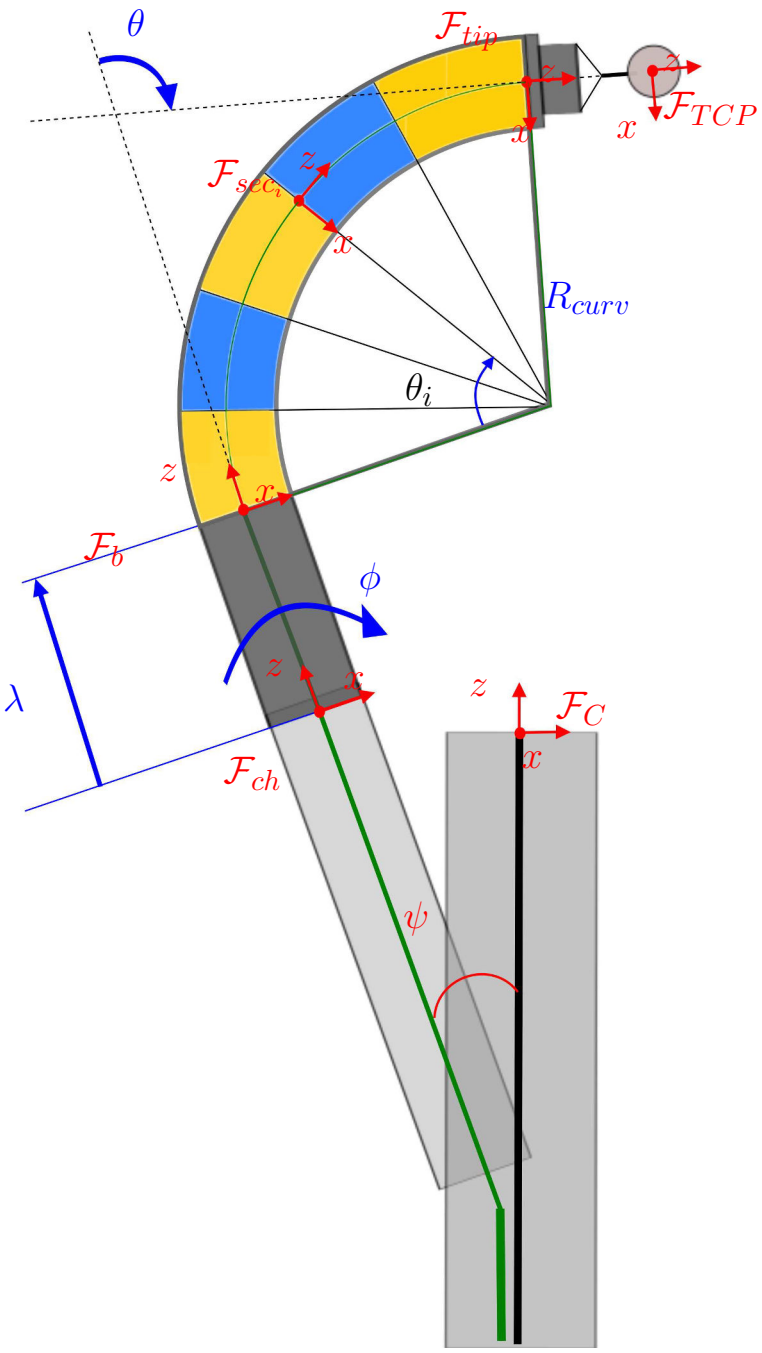
[Zhang 1999] Zhengyou Zhang. *Flexible camera calibration by viewing a plane from unknown orientations.* In Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, volume 1, pages 666–673. Ieee, 1999. (Cited on page 56.)

[Zhao 2010] Qiang Zhao and Fang Gao. *Design and analysis of a kind of biomimetic continuum robot.* In 2010 IEEE International Conference on Robotics and Biomimetics (ROBIO), pages 1316–1320, December 2010. (Cited on page 28.)

[Zorrón 2007] Ricardo Zorrón, Marcos Filgueiras, Luís Carlos Maggioni, Luciana Pombo, Gustavo Lopes Carvalho and Andre Lacerda Oliveira. *NOTES. Transvaginal cholecystectomy: report of the first case.* Surg Innov, vol. 14, no. 4, pages 279–283, December 2007. (Cited on page 8.)


ϕ
ψ
C
$\mathcal{F}_{TCP}$
x
z
$\mathcal{F}_{ch}$
$\mathcal{F}_b$
θ
λ
z
$\mathcal{F}_c$
μ
y
$\mathcal{F}_c$
x
$\mathcal{F}_{ch}$
z
y

Tip of STRAS.



Scheme of the tip of STRAS with the left instrument.

Nomenclature

λ	Instrument translation (Natural DOF)
ϕ	Instrument rotation around channel axis (Natural DOF)
θ	Instrument deflection angle (Natural DOF)
(x_{ch}, y_{ch})	Channel position on the camera plane (Spurious DOF)
ψ	Opening angle of the instrument wrt the Camera axis (Spurious DOF)
μ	Tilt angle of the instrument wrt the Camera axis (Spurious DOF)
$\mathcal{F}_{[\cdot]}$	Reference frame associated to the different parts of the kinematic/geometrical chain.
$t^c_{a,b}$	Column vector describing the translation from $\mathcal{F}_a$ to $\mathcal{F}_b$ expressed wrt to $\mathcal{F}_c$
R^b_a	Matrix representing the orientation of $\mathcal{F}_a$ wrt $\mathcal{F}_b$
$R_x(\cdot), R_y(\cdot), R_z(\cdot)$	Matrix expressing the rotation around (respectvely) x, y and z of the angle indicated as argument
$\mathbf{P_i}^a$	3D point corresponding to the i-th section of the instrument bendable part (determined by the markers) expressed wrt $\mathcal{F}_a$
$\mathbf{P_{i,u}}$ $\mathbf{P_{i,l}}$	3D point corresponding to the i-th *upper (lower)* apparent corner defined by the i-th section
$\mathbf{p_{i,u}}$ $\mathbf{p_{i,l}}$	i-th *upper (lower)* apparent corner defined by the i-th section
$\hat{\cdot}$	estimated value of the argument

K	Number of clusters
M	Number of elements in the training set
N_k	Number of elements in each cluster: $\sum_k N_k = M$
$\mathbf{C}_{k,i}$	Center of gravity of each cluster k in the estimation problem of the i-th 3D coordinate.
$\gamma_{k,i}$	Coefficient weighting the influence of the k-th kernel for the estimation of the i-th 3D coordinate.
$\mathbf{y}$	function output: in our case the three coordinates of the TCP 3D position (X, Y, Z)
$\mathbf{x}$	function input: in our case the coordinates of the n corners position in the image $(x_1, \cdots, x_{2n})$
$\mathbf{x_m}$	Input samples of the training set $m = 1 \cdots M$
$\mathbf{x_{m_k}}$	Elements of the training set pertaining to cluster k
$\hat{y}_i$	Estimation of the i-th coordinate $y_1 = X$, $y_2 = Y$ and $y_3 = Z$
$\hat{y}_{m,i}$	i-th component of the output sample corresponding to $\mathbf{x}_m$

UNIVERSITÉ DE STRASBOURG

UNIVERSITÉ DE STRASBOURG

RÉSUMÉ EN FRANÇAIS DE LA THÈSE DE DOCTORAT

Discipline : Signal, Image, Automatique, Robotique

Spécialité (facultative) :

Titre : **Mesure par vision de la position d'instruments médicaux flexibles pour la chirurgie endoscopique robotisée.**

Unité de Recherche : UMR CNRS 7357 - Lab. ICube Equipe AVR

Directeur de thèse : DOIGNON Christophe, PR

Encadrant de thèse : NAGEOTTE Florent, PhD MdC

Localisation : 300 bd Sébastien Brant - CS 10413 - F-67412 Illkirch Cedex

ECOLES DOCTORALES

cocher la case

☐ ED 519-Sciences Humaines et Sociales	☑ ED 269-Mathématiques, Sciences de
☐ ED 520-Humanités	l'Information et de l'Ingénieur
☐ ED 101-Droit, Sciences politiques et histoire	☐ ED 270-Théologie et sciences religieuses
☐ ED 182-Physique et chimie physique	☐ ED 413-Sciences de la terre, de l'univers et
☐ ED 221-Augustin Cournot	de l'environnement
☐ ED 222-Sciences chimiques	☐ ED 414 - Sciences de la vie et de la santé

1 Introduction

Avec les progrès des techniques de chirurgie minimalement invasive, les systèmes flexibles et à sections continues sont maintenant utilisés pour des opérations chirurgicales. En chirurgie endoscopique flexible, les instruments utilisés sont longs, fins et leur extrémité distale pliable est habituellement actionnée par câbles. Leur manipulation est complexe et une assistance robotique est intéressante pour l'utilisation pendant des tâches chirurgicales [5].

La commande robotique de ces systèmes flexibles continus est difficile. En effet, l'information fournie par les capteurs proximaux n'est pas fiable en raison des interactions mécaniques complexes entre câbles de transmissions et gaines. Par ailleurs, intégrer des modèles de ces interactions dans la commande résulte en général en des solutions très spécifiques difficilement transférables à d'autres systèmes.

Pour implémenter des modes de commande efficaces, il serait souhaitable de fermer la boucle en utilisant des capteurs externes. Cela est notamment nécessaire pour permettre des mouvements autonomes et améliorer les modes de télémanipulation des robots flexibles.

En raison de contraintes liées à la taille, à la compatibilité avec l'environnement chirurgical et à la résistance à la déformation, il n'existe pas de capteurs standard qui puissent être utilisés in vivo pour les systèmes flexibles. Par conséquent, utiliser la caméra endoscopique embarquée semble être une façon intéressante de mesurer la position des instruments, tout en fournissant des informations relatives à l'environnement. Cette approche a été utilisée précédemment dans [9]. Toutefois, les méthodes décrites dans ce travail ne fonctionnent correctement que si le modèle géométrique du système est parfaitement connu, ce qui n'est pas le cas en pratique.

1.1 Objectifs

Notre objectif est de proposer des outils pour l'estimation de la position 3D d'instruments flexibles, qui puissent être utilisés dans un environnement chirurgical réel, et qui se basent principalement sur les images de la caméra endoscopique embarquée. De plus on souhaite qu'ils soient adaptés aux contraintes suivantes :

- système de vision monoculaire,
- instruments flexibles à section continue béquillable non texturée,
- point de vue défavorable (cas des endoscopes classiques utilisés en chirurgie) : azimut et élévation faibles entre les instruments et l'axe optique de la caméra,
- transmission mécanique par câbles sur de grandes longueurs, ce qui signifie que les mesures données par les codeurs proximaux ne représentent pas correctement la position distale,
- le bout de l'instrument n'est habituellement pas visible dans les images endoscopiques car il se trouve en contact avec les tissus.

Ces conditions sont notamment représentatives des systèmes flexibles utilisés actuellement en chirurgie mini-invasive (aussi bien manuelle que robotique), comme par exemple le système robotique STRAS (ICube / Storz) présenté en Fig. 1(a) [4].

Après avoir discuté la modélisation cinématique des instruments flexibles et le banc de test expérimental utilisé dans ce travail, nous présentons les trois principales parties de la thèse : la détermination et l'extraction de primitives pour l'estimation de pose (sec. 3), une approche basée modèle permettant une estimation de pose robuste aux variations de modèle (sec. 4) et l'étude d'une approche alternative, sans modèle, basée sur

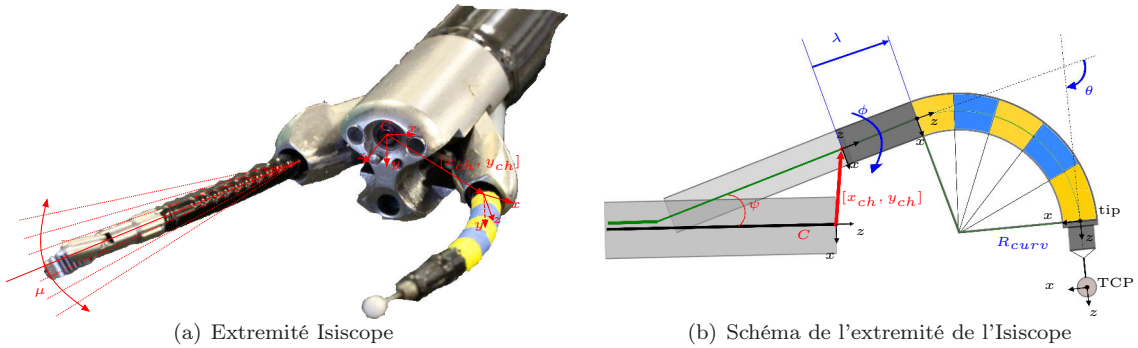
l'apprentissage de la relation entre des primitives visuelles (entrée) et la position 3D de l'instrument (sortie) (sec. 5).

2 Cellule expérimentale et système cible

La partie distale du système Anubis est composée d'une caméra endoscopique et de deux instruments qui sortent de deux canaux inclus dans le corps de l'endoscope (Fig. 1(a)). Ces instruments ont trois degrés de liberté (DDL) : translation et rotation par rapport à l'axe du canal et flexion (indiqués respectivement par λ, ϕ and θ sur la Fig. 1(b)). La partie pliable de l'instrument est composée de multiples vertèbres parcourues par deux câbles antagoniste : quand le câble attaché à la vertèbre la plus distale est tiré, toute la partie pliable change de courbure de sorte à minimiser son énergie mécanique. Ces systèmes flexibles sont normalement modélisés par des tronçons de tore avec un angle de courbure (θ) variable [10].

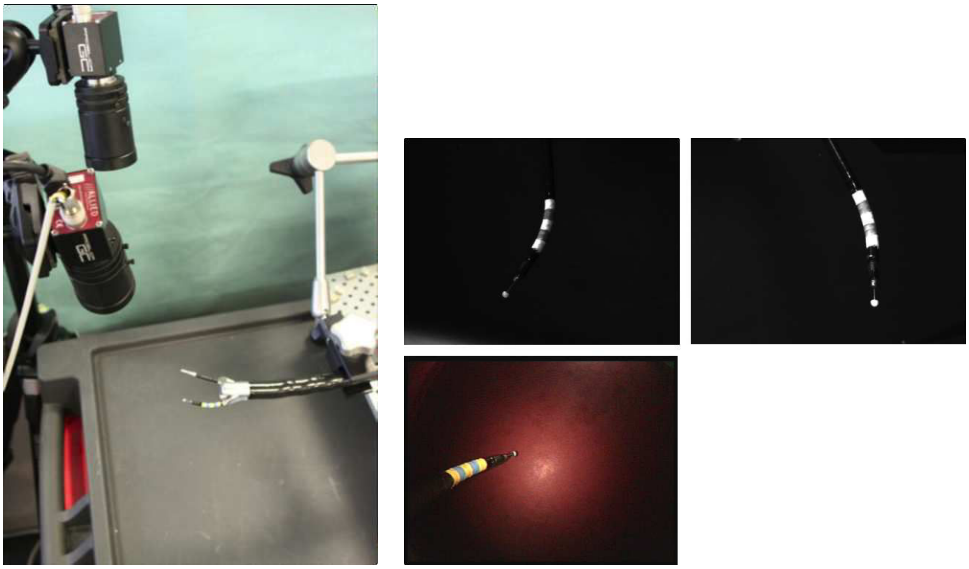
Pour décrire entièrement le système, quatre autres paramètres (appelés *mécaniques*) ont été introduits : la position de la terminaison du canal de l'instrument ($[x_{ch}, y_{ch}]$) et l'orientation de ce dernier par rapport à l'axe optique (z) de la caméra, décrite comme la composition de deux rotations consécutives : la première d'angle ψ autour de l'axe y de la camera et la deuxième d'un angle μ autour de l'axe x résultant.

Afin d'évaluer la précision des méthodes présentées par la suite, une cellule expérimentale a été conçue, composée de la plateforme STRAS [4] et deux caméras conventionnelles positionnées au-dessus du système robotique et formant un système de vision stéréoscopique. Pour chaque configuration d'un des instruments flexibles, une image est prise par la camera endoscopique et simultanément par les caméras du système stéréo (Fig. 2). La précision de mesure endoscopique est évaluée comme la distance 3D entre la position du bout de l'instrument (TCP) estimée à partir des indices visuels de l'image endoscopique et la vérité terrain donnée par le système stéréo. Le choix d'utiliser un point de référence distant de l'extrémité de la partie pliable (la boule blanche visible en Fig. 1(a)) permet de valider non seulement l'estimation de la position mais aussi de l'orientation du corps de l'instrument.



(a) Extremité Isiscope (b) Schéma de l'extremité de l'Isiscope

FIGURE 1 – En (a) la photo de la pointe de l'endoscope avec les instruments et en (b) le schéma correspondant.



(a) (b)

FIGURE 2 – (a) : cellule expérimentale composée de la plateforme STRAS et deux caméras conventionnelles positionnées au-dessus du système robotique et formant un système de vision stéréoscopique ; (b) : pour chaque configuration une image est prise par la camera endoscopique (bas) et simultanément par les caméras du système stéréo (haut).

3 Choix des primitives visuelles pour un instrument flexible

La précision des méthodes basées image utilisant une caméra monoculaire dépend fortement de la capacité des primitives choisies à décrire l'effet perspectif. En ces termes, le concept de pertinence d'une primitive pour l'estimation de pose peut être défini en se basant sur la sensibilité de l'estimation de pose 3D par rapport à l'erreur sur la segmentation des primitives dans l'image.

Une étude préliminaire a été menée afin de déterminer les primitives les plus adaptées pour ce type de problèmes. Considérant le fait que le bout de l'instrument est souvent caché pendant l'opération, seules des primitives attachées à la section flexible de l'instrument ont été prises en comptes, tout en incluant la possibilité d'utiliser des marqueurs.

Cette étude de propagation d'incertitude a suggéré l'adoption d'indices visuels liés aux bords apparents de l'instrument, à condition qu'ils aient une véritable signification physique par rapport à l'instrument. Le choix d'utiliser des marqueurs est assez commun dans le contexte des opérations in-vivo en raison du besoin de robustesse à la présence de fumée ou de fluides. Des marqueurs de dimensions connues enroulés autour du corps de l'instrument ont donc été utilisés et leurs coins ont été sélectionnés pour servir de primitives. Ces marqueurs permettent la segmentation des primitives quelle que soit la rotation de l'instrument.

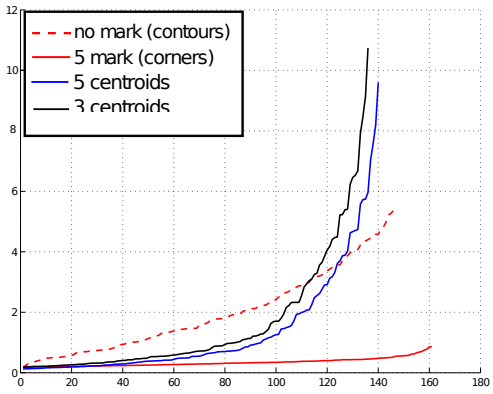


FIGURE 3 – Pour chaque configuration du robot, la direction avec la plus grande variance sur l'estimation de la position du bout de l'instrument est sélectionnée et les normes des erreurs obtenues sont ordonnées. Le résultat pour différentes primitives est montré, la meilleure solution consistant à considérer les coins apparent des différents marqueurs (5 marqueurs dans ce cas).

4 Extraction des primitives d'un instrument flexible

Une fois les primitives sélectionnées, l'autre point clef pour obtenir un bon résultat d'estimation est d'atteindre une robustesse et une précision d'extraction suffisantes. Cette tâche présente plusieurs difficultés, en particulier pour des instruments déformables et dans des conditions in-vivo. Comme cela est montré dans la fig. 4, l'illumination intense (nécessaire pour avoir une bonne visibilité de toute la scène) provoque de fortes spécularités qui effacent toute l'information de couleur ou de texture présente sur le corps de l'instrument. En outre, les reflets qui se produisent sur les organes peuvent amener à la détection de faux positifs. Enfin, le mouvement respiratoire rend inefficace les techniques de soustraction de fond.

A notre connaissance, il n'existe pas de méthodes de segmentation pour ce type d'instruments et, en particulier, dans ces conditions. Nous avons donc proposé une nouvelle approche composée de 4 étapes (Fig. 5). Après une segmentation basée couleur qui extrait des régions candidates (1ère étape), une méthode basée sur des graphes permet d'assigner, selon des critères topologiques et de forme, une étiquette indiquant à quelle classe (c'est-à-dire à quel marquer) appartient chaque région (2ème étape). Les points des bords apparents supérieur et inférieur sont alors extraits (3ème étape) et filtrés en leur ajustant une courbe de

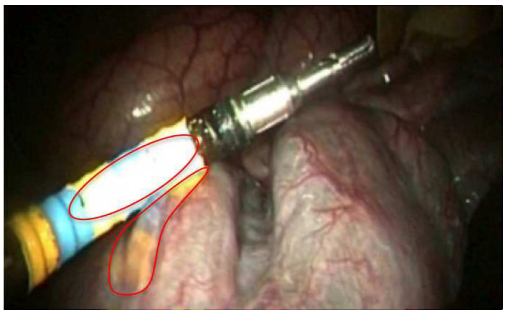


FIGURE 4 – Exemple de scène in-vivo capturée par la caméra endoscopique. On peut observer la forte spécularité au milieu de l'instrument due à l' illumination intense et le reflet de l'instrument sur l'organe.

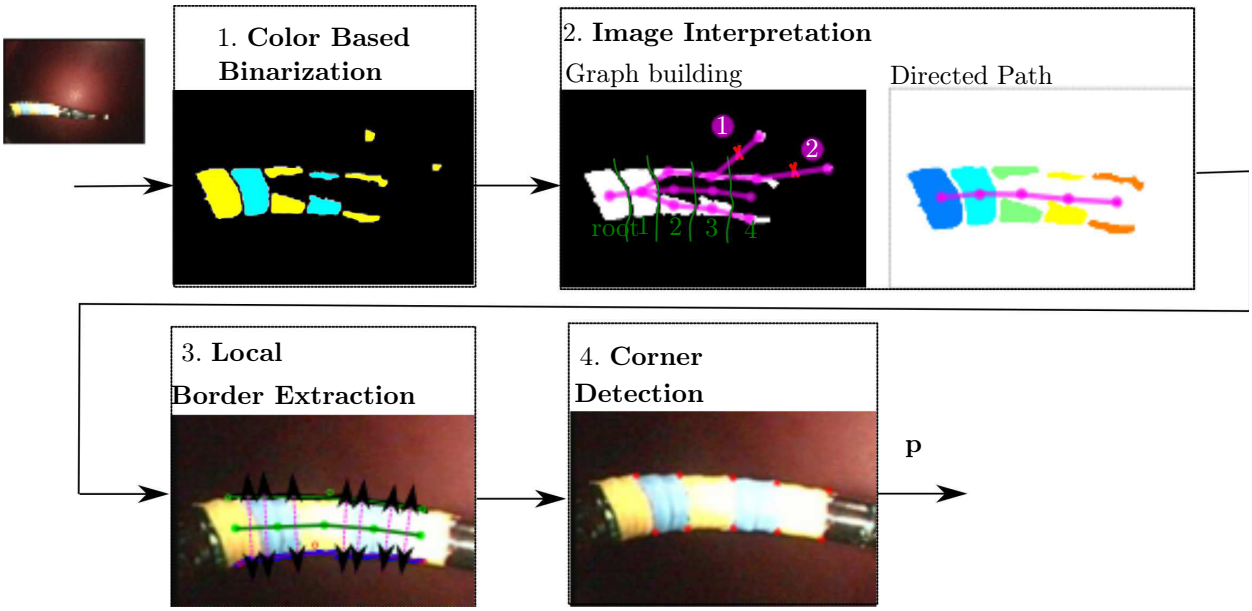


FIGURE 5 – Étapes de la méthode de segmentation. (1) Le résultat donné par les modèles Gaussiens des couleurs est binarisé pour déterminer les régions jaune et bleu dans l'image. (2) Création d'un arbre en utilisant la connectivité spécifiée : la branche 2 n'est pas connectée parce qu'elle ne respecte pas la condition 1 ; la branche 1 ne respecte pas la condition 2 parce que l'orientation de la branche diffère sensiblement de l'orientation des branches qui la précèdent. Finalement, l'arbre est traité pour obtenir un chemin orienté. (3) Le squelette est utilisé pour chercher localement les points de bords, en recherchant les maximas locaux du gradient le long des perpendiculaires à la section du squelette. Ces points candidats sont ensuite décrits avec des courbes de Bézier . Les coins apparents des marqueurs sont finalement en recherchant les discontinuités de couleur le long des courbes de Bézier (4).

Bézier , en se basant sur la théorie des M-estimateurs. Finalement, les coins des marqueurs sont extraits comme les transitions entre les couleurs bleu et jaune le long des courbes de Bézier estimées (4ème étape). Grâce à cette définition continue des bords apparents, une précision sub-pixelique de la position des coins peut être attendue.

4.1 Interprétation de l'image

Le coeur de ce procédé est la deuxième étape, qui permet de recomposer le corps de l'instrument dans l'image, en assignant à chaque région candidate une étiquette représentant le marqueur auquel elle appartient. Cette phase se base sur le concept qu'un objet peut être représenté comme un ensemble de parties connectées entre elles. Dans notre cas, la partie pliable de l'instrument peut être vue comme la séquence ordonnée des cinq marqueurs. Toutefois, en raison des fortes spécularités, le résultat de la segmentation basée couleur présente plus de régions candidates que de marqueurs (cf. stage 1 fig. 5).

Cette étape vise donc à résoudre ce problème en réassemblant les régions dans les parties originales de l'objet de sorte à obtenir l'objet complet. Elle est composée de deux phases : la construction de l'arbre et son traitement.

Pour mener à bien la première phase, il n'est pas suffisant de considérer une connectivité classique à 4

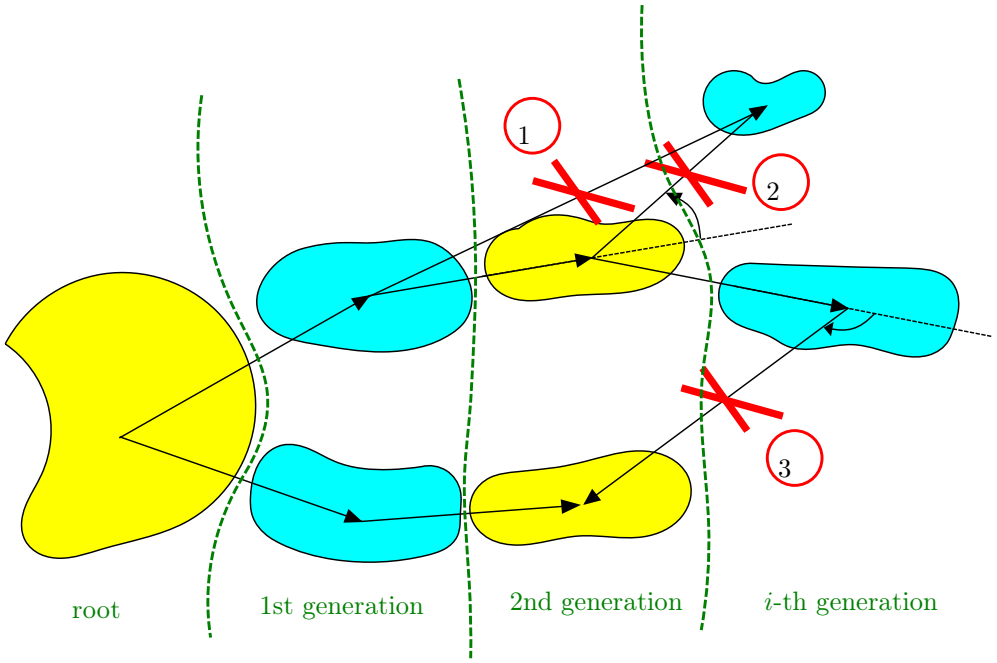


FIGURE 6 – Construction de l'arbre avec la connectivité spécifiée. Dans le premier cas, la branche n'est pas créée parce qu'elle ne respecte pas la condition 1. Dans le cas 2 et 3, c'est la condition 2 qui n'est pas respectée, parce que l'orientation de la branche considérée diffère sensiblement de l'orientation donnée par les branches qui la précèdent. Cela permet de ne pas connecter des faux-positifs (cas 2) ou de former des boucles internes qui correspondent à configurations qui n'ont aucune signification physique (cas 3).

ou 8 voisins pour décrire la relation entre les régions. Une nouvelle connectivité a donc été définie, basée sur deux caractéristiques structurelles de la section béquillable, c'est à dire :

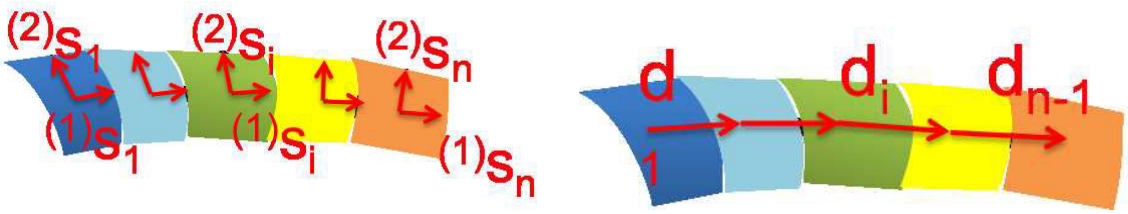
1. Les marqueurs sont adjacents l'un avec l'autre et les couleurs bleu et jaune s'alternent ;
2. Les centres des différents marqueurs dans l'image doivent définir une courbe qui ne présente pas de fortes discontinuités de direction.

Plus spécifiquement, en commençant de la racine, un nouveau nœud est connecté a son parent s'il respecte ces deux propriétés (fig. 6). Construire l'arbre de cette manière permet d'ordonner les régions de la base au bout et d'obtenir, à chaque niveau (cfg. fig. 6), toutes les régions candidates pouvant être associées à un même marqueur.

Une fois l'arbre obtenu, pour chaque niveau les nœuds sont considérés deux à deux et un nœud artificel est créé, qui représente la région résultant de la fusion possible entre les deux régions correspondant aux nœuds initaux. On obtient ainsi l'arbre montré en fig. 8 (en haut).

Cependant, l'objectif final est d'obtenir un chemin dirigé composé par les marqueurs de la partie pliable de l'instrument. Un critère de cohérence temporelle est utilisé pour traiter l'arbre et en extraire ce chemin. En supposant que, d'une image à la suivante, la forme de l'instrument reste substantiellement la même, l'apparence de chaque marqueur peut être décrite par deux facteurs :

– un *facteur de forme* dépendant du moment de second ordre de la région (calculé selon ses axes principaux) : $s^{(1)}$ et $s^{(2)}$;



(a) (b)

FIGURE 7 – L'apparence de chaque marqueur peut être décrite par deux facteurs : un facteur de forme et une facteur topologique. Le premier est défini comme la longueur et largeur de chaque marqueur selon ses axes principaux (gauche) et le deuxième comme la distance du marqueur considéré à son prédécesseur (droite).

– un *facteur topologique* dépendant de la distance du nœud considéré (réel ou artificiel) à son nœud parent (ou région) : d.

Ces deux facteurs ont été modélisés comme des variables stochastiques. En définissant a comme l'événement qu'une région soit un marqueur particulier A, s l'événement qu'elle ait la forme S (facteur de forme) et d l'événement que la distance à son nœud parent soit D (facteur topologique), la probabilité conditionnelle est définie par : $p(A|S,D)$.

En première approximation, nous pouvons considérer que, connaissant le marqueur A, les événements de forme et de topologie sont des variables Gaussiennes qui peuvent être définies comme :

$$p(S^{(1)}|A) \sim \frac{1}{2\pi\sigma_{s_a}^{(1)}} exp(-\frac{(s_i-\mu_{s_a}^{(1)})^2}{2[\sigma_{s_a}^{(1)}]^2}) \quad (1)$$

$$p(S^{(2)}|A) \sim \frac{1}{2\pi\sigma_{s_a}^{(2)}} exp(-\frac{(s_i-\mu_{s_a}^{(2)})^2}{2[\sigma_{s_a}^{(2)}]^2}) \quad (2)$$

pour la forme (fig. 7(a)) et

$$p(D|A) \sim \frac{1}{2\pi\sigma_{d_a}} exp(-\frac{(d_i-\mu_{d_a})^2}{2\sigma_{d_a}^2}) \quad (3)$$

pour la topologie (fig. 7(a)), où μ et σ sont respectivement la moyenne et l'écart type des valeurs observées dans un passé proche :

– du moment central de second ordre calculé le long des composantes principales du marqueurs en question $(\mu_{s_a}^{(1)}, \sigma_{s_a}^{(1)})$ et $(\mu_{s_a}^{(2)}, \sigma_{s_a}^{(2)})$,
– de la distance Euclidienne depuis le centre de la région parent $(\mu_{d_a}, \sigma_{d_a})$.

Ces valeurs sont déterminées pour chaque marqueur (de 1 à 5). Pour tenir en compte de la variation de taille et de forme perçues dans l'image pendant le mouvement de l'instrument, ces paramètres sont actualisés à chaque image en considérant les observations sur une fenêtre glissante des 10 dernières images.

A partir des caractéristiques d'un nœud (réel ou artificiel), on peut alors calculer la valeur de la probabilité de ce nœud d'être le marqueur du niveau considéré :

$$p(A|S,D) \sim \frac{1}{2}(p(S^{(1)}|A) + p(S^{(2)}|A))\, p(D|A). \quad (4)$$

A l'exception de la racine (le marqueur de la base de la partie pliable), cette probabilité conditionnelle peut être considérée comme une récompense associée à chaque branche de l'arbre en provenance du niveau

supérieur. L'arbre peut être traité niveau par niveau (cf. Fig. 8) en sélectionnant la branche avec la probabilité la plus haute. Dans le cas d'un nœud artificiel, les facteurs de forme et topologique sont calculés en considérant une région résultant de la fusion des deux régions candidates impliquées. Si la région avec la plus haute probabilité est celle associée à un nœud artificiel, les deux nœuds impliqués sont éliminés et remplacés par le nœud artificiel.

Si le niveau présente plus de deux régions, le même procédé peut être appliqué itérativement jusqu'au moment où aucune meilleure solution ne peut être trouvée.

A la fin de ce procédé, l'arbre est une séquence de 5 centres, ordonnés de la base au bout de la section bequillable de l'instrument (deuxième étape en fig. 5).

Ce traitement définit une sorte de squelette de l'instrument, qui sert ensuite à détecter les points des bords supérieur et inférieur comme décrit auparavant. Ces points de bord sont ensuite approchés par des courbes de Bézier , puis les coins des marqueurs sont extraits en analysant les variations de couleurs le long de ces courbes.

Cette méthode a fourni de bons résultats, même en conditions in-vivo, lorsque l'instrument se déplace dans un espace libre et sans occultation. La méthode est robuste aux fortes spécularités et est capable de discriminer l'instrument de ses éventuels reflets sur les organes (Fig. 11(c)).

Comme cette méthode ne prend pas en compte de modèles de l'instrument, l'algorithme est peu robuste aux occlusions, ce qui pourrait être pris en considération dans des travaux futurs.

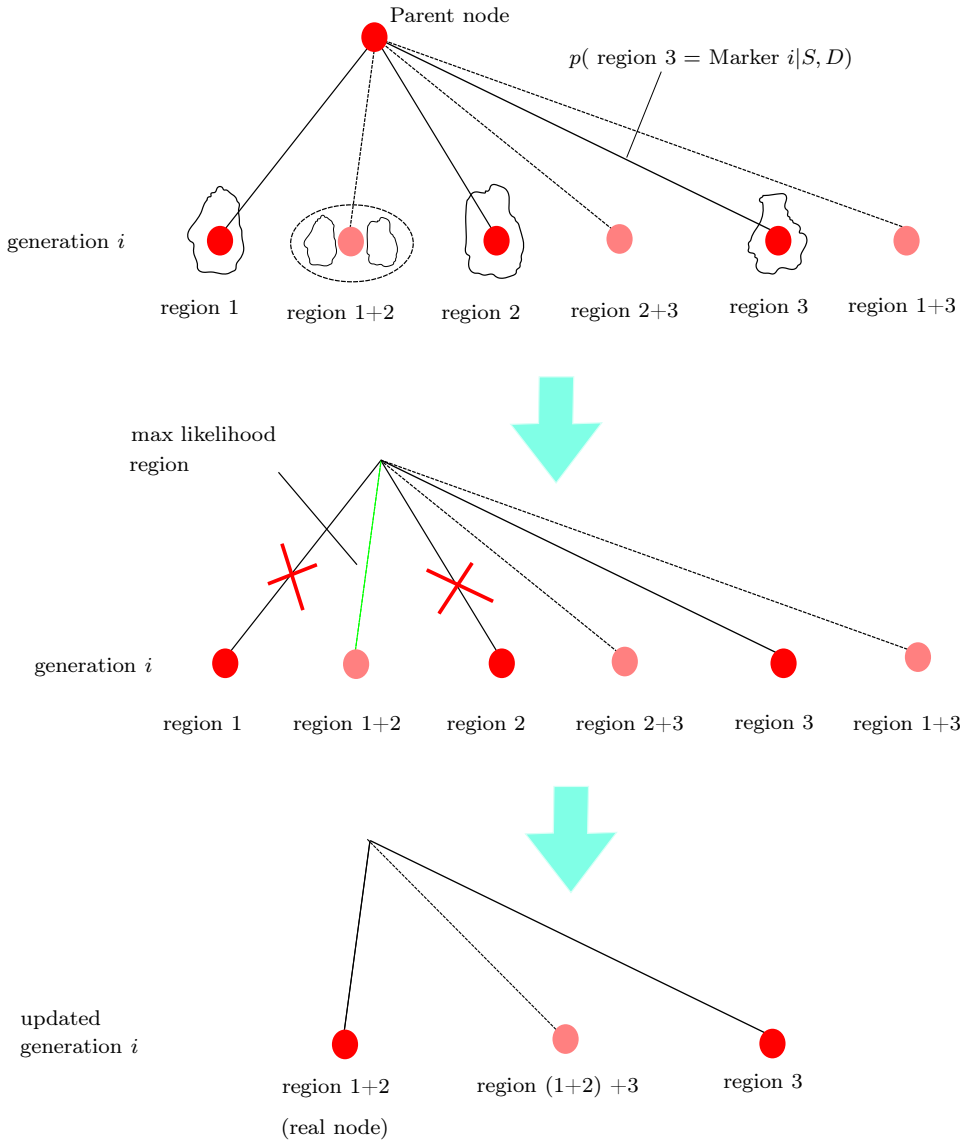


FIGURE 8 – Schéma du procédé de traitement de l'arbre. Pour chaque niveau, la branche avec la valeur la plus grande (celle associée à la région avec la plus grande probabilité d'être le marqueur du niveau considéré) est sélectionnée (colorée en vert). Si, comme dans le cas en question, la branche la plus probable est celle d'un nœud artificiel (le nœud qui représente l'union de deux régions candidates), les nœuds associés à ces deux régions sont effacés et le nœud artificiel est converti en un nœud réel. Si le niveau présente plus de deux régions, le même procédé est appliqué itérativement jusqu'à obtenir un seul nœud.

(a) (b)

FIGURE 9 – Résultats in-vivo de la méthode présentée. En (a), malgré les fortes spécularités, la structure de l'instrument est retrouvée. En (c), grâce à l'extraction locale des bords et à l'approximation robuste utilisant une fonction de coût Beaton-Tukey, la plupart des points aberrants sont éliminés et l'instrument est correctement discriminé de son reflet sur l'organe.

5 Approche basée modèle

La première approche que nous avons proposé pour estimer la pose de l'instrument à partir des primitives image consiste à utiliser un modèle géométrique paramétré du système.

En supposant que les modèles géométriques de l'instrument et de la caméra sont connus, une image synthétique (virtuelle) de la scène peut être calculée pour une configuration donnée de l'instrument. En conséquence, le problème de l'estimation de la pose 3D peut être formalisé comme un problème d'optimisation, dont l'objectif est de trouver la meilleure configuration de l'instrument selon un "critère visuel ", par exemple basé sur la différence entre l'image virtuelle courante (position des points dans l'image calculée) et l'image de référence (position des indices visuels dans l'image vidéo acquise) [7].

Nous avons montré que l'hypothèse considérant que l'état actuel de l'instrument dépend seulement de 3 DDL (comme dans [9]) n'est pas adaptée : en effet, certaines parties du modèle peuvent varier en raison, par exemple, des jeux mécaniques entre les instruments et les canaux qu'ils traversent. Ces jeux ne sont pas négligeables ni évitables puisque ils sont indispensables afin que l'instrument puisse se déplacer à l'intérieur du canal.

Nous avons donc décidé de considérer une tolérance sur les *paramètres mécaniques* et, en conséquence, nous avons développé une nouvelle méthode d'optimisation se fondant sur des fonctions de pondération, qui aide à contraindre l'estimation de ces paramètres dans un voisinage de leurs valeurs nominales. Le critère d'optimisation, donc, n'est plus strictement visuel, mais tient compte de ces contraintes sur les paramètres mécaniques :

$$\chi^2 = \frac{1}{2}\sum_{i=1}^{N} w_i \left[\mathbf{p_i} - \hat{\mathbf{p}}_\mathbf{i}\right]^2 + \rho_{x_{ch}}(x^*_{ch} - \hat{x}_{ch}) + \rho_{y_{ch}}(y^*_{ch} - \hat{y}_{ch}) + \rho_\psi(\psi^* - \hat{\psi}) + \rho_\mu(\mu^* - \hat{\mu}) \tag{5}$$

où $\mathbf{p_i}$ est l'état de référence et $\hat{\mathbf{p}}_\mathbf{i}$ est l'état estimé. Les astérisques indiquent les valeurs nominales des paramètres mécaniques et $\hat{}$ les valeurs estimées. Le choix de la forme de ρ dérive de la nature des tolérances.

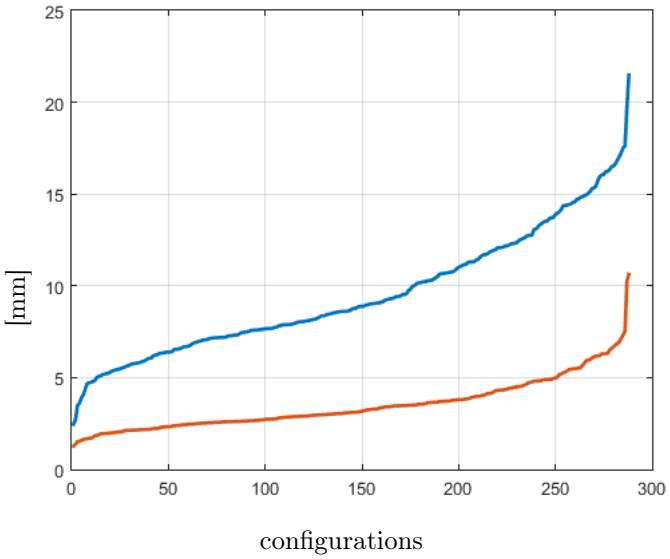


FIGURE 10 – Comparaison entre les deux approches basées modèle. Les deux courbes représentent les normes des erreurs commises sur l'estimation de la position 3D pour différentes configurations dans l'espace de travail et ordonnées de façon croissante. La considération des jeux mécaniques (courbe rouge) permet obtenir une estimation nettement meilleure à celle obtenue avec une méthode plus classique comme celle de Reilink (courbe bleu).

En fait, un jeu mécanique laisse un mouvement presque libre autour de la position de repos, alors que les contraintes deviennent fortes lorsque l'instrument s'éloigne de son état de repos (décrit par la valeur nominale fournie par le fabricant). Pour répliquer ce comportement, la fonction adéquate doit présenter une zone morte autour des valeurs nominales et une valeur croissant rapidement en dehors de cet intervalle. De plus, elle doit être au moins $\mathcal{C}^1$ dans tout son domaine pour pouvoir procéder à l'optimisation. Nous avons donc proposé d'utiliser le module d'une fonction cubique :

$$\rho(u) = \frac{k_w}{3a^3}|u|^3,\ a \in \mathbb{R}^+,\ k_w \in \mathbb{R}^+ \tag{6}$$

où k_w détermine la pente en dehors de la zone morte dont la largeur est définie par le paramètre a.

Dans le contexte robotique, les informations visuelles ont été fusionnées avec les données moteur de sorte à les compléter et les renforcer. Ceci est réalisé au travers d'un filtre de Kalman dont les confiances sur la mesure et l'état sont basées sur la quantité disponible d'information visuelle. En effet, si l'information visuelle n'est que partiellement disponible (i.e. seulement quelques coins sont détectés), la confiance du modèle devrait être augmentée au détriment de la confiance sur la mesure (résultat de l'asservissement virtuel).

Cette approche a été testée sur une séquence de 295 images acquises sur le dispositif expérimental. Le processus d'extraction de caractéristiques a échoué sur 16 images. L'erreur moyenne quadratique (RMS) obtenue sur chaque coordonnée (en mm) est : 1.87 ± 1.65 sur x , 2.12 ± 2.07 sur y et 2.69 ± 2.57 sur l'axe z. Ces erreurs ne diminuent que légèrement si les poses pour lesquelles l'information visuelle est incomplète sont exclues, ce qui démontre la capacité de cette approche à gérer des informations partielles. La précision est bien meilleure que celle qui peut être obtenue à partir des approches plus classiques (fig. 10).

Cette même méthode a été testée sur deux séquences vidéo in-vivo acquises dans l'abdomen d'un cochon.

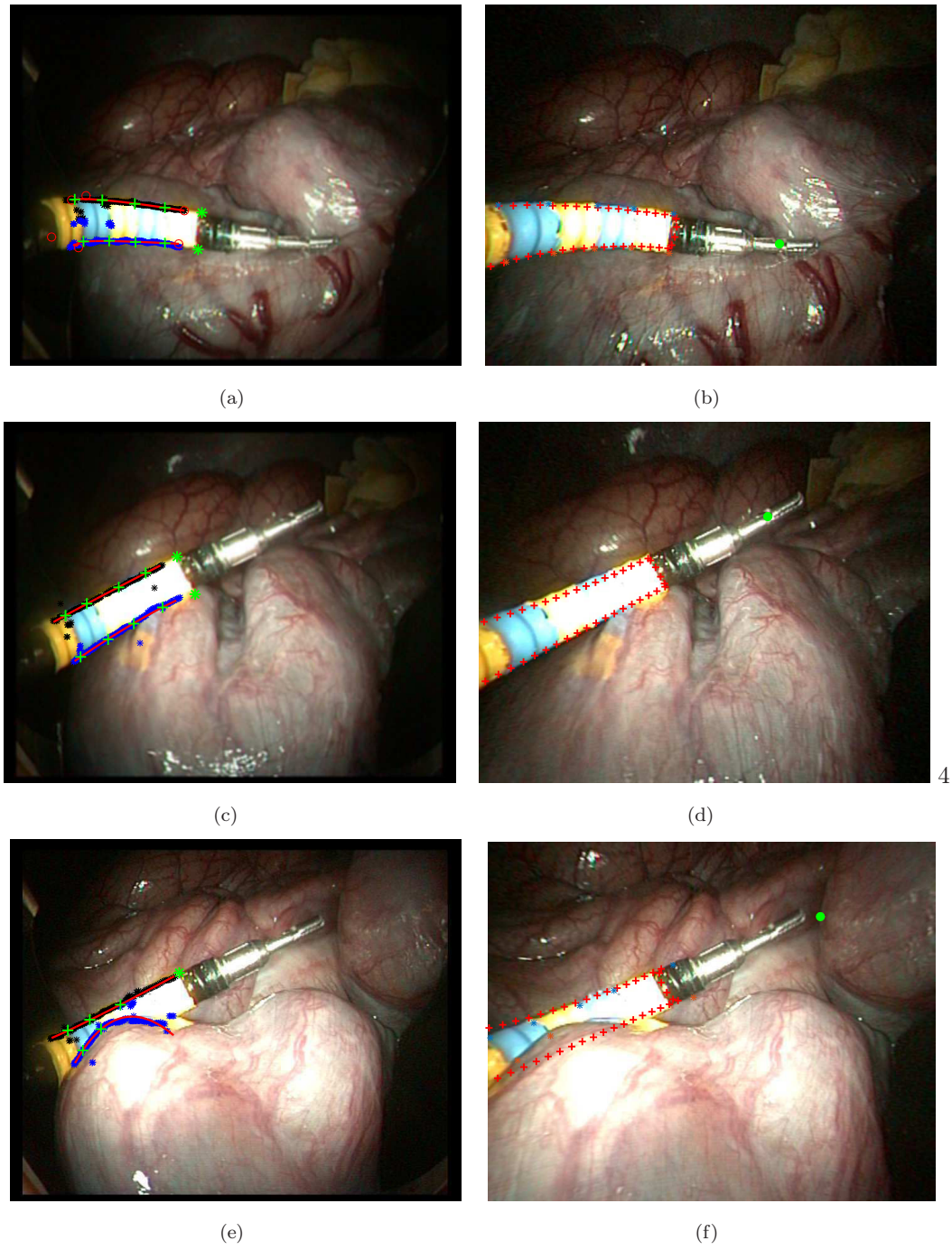
Sur la figure 11 quelques exemples significatifs sont montrés, où les instruments sont premièrement détectés (comme expliqué dans la section précédente) puis où la pose 3D est calculée en se basant sur les coins extraits (croix vertes). L'image virtuelle de l'instrument est reprojetée dans l'image originale avec la position estimée du centre de la pince (point vert).

Ces résultats semblent confirmer la capacité de la méthode proposée à estimer la pose d'un instrument béquillable en se basant sur les coins des marqueurs colorés.

En outre, l'introduction des tolérances pour les paramètres mécaniques permet au procédé de converger même si l'instrument est en train de pousser sur les organes et que, en conséquence, sa position dans le canal est modifiée (fig. 11(a) et (b)).

L'adaptabilité du schéma proposé est aussi démontrée par le fait que le système flexible utilisé pour les expériences *in vivo* est différent de celui employé en laboratoire. Néanmoins l'estimation de la pose fonctionne sans aucune modification de l'algorithme.

Finalement, même si l'information visuelle est partielle (fig. 11(e)), la pose 3D calculée est acceptable en termes de reprojection dans l'image.



(a) (b)

(c) (d)

(e) (f)

FIGURE 11 – Résultats *in-vivo* de la méthode présentée : (gauche) le résultat de la segmentation et (droite) la reprojection de l'instrument à partir de la pose estimée. Le point vert est la reprojection du centre de la pince. Dans (a), les jeux mécaniques considérés permettent d'approcher l'instrument même s'il est en train de pousser sur l'organe. Dans (c), l'instrument réel est correctement discriminé de son reflet sur l'organe. Même si la segmentation échoue partiellement (e), la pose estimée peut être considérée acceptable du point de vue qualitatif en termes d'erreur de reprojection (cf. (f)).

6 Approche basée apprentissage

Les erreurs résiduelles de l'approche basée modèle peuvent être expliquées en partie par l'utilisation d'un modèle paramétrique du système. En effet, la tâche de déterminer et décrire tous les aspects qui influent sur le comportement de l'instrument est ardue et le résultat serait un modèle complexe et difficilement exploitable ou transférable à d'autres systèmes. Pour éviter l'utilisation de modèles, nous avons étudié une deuxième approche où l'idée de base est d' *apprendre* la fonction qui relie les indices visuels dans l'image avec la position 3D du TCP.

La forme de ce type de fonctions est difficilement déterminable a priori. Pour cette raison nous avons choisi d'étudier les potentiels de plusieurs méthodes générales parmi lesquelles nous avons choisi les deux qui ont donné les meilleurs résultats : un réseau de fonctions à base radiale (RBF) et une régression localement pondérée (LWR). La première approche modélise la fonction comme une somme pondérée de K noyaux gaussiens φ_k (modèle local) dont les poids sont calculés d'après un critère global [3, 8]. Plus en détails, étant donné $\mathbf{x} = [x_1, \cdots, x_{2n}]^T$ le vecteur d'entrée, l'estimation d'une sortie $\hat{y}$ obtenue par un réseau de fonctions à base radiale peut être écrite comme :

$$\hat{y} = h(\mathbf{x}) = \sum_{k=1}^{K} \gamma_k \varphi_k(\mathbf{x}) + \gamma_{K+1} \tag{7}$$

où γ_k sont les coefficients appris qui décrivent la contribution du k-ème noyau à l'estimation de la sortie. Dans le cas de noyaux gaussiens, φ_k est défini comme [2, 6] :

$$\varphi_k(\mathbf{x}) = \mathcal{Y}_k \, e^{-\frac{1}{2}(\mathbf{x}-\mathbf{C}_k)^T \Sigma_k^{-1} (\mathbf{x}-\mathbf{C}_k)} \tag{8}$$

où $\mathbf{C}_k$ est le centre de chaque noyau, Σ_k sa variance et $\mathcal{Y}_k$ son amplitude.

Cette méthode nécessite une étape initiale d'apprentissage divisée en deux pas. Le premier pas répartit les donnée d'entrée en clusters (en utilisant un algorithme type k-means) et, selon les composants de chaque cluster, calcule la forme et la position des gaussiennes ; le deuxième pas détermine la pondération des noyaux qui permet de décrire au mieux la relation entrée/sortie.

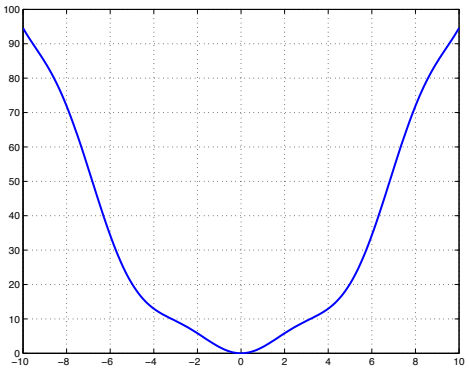
La seconde approche effectue une approximation linéaire en fonction de l'information locale de l'ensemble d'apprentissage [1]. Cette méthode ne nécessite pas d'étape d'entraînement car toutes les données de l'ensemble d'apprentissage sont utilisées lors d'une nouvelle estimation. Comme visualisé en figure 13(a), la valeur estimée de la sortie pour un nouvel échantillon d'entrée (*query point*) est le résultat d'une approximation linéaire où la contribution de chaque échantillon de l'ensemble d'entraînement (x_m) est déterminée selon sa distance à la nouvelle entrée (query point) $\mathbf{x}_q$. En appliquant ce critère, la fonction de coût qui en résulte est :

$$\chi^2 = \frac{1}{2} \sum_{m=1}^{M} w_m(\mathbf{x}_q) \, ||y_{m,i} - \gamma^T \mathbf{x_m}||^2 \tag{9}$$

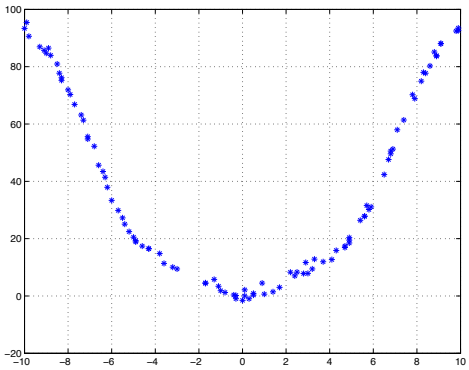
où les poids peuvent être exprimés comme

$$w_m(\mathbf{x}_q) = \exp\left(-\frac{1}{2} (\mathbf{x_m} - \mathbf{x}_q)^T D (\mathbf{x_m} - \mathbf{x}_q) \right) \tag{10}$$

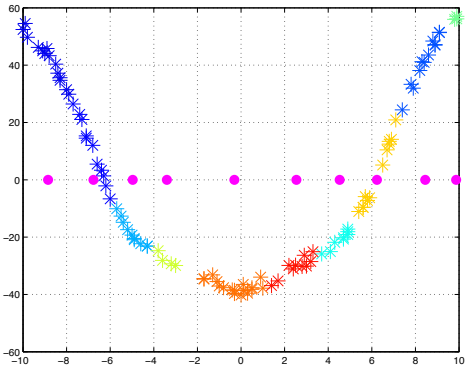
où D est une matrice semi-définie positive, qui définit une métrique de distance determinant l'influence de chaque element de l'ensemble d'entraînement dans le calcul du model de regression.



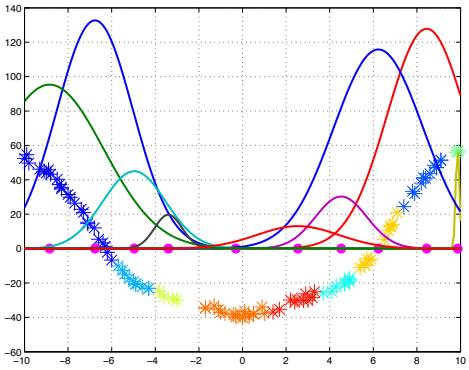
(a)



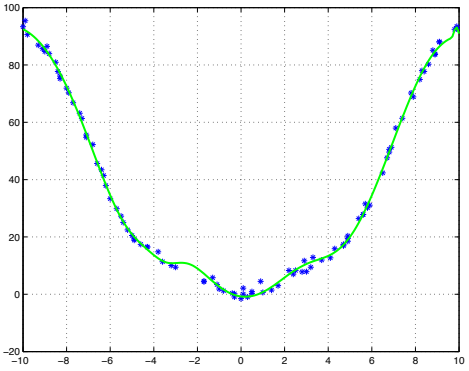
(b)



(c)

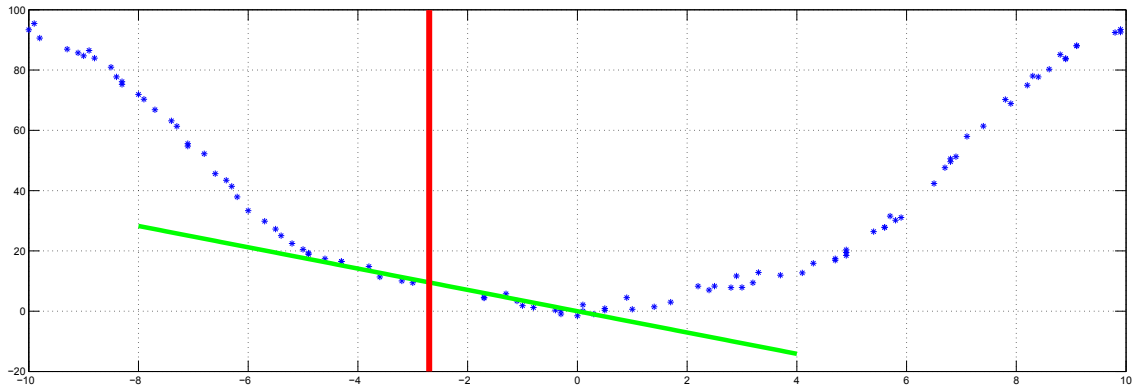


(d)

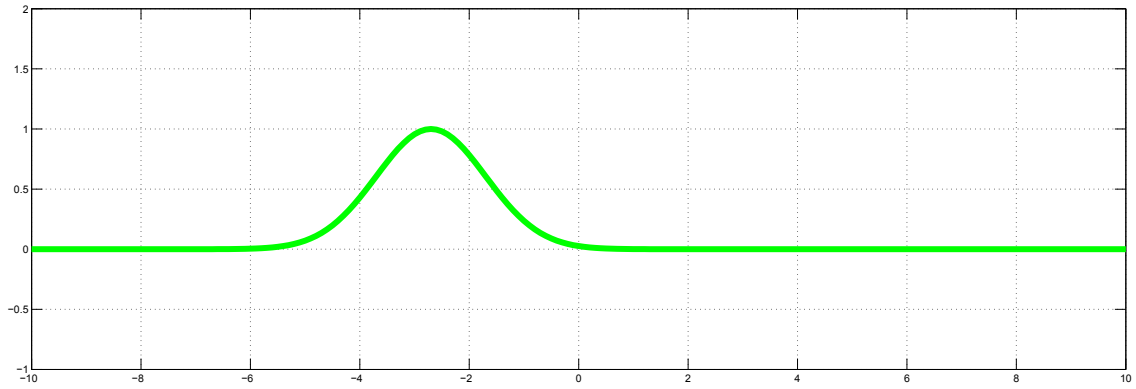


(e)

FIGURE 12 – Illustration de l'entraînement d'un réseau RBF dans le cas de la fonction montrée en (a) avec une seule entrée et sortie. L'ensemble d'entraînement, composé par plusieurs échantillons d'entrées et sorties (dans notre cas, coordonnées des centres et position 3D de l'instrument respectivement) (b), est réorganisé en clusters (c) à travers un algorithme d'apprentissage non supervisé de type k-means. Ensuite, selon les composantes de chaque clusters, la forme et les centres de chaque noyau sont calculés (d). Finalement, la meilleure pondération est déterminée par la méthode des moindres carrés.



(a)



(b)

FIGURE 13 – Principe de l'approche LWR pour un cas monodimensionnel. Pour une nouvelle entrée (*query point*) (red line in (a)), les poids associés à tous les échantillons de l'ensemble d'entraînement sont calculés en fonction de la mesure de distance adoptée (comme celle proposée en (b)) et par rapport à l'entrée considérée (*query point*). Ensuite, l'équation (11) est utilisée pour obtenir l'approximation linéaire, qui est valide seulement dans le voisinage du *query point*.

Les coefficients de régression peuvent ensuite être calculés comme :

$$\gamma = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y} \tag{11}$$

où $\mathbf{W}$ est une matrice diagonale contenant les poids w_m.

Pour limiter la redondance d'information qui semble s'opposer à l'efficacité de ces méthodes, la dimension du domaine d'entrée a été réduite. Pour cela nous avons utilisé les 5 centres des marqueurs au lieu des coins de ceux-ci.

En outre, deux améliorations ont été apportées par rapport aux méthodes originales. La première vise à mitiger l'effet du bruit en l'intégrant dans l'ensemble d'entraînement. L'idée est d'enrichir l'ensemble de données d'entraînement (composé par les coordonnées des centres $\mathbf{x}$ et la position donnée par la vision 3D stéréo Y) en tenant en compte des possibles erreurs de segmentation. Cet enrichissement est obtenu en ajoutant, pour chaque élément de l'ensemble d'entraînement $X = (\mathbf{x}, Y)$, deux (ou plus) échantillons artificiels ayant la même sortie Y (position 3D) que l'élément original et comme entrées les coordonnées originales des centres perturbées par un bruit uniforme $(\mathbf{x} + \Delta\mathbf{x})$.

La deuxième amélioration concerne le réseau de RBF et plus spécifiquement le clustering initial. En effet, si le clustering est réalisé en ne considérant que les données d'entrée, il est possible que les données de sortie associées à chaque cluster ne soient pas représentables par une gaussienne. Un exemple simple de ce phénomène se visualise quand l'instrument est plié et tourne autour de son axe principal. Si on considère, comme entrée de notre fonction, la coordonnée x du dernier marqueur et comme sortie la distance du bout de l'instrument par rapport à la camera, on obtienne une tendance décrite par la fig. 14(b). Ce comportement serait bien décrit par une gaussienne très ample comme montré en fig. 14(d). En revanche, si l'algorithme de cluster se base seulement sur le domaine d'entrée, le risque est d'obtenir un résultat de la forme donnée à la fig. 14(c). Dans ce cas, la combinaison des fonctions à base radiale n'arrive pas à décrire correctement la tendance de la fonction dans cette portion du domaine. Nous avons donc proposé d'appliquer l'algorithme de clustering dans un domaine conjoint entrée/sortie afin que la distance des données dans le domaine de sortie soit prise en compte pour le clustering. L'objectif est de favoriser un résultat similaire à celui de la fig. 14(d) où les données proches dans le domaine de sortie sont rassemblées dans le même sous-ensemble. Dans ce cas, la gaussienne qui décrit le cluster prend en compte la variance de ces nouveaux clusters en prenant une forme plus aplatie. Ce domaine étendu est composé par les échantillons obtenus en rajoutant les variables de sortie $Y = (y_1, y_2, ...y_m)$ aux coordonnées d'entrée $\mathbf{x}$. On obtient alors $\mathbf{x}' = [x_1, \cdots, x_{2n}, y_1, y_2, ..., y_m]^T$. Dans notre cas, cela consiste à ajouter les coordonnées 3D de l'instrument à la séquence des coordonnées 2D (image) des centres des cinq marqueurs.

Comme montré en figure 15, ces deux modifications apportent des bénéfices aux deux approches en améliorant la précision de l'estimation de la position 3D du bout de l'instrument.

La validation a été effectuée sur le même ensemble d'images que pour la solution basée modèle, en utilisant une partie des données pour l'apprentissage. L'erreur RMS entre les valeurs estimées et les coordonnées de référence est (respectivement pour les coordonnées x, y et z du TCP) : 0.92 ± 0.61, 1.2 ± 0.62 et 2.49 ± 1.49 mm pour le réseau de RBF et 0.7 ± 0.43, 0.74 ± 0.49 et 1.74 ± 1.08 pour le LWR (fig. 16).

Les méthodes proposées donnent des résultats comparables et offrent une précision supérieure aux approches précédemment proposées dans la littérature. Néanmoins chacune d'entre elles présente des inconvénients : d'un côté le traitement d'image nécessaire à l'approche basée modèle est critique. D'un autre côté, les besoins pour l'approche d'apprentissage sont plus exigeants, car elle nécessite un système spécifique pour la création d'un ensemble d'entraînement fiable contenant la position 3D du TCP. En outre, ce dernier type d'approches fournit seulement la position 3D de l'outil alors que l'approche basée modèle rend également compte de son orientation qui peut être utile dans certaines opérations. Toutefois, dans le contexte considéré, ce dernier aspect n'est pas critique puisque l'instrument possède seulement 3 DDL et seule sa position dans l'espace peut être commandée, l'orientation dépendant de celle-ci. Le choix final dépendra donc fortement de l'utilisation recherchée.

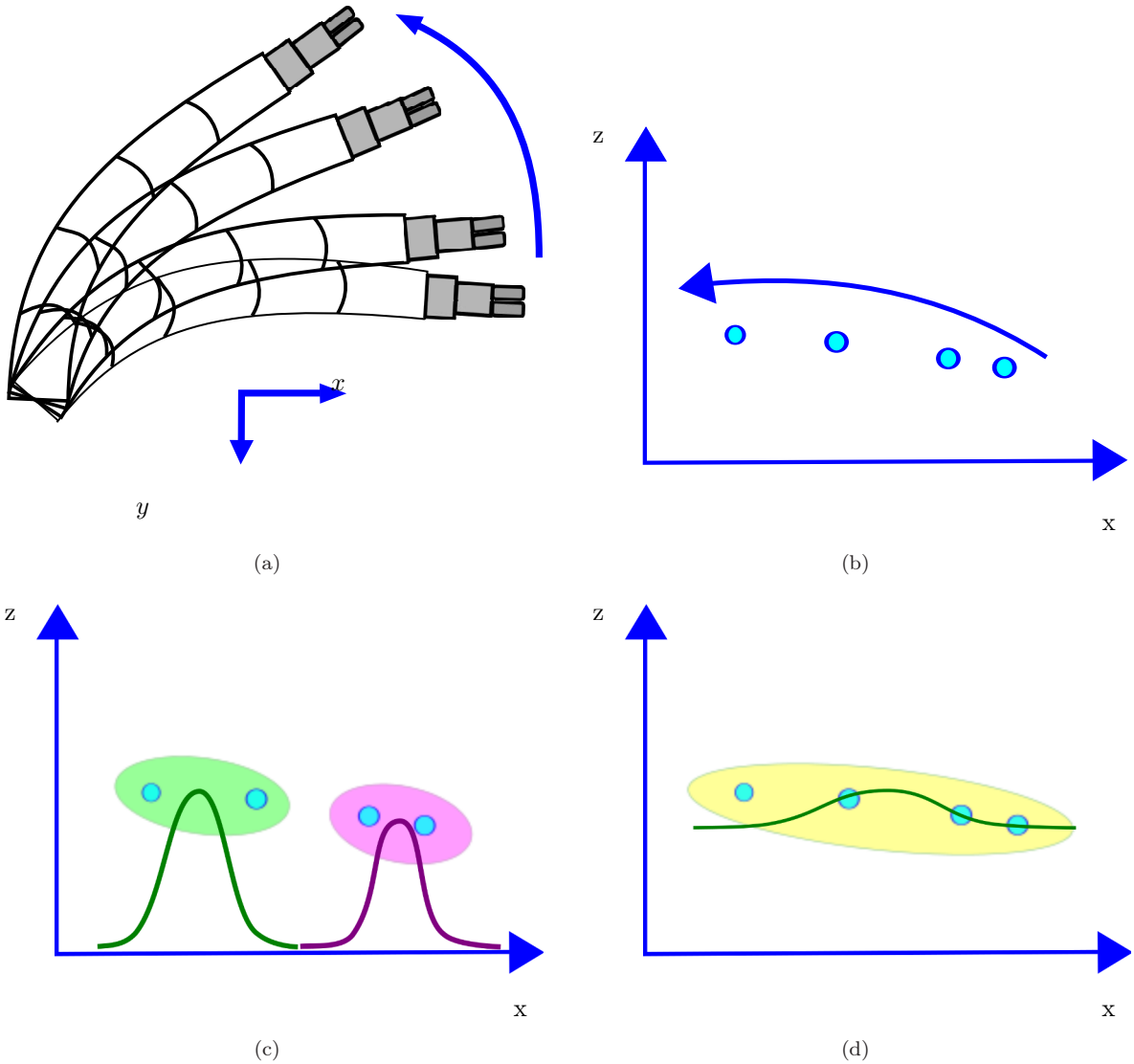


FIGURE 14 – Exemple illustrant la problématique liée au clustering initial avec l'approche RBF. En supposant que l'instrument est plié et tourne autour de son axe (a) et que nous voulons estimer la relation entre la coordonnée x du bout de l'instrument et sa distance à la camera z (b), l'algorithme de clustering (type k-means) utilisé en considérant seulement les données d'entrée (coordonnée x) produit un résultat qui amène à une mauvaise approximation (c). Si, au contraire, on prend en compte aussi les données de sortie (la coordonnée z qui décrit la profondeur), il sera plus probable d'obtenir un résultat de clustering comme celui de la fig. (d) qui amène à une gaussienne dont la forme décrit mieux la tendance de la relation position-profondeur.

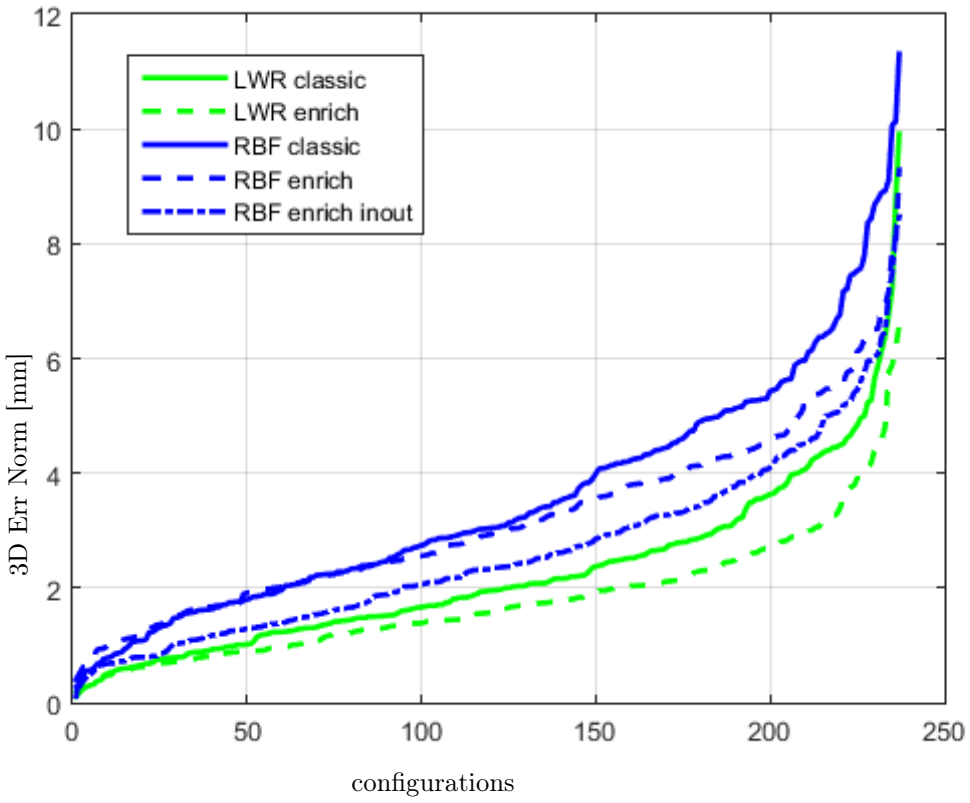


FIGURE 15 – Effet des amélioration apportées aux deux méthodes basées apprentissage. Sur le graphique, nous pouvons observer que l'enrichissement de l'ensemble d'entraînement améliore la capacité d'approximation des RBF (cf. courbes *RBF classic* vs *RBF enrich*) et du LWR (cf. courbes *LWR classic* vs *LWR enrich*). Finalement, dans le cas du RBF, le clustering réalisé dans le domaine conjoint entrée/sortie (courbe *RBF enrich inout*) améliore le résultat obtenu avec l'enrichissement seul.

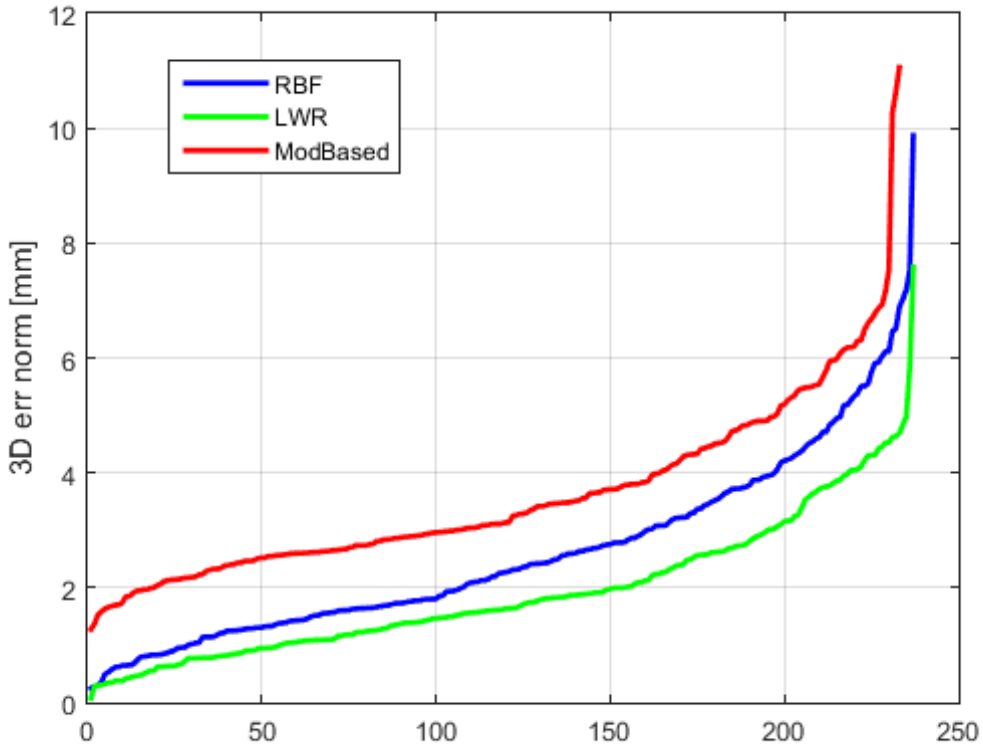


FIGURE 16 – Comparaison entre les trois approches proposées pour l'estimation de la position 3D du bout de l'instrument. Les trois courbes représentent les normes des erreurs commises sur l'estimation de la position 3D pour différentes configurations dans l'espace de travail et ordonnées. La méthode de régression localement pondérée (LWR) semble donner le meilleur résultat par rapport aux RBF et la méthode basée modèle.

7 Conclusion

Dans ce travail, le problème de l'estimation de la position 3D d'un instrument flexible de chirurgie endoscopique a été traité dans son ensemble : choix et segmentation de primitives visuelles et obtention d'informations 3D à partir de celles-ci. Une méthode spécifique de segmentation d'image a été présentée dans le cadre de l'endoscopie flexible, qui permet d'extraire le squelette et les coins apparents de marqueurs colorés sur un instrument flexible en environnement in-vivo. Deux solutions pour l'estimation de la pose à partir de l'image ont été décrites : une méthode basée modèle et une méthode sans modèle, basée apprentissage. La première permet d'estimer la configuration complète du robot et est robuste aux changements de la géométrie du système dus, par exemple, aux jeux mécaniques entre l'instrument et les canaux. La deuxième permet de surmonter les problèmes dérivants d'une connaissance partielle du système et donne une meilleure estimation de la position 3D du TCP. En revanche, aucune information sur la configuration n'est obtenue.

Une suite logique à ce travail serait l'implémentation d'un système de suivi multi-images pour rendre la détection des primitives visuelles plus robuste, ainsi que l'estimation de la position 3D. Cela permettrait aussi de rendre tout le procédé plus rapide pour ensuite utiliser l'estimation de pose pour commander un instrument flexible en boucle fermée.

Enfin, une extension intéressante pourrait être d'utiliser la forme et la position détectées par la vision pour déterminer (et éventuellement estimer) les forces externes (par exemple dues à l'interaction avec les tissus) appliquées sur l'instrument.

Références

[1] Christopher G. Atkeson, Andrew W. Moore, and Stefan Schaal. Locally Weighted Learning. In David W. Aha, editor, *Lazy Learning*, pages 11–73. Springer Netherlands, 1997.

[2] Adrian Gheorghe Bors. Introduction of the radial basis function (rbf) networks. In *Online Symposium for Electronics Engineers*, 2001.

[3] D.S. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2 :321–355, 1988.

[4] Antonio De Donno, Lucile Zorn, Philippe Zanne, Florent Nageotte, and Michel de Mathelin. Introducing STRAS : a new flexible robotic system for minimally invasive surgery. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 1213–1220. IEEE, 2013.

[5] G. Dogangil, B.L. Davies, and F. Rodriguez y Baena. A review of medical robotics for minimally invasive soft tissue surgery. In *Proc. of the Institution of Mechanical Engineers*, volume 224, pages 653–679, May 2010.

[6] A. Lendasse, J. Lee, E. de Bodt, V. Wertz, and M. Verleysen. Approximation by radial basis function networks - application to option pricing. In *Connectionist Approaches in Economics and Management Sciences*, pages 201–212. Kluwer academic publishers, 2003.

[7] E. Marchand and F. Chaumette. Virtual visual servoing : a framework for real-time augmented reality. *Eurographics*, 21(3), 2002.

[8] John Moody and Christian J. Darken. Fast learning in networks of locally-tuned processing units. *Neural Comput.*, 1(2) :281–294, 1989.

[9] R. Reilink, S. Stramigioli, and S. Misra. Pose reconstruction of flexible instruments from endoscopic images using markers. In *IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 2938–2943, 2012.

[10] Robert J. Webster and Bryan A. Jones. Design and kinematic modeling of constant curvature continuum robots : A review. *The International Journal of Robotics Research*, 29(13) :1661–1683, 2010.

Paolo CABRAS

3D Pose Estimation for the Control of Flexible Instruments in Robotic Endoscopic Surgery.

Résumé

Grâce à leur dextérité, les systèmes flexibles et endoscopiques peuvent atteindre des zones distales à l'intérieur du corps, d'une manière vraiment micro-invasive à travers des orifices naturels, permettant ainsi la réalisation d'un geste chirurgicale sans ou avec une seule cicatrice. Etre capable de mesurer la position 3-D de ces instruments peut être utile pour diverses tâches, telles que le contrôle automatique des instruments robotisés ou le guidage des gestes du chirurgien. Dans cette thèse, nous présentons plusieurs méthodes automatiques pour déduire la pose 3-D d'un instrument à section de pliage unique en utilisant seulement les images fournies par la caméra monoculaire intégrée à l'extrémité de l'endoscope. Les deux méthodes que nous proposons reposent sur l'emploi de marqueurs de couleur attachés à la section pliable de l'instrument. L'image acquise est segmentée en utilisant une méthode basée sur les graphes et, successivement, les coins des marqueurs sont extraits par la détection de la transition de couleur le long des courbes de Bézier qui modélisent les contours. Ces primitives visuelles servent alors à l'estimation de la pose 3-D de l'instrument à l'aide d'un modèle adaptatif de ce dernier qui prend en compte le jeu entre l'instrument et son canal de logement. Les fortes incertitudes sur la modélisation géométrique d'un tel instrument robotisé et les nombreuses variabilités inhérentes à l'environnement in-vivo affectent grandement la précision du résultat. Par conséquent, deux approches basées sur de l'apprentissage ont été étudiées pour bien refléter la relation entre la position 3-D du bout de l'instrument et l'image associée à sa projection: un réseau de fonctions à base radiale (RBF) et une régression localement pondérée (LWR). La première approche modélise la fonction comme une somme pondérée de gaussiennes dont les poids sont calculés d'après un critère global. La seconde approche effectue une approximation linéaire en fonction de l'information locale de l'ensemble d'apprentissage. Toutes les méthodes proposées ont été validées sur une cellule expérimentale robotique au laboratoire ICube et sur des séquences in-vivo. De l'analyse des résultats obtenus, les avantages et inconvénients de chaque méthode sont énoncés et décrits.

Mots-clés: Robotique Medicale Flexible, Estimation de la pose 3D à partir d'une seule image, extraction des primitives pour instrument flexible in *in-vivo*, apprentissage.

Abstract

Thanks to their dexterity and compliance, flexible systems can reach distal body zone in a real non-invasive way through natural orifice allowing no-scar or single scar surgery. Being able to measure the 3D position of such instruments can be useful for various tasks, such as controlling automatically the robotized instruments or providing gesture guidance. In this thesis, we propose two automatic methods to infer the 3D pose of a single bending section instrument using only the images provided by the monocular camera embedded at the tip of the endoscope. Both methods relies on colored markers attached onto the bending section. The image of the instrument is segmented using a graph-based method and the corners of the markers are extracted by detecting the color transition along Bézier curves fitted on edge points. These features are then used to estimate the 3D pose of the instrument using an adaptive model that takes into account backlash between the instrument and its housing channel. Strong model uncertainties can affect the result of such model-based method. Therefore, two learning approaches have been studied that approximate image features-to-3D function according to a training set: Radial Basis Function (RBF) Network and Locally Weighted Regression (LWR). The first proposal approximate the function as a weighted sum of Gaussians whose weights are computed according to a global criteria. The other approach, instead, performs a linear approximation according to local information of the training set. The proposed methods are validated on a robotic experimental cell and in *in-vivo* sequences. Advantages and inconveniences of each method are presented and commented.

Key-words: Medical Flexible Robotics, Single-Image-Based 3D Pose Estimation, *In-vivo* Flexible Instrument Segmentation, Learning Regression.